COMPENDIUM OF ADMINISTRATIVE DATA SOURCES FOR SELF-SUFFICIENCY RESEARCH
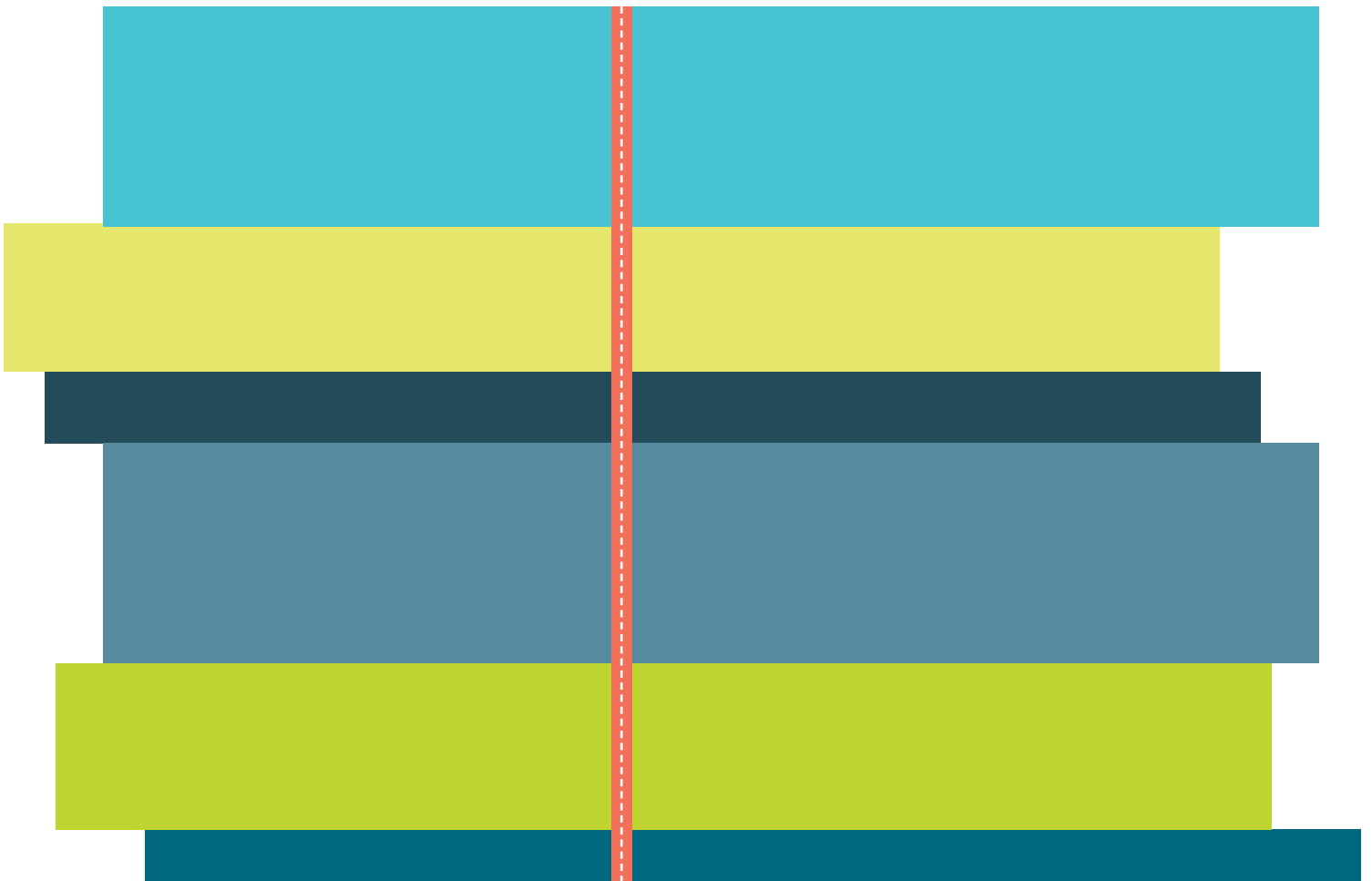
Compendium of Administrative Data Sources for Self-Sufficiency Research

Editors: Daron Holman, Alexandra Pennington, Kelsey Schaberg, Andrew Rock

This report and other reports sponsored by the Office of Planning, Research, and Evaluation are available at www.acf.hhs.gov/opre.

[Sign-up for the OPRE Newsletter](#)

Follow OPRE on Twitter
[@OPRE_ACF](#)

Like OPRE on Facebook
[facebook.com/OPRE.ACF](#)

Follow OPRE on Instagram
[@opre_acf](#)

OPRE

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

Table of Contents

Acknowledgments

Introduction

The *Compendium of Administrative Data Sources for Self-Sufficiency Research* is an effort to describe promising administrative data sources for evaluations of economic and social interventions. The Compendium was created as part of the Assessing Options to Evaluate Long-Term Outcomes Using Administrative Data (LTO) project funded by the Administration for Children and Families' Office of Planning, Research, and Evaluation (ACF/OPRE) in the U.S. Department of Health and Human Services.

LTO Project Overview

Many social programs are designed to have long-term benefits for participants, but evaluations of these programs rarely track outcomes in the long term, often because of limited resources. Administrative data present a potentially low-cost opportunity for tracking long-term effects. Efforts related to making these data more accessible for such purposes are gaining traction, with many federal initiatives emerging to support leveraging these data for research, particularly in light of the recommendations by the Commission on Evidence-Based Policymaking,[1] and the passage of the Foundations for Evidence-based Policymaking Act of 2018.[2]

The LTO project is helping ACF/OPRE understand the feasibility of linking data sets for a set of major evaluations. Evaluations and administrative data sources are being selected and reviewed to assess the feasibility of linking data for long-run follow-up. This Compendium is part of this project and is intended to provide information on administrative sources to facilitate linkages to measure impacts of social programs in both the medium and long terms.

Purpose of This Compendium

There is growing interest in making better use of administrative data — which are collected primarily to manage programs — to support research on program effectiveness and evidence-based policymaking. The first step toward this goal is to better understand (1) the types and extent of data available and where they reside; (2) the process for obtaining data access; and (3) the feasibility of linking these data sources with evaluation data to measure the impact of government-funded programs. This Compendium includes such information for a variety of national, federal, and state-level administrative data sources that can be enlisted to support this federal priority. It is not an exhaustive list of all the data sources that might be useful for this purpose.

About the Information-Gathering Process

Data sources were selected for this Compendium based on the following ACF priorities:

- those commonly used for federally funded evaluations of economic and social interventions
- those capable of facilitating research on a variety of social and economic outcomes
- those that can be linked to other data sources, especially program evaluations
- those that contain the data needed to measure long-term outcomes
- other ACF-recommended sources, including those with potential to be used for research purposes

---

[1][Commission on Evidence-Based Policy Making (CEP) Final Report](#)
[2][Evidence-based Policymaking Act of 2018](#)

A two-stage process was used to compile information on each source. In the first stage, publicly available resources were scanned for the following metadata: (1) data content, coverage, and availability; (2) the process for obtaining access to these data (including any restrictions regarding who can access the data and how); and (3) the steps involved in linking evaluation data to the source (if such linkages are possible). In the second stage, drafts summarizing the metadata were submitted to data providers for review to ensure accuracy and completeness. For this reason, the information contained here is only a snapshot of these data sources as of fall 2018 and is subject to change if updates to the data sources and/or the procedures for linkage are made. Some of the information is listed as "Not available," which signifies that the data provider left this field blank. In other cases, the data provider noted some information as "NA" if the field was deemed not applicable.

This Compendium is not comprehensive. There are far too many data sources to cover all of them, and some sources are not amenable to these research purposes. Other compendiums that are either currently available or in progress cover some of the same and some additional data sources.[3]

In the short term, this resource can help provide important reference material for data consumers, including government agencies and their research partners. With increased interest and use of administrative data, and policy changes affecting access to these data, this contribution should — in the long term — help to advance the development of a much broader and sustainable repository of metadata on administrative data sources for self-sufficiency research.

How to Read This Compendium

Data sources are grouped and presented as follows:

- *National and federal data sources*, which are typically focused on a single domain but have national coverage (Part 1)
- *State-level sources,* which cover a single domain within the state (Part 2)
- *Data centers*, which maintain information across multiple domains (for example, employment, health, public assistance, and so forth) and from multiple data sources (Part 3)

Each part begins with background information as well as a glossary of terms that describe the types of information gathered (if available) about each data source.

---

[3]For example, see the forthcoming Compendium of Administrative and Survey Data Resources in the Administration for Children and Families and the Abdul Latif Jameel Poverty Action Lab's Catalog of Administrative Data Sets.

Part 1: National/Federal Administrative Data Sources

This section features data sources with national coverage. All but two of these data sources — the Eviction Lab and the National Student Clearinghouse — contain data from a federal program, policy, or survey and are overseen by a federal agency.

The data sources overseen by federal agencies typically have more restrictions around who can access the data and require lengthier application and acquisition processes than nonfederal sources. These restrictions are often due to federal laws and regulations. For example, many sources, including some in this collection, require that matched data files reside on federal servers at all times, and some sources remove all personally identifiable information before returning data files to researchers.

One benefit of national data sources is that they contain data from all or most states and jurisdictions. Researchers working on studies with sites in multiple states would be able to obtain data for all sites through one source rather than going through multiple individual states. Because many national sources rely on and contain state-reported data, however, there can be significant time lags before the data are available. The availability of up-to-date information across states could take a while. Additionally, some national data sources may have issues with data coverage if, for example, states are not required to report all of their data.

Some of the data sources in this section are routinely used to track the outcomes and impacts of major social interventions and — as such — have well-defined data acquisition procedures for research purposes, while other sources do not allow or do not currently have the capacity to consider research data requests.

Glossary — National/Federal Data Sources

Overview

Content/domain A general description of the domains included in the data source — for example, demographic, public assistance, employment, or health data.

Ownership/authorizing agency The organization or agency that has the authority to provide access to the data for research purposes. This may not be the organization or agency that originally collected the data.

Data Availability

Data element categories The types or categories of data available — for example, personally identifiable information (PII), geographic information, or quarterly wages.

Geographic coverage Geographic location(s) the data cover — for example, national data or state- or county-based data.

Population Who is covered by the data source — for example, all Temporary Assistance for Needy Families cash assistance recipients or all individuals with an assigned Social Security Number (SSN).

Time period coverage The time period (years or months) for which data from each data set were available as of fall 2018. If data are collected for only parts of a year, a description may be provided.

Periodicity How often the data are collected and/or in what time interval the data elements are aggregated — for example, annually or quarterly; Supplemental Nutrition Assistance Program payments are reported monthly.

Lag The amount of time between when the data are collected and when they are available to researchers (including other state or federal agencies and their contractors). For example, if wage records for Quarter 1, 2014, are available to researchers in Quarter 3, 2014, the lag is two quarters.

Data retention, update, and deletion schedule(s) The time period for which data are retained before purging or archiving occurs, how often the data are updated, and/or on what schedule data are deleted (if applicable) — for example, data may be retained for 24 months before purging and updated on a quarterly basis.

Data Access

Data acquisition process Description of how researchers request access to the data source, including a summary of the process to obtain approval and access — for example, a research request might need to be submitted for approval, evidence of approval by an institutional review board (IRB) or study participant consent (and/or waiver of confidentiality) might be required, or a background check might be conducted.

Are there restrictions on who can access the data? Yes/no answer to whether there are any restrictions on who can apply for access to the data for research purposes. If yes, a summary of the restrictions may be provided — for example, third-party researchers might be eligible; access may be limited to (local,

state, or federal) government contractors or research partners; or those wanting to access the data might need to sign a nondisclosure agreement.

How long does the data acquisition process take? A (rough) timeframe for how long the data acquisition process usually takes from the point when researchers first apply for access until the point when they are able to work with the data. If applicable, a timeframe may be given for each step in the acquisition process — for example, researchers might need to obtain Special Sworn Status before data can be accessed, a process that can take 4 to 6 weeks.

Is a data-sharing agreement or memorandum of understanding required? Yes/no answer to whether a data-sharing agreement of memorandum of understanding is required to access the data center. If yes, a description of the requirements may be provided.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes/no answer to whether evidence of an IRB review is required to permit access to the data. If yes, a description of the requirements may be provided.

Access location(s) The physical or virtual location(s) where approved researchers may access the data, including any data centers that house the data source. If there are physical locations, a summary of how many there are and where they are located may be provided. If the data can be transmitted securely to researchers (for example, via Secure File Transfer Protocol, or SFTP), this information may be provided.

Matching and Working with Data

Can a research sample be matched to the data source? Yes/no answer to whether researchers can submit a file with personal identifying information to the authorizing agency — for example, researchers may be able to send a list of SSNs for 1,000 study sample members, and that list is used to match to the data source.

Matching process Summary of the process by which researchers initiate a match to the data source after they have been granted approval to work with the data. This may include a list of the steps in the matching process with links to any online resources that describe the steps, as appropriate — for example, how researchers submit a data file and how the data are linked.

What identifiers are needed for matching? A list of the person-level identifiers required to link/match a research sample to the data available through the data center — for example, SSNs. If a data center allows for fuzzy matching, a list of the fields — such as name, date of birth, and phone number — used in the algorithm may be provided.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Yes/no answer to whether there are any restrictions on the data file that is returned to researchers. If yes, a list of the restrictions may be provided — for example, researchers may be allowed to work only with deidentified data or aggregate data.

Other restrictions for matching and working with data Summary of the limitations around accessing, matching, or using the data. These could be limitations around specific data sets or elements — for example, perhaps researchers must work with the data on-site, or researchers are given access only to deidentified data.

Cost

Is there a fee associated with accessing the data? Yes/no answer to whether researchers are charged a fee to access the data. If yes, an overview of the fee structure may be provided.

Documentation

Links to any publicly available documentation. These sources could include how to apply for access to the data, standard formats or record layouts of input and output files, examples of research publications that feature analysis of matched data, or other relevant information.

Child Care and Development Fund (CCDF) Case-Level Administrative Data (ACF-801 Data)

Overview

Content/domain Child care assistance and use.

Ownership/authorizing agency Office of Child Care (OCC), Administration for Children and Families (ACF).

Data Availability

Data element categories Program participation data about families that have received assistance from the CCDF and related child care programs. Measures include information on the reasons for child care, the assistance provided (months and amounts), characteristics of the family and child receiving the child care subsidy, and the child care provider.

Data are available at the family, child, setting, and provider levels:[4]

- Family level: head of household single status, family size, primary language spoken at home, income, homeless status, reason for receiving child care assistance, income sources, child care copayment amount, FIPS (county) code, zip code
- Child level: age, sex, disability, month and year of birth, race, ethnicity
- Setting level: type of child care, hours of child care provided, amount paid to provider
- Provider level: Quality Rating and Improvement System (QRIS) participation and rating, accreditation, whether provider is subject to state pre-K and Head Start/Early Head Start standards, zip code

Geographic coverage National, but case-level data are reported to ACF by states, territories, and the District of Columbia (referred to as States).

Population Children and families receiving assistance through their states funded with CCDF dollars. States are allowed to report full-population or sample data.[5]

---

[4]Data quality may vary at the state and/or family level. In addition, variations in data quality across States may be a result of differences in state policies. More information on which variables this applies to can be found in the ACF Administrative and Survey Data Resources Compendium.

[5]States can choose to report either all subsidy cases that meet CCDF eligibility requirements or just the CCDF subsidy cases. This flexibility allows States that do not have the capacity in their data systems to identify the subsidy funding source for each child served to be able to report case-level and aggregate data. OCC has asked States to report a "pooling factor" in the annual aggregate ACF-800 report. The pooling factor enables OCC to report the number of families and children served by CCDF funds alone (as opposed to those served with CCDF funds combined with other sources of funding). The States that pool funds have to report all families and children

Time period coverage 2001 to present.

Periodicity Case-level data are reported monthly. Data are aggregated annually (on a fiscal year basis) and made available on an annual basis due to the interdependency of the ACF-801 and ACF-800 data (pooling factor).[6]

Lag Generally 10 to 11 months after the end of the fiscal year.[7]

Data retention, update, and deletion schedule(s) OCC releases a sample public-use file (with sample case-level records) that is available to researchers after OCC has published annual fiscal year reports.

Data Access

Data acquisition process Following the release of national reports, OCC makes a sample public-use data set available at Research Connections. Researchers interested in the public-use data can consult the archive's processes.

Are there restrictions on who can access the data? No. The sample public-use data file can be accessed by the general public.

How long does the data acquisition process take? The sample public-use data file can be downloaded from Research Connections at any time.

Is a data-sharing agreement or memorandum of understanding required? Researchers do not need a separate memorandum of understanding (MOU) with OCC to access the sample public-use data file.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? No.

Access location(s) Research Connections; OCC, ACF; and/or Federal Statistical Research Data Centers (FSRDCs).

Matching and Working with Data

Can a research sample be matched to the data source? Yes, for any data covering the period(s) prior to federal Fiscal Year 2015.[8]

---

receiving child care services who meet CCDF eligibility requirements. For these States, the pooling factor will be less than 100, and should be calculated based on the direct services costs for only those children and families reported (i.e., those meeting CCDF eligibility requirements). OCC will apply the pooling to adjust the counts of families and children served with CCDF funds after the State has reported its data. The States that do not pool funds have to report all families and children served with CCDF funds. For these States, a pooling factor of 100 must be reported. OCC will not make any adjustment to family and child counts after the State has reported its data.

[6]ACF-800 provides state-level aggregate reporting requirements, and ACF-801 provides child care monthly case record reporting requirements.

[7]OCC generally anticipates publishing national reports 10 to 11 months after the end of a fiscal year due to the interdependency of the ACF-801 and ACF-800 data (pooling factors).

[8]The Office of the Assistant Secretary for Planning and Evaluation (ASPE) is currently undertaking a project to compare variables from the child care data with data from the American Community Survey (ACS). More

Matching process Not available.

What identifiers are needed for matching? SSN or case identifier.[9]

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Not available.

Other restrictions for matching and working with data Not available.

Cost

Is there a fee associated with accessing the data? No fee is charged for accessing the sample public-use data. There may be a cost for accessing data that includes identifiers.

Documentation

➢ Child Care and Development Fund Statistics
➢ Research Connections: Child Care and Development Fund (CCDF) Administrative Data Series
➢ ICPSR: Child Care and Development Fund (CCDF) Administrative Data Series
➢ Office of Child Care: Reports to Congress
➢ Researching the CCDF Program by Linking Administrative Data with Data from the CCDF Policies Database: A How-To Guide
➢ Is Subsidized Childcare Associated with Lower Risk of Grade Retention for Low-Income Children? Evidence from Child Care and Development Fund Administrative Records Linked to the American Community Survey (conference presentation)
➢ ACF Administrative and Survey Data Resources Compendium[10]

---

information can be found here: Is Subsidized Childcare Associated with Lower Risk of Grade Retention for Low-Income Children?

[9]SSN was an optional field prior to federal Fiscal Year 2015, after which it was no longer collected. Given these parameters, match results may vary.

[10]ACF will release a compendium of administrative and survey data resources at a future date.

The Eviction Lab

Overview

Content/domain Formal eviction records.

Ownership/authorizing agency The Eviction Lab at Princeton University.

Data Availability

Data element categories Formal eviction records include information related to eviction court cases, such as defendant and plaintiff names, defendant addresses, monetary judgment information, and case outcomes.[11]

Geographic coverage National — publicly reported at state level through block group level, as well as census place.

Population Individuals who have faced formal eviction.

Time period coverage 2000-2016.

Periodicity The Eviction Lab receives annual updates to the database. Estimates are reported for each geography by year.

Lag Researchers receive annual updates in early fall of each year. Additional available data/estimates are made public in subsequent months.

Data retention, update, and deletion schedule(s) Records are retained indefinitely, with no deletion of records absent notification of an error in the record by the original data vendor. Data are updated on an annual basis.

Data Access

Data acquisition process The primary access point for data requests is https://evictionlab.org/get-the-data/. Researchers may request additional data via email to data.merge@evictionlab.org. In the email, they should provide the following information:

1. A description of the data.
2. Why they want to match with the eviction records and any research questions.
3. A description of the security measures their organization will take to secure the data provided and how confidentiality will be kept.
4. A copy of IRB approval/exemption for the project.

See more information on the Eviction Lab's data request application on its Data Request Application page.

Are there restrictions on who can access the data? Researchers who are affiliated with a university are encouraged to make requests. Researchers looking for data merges/custom data requests who are not

---

[11] FAQ on the Eviction Lab's methods

affiliated with a university should consult the Eviction Lab to determine if they are eligible by emailing data.merge@evictionlab.org.

How long does the data acquisition process take? Requests are approved within 4 to 6 weeks, depending on their complexity.

Is a data-sharing agreement or memorandum of understanding required? A data agreement, which outlines data access and crediting requirements, must be signed prior to execution of the request.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes. Researchers must attach a copy of their institution's IRB approval/exemption to the data merge application.

Access location(s) Data will be shared via Secure File Transfer Protocol (SFTP). Data must be stored in a secure location, accessible only by research personnel listed on the IRB approval/exemption. This information should be contained in the data request application mentioned above.

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process The Eviction Lab will merge data sets with eviction records by research subjects' names and addresses. It will then return deidentified data to the researcher, listing only case IDs and eviction information.[12]

What identifiers are needed for matching? Name and address.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Deidentified data, listing case IDs and eviction information, will be returned alongside other variables provided by the researcher that carry no risk of reidentifying tenants. Information that could potentially be used to reidentify individuals (for example, enrollment dates, geocoded information) will not be accepted or returned to researchers in the deidentified data set.

Other restrictions for matching and working with data For questions about the ability to merge certain variables, please email data.merge@evictionlab.org.

Cost

Is there a fee associated with accessing the data? No.

Documentation

➢ Eviction Lab website
➢ Information on the Eviction Lab's Data Request Application
➢ FAQ on the Eviction Lab's methods

---

[12] Information on the Eviction Lab's Data Request Application

Department of Housing and Urban Development Inventory Management System (IMS)/Public and Indian Housing Information Center (PIC)

Overview

Content/domain Housing and tenant data.

Ownership/authorizing agency U.S. Department of Housing and Urban Development, Office of Public and Indian Housing.

Data Availability

Data element categories 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. 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.

Geographic coverage National, plus external territories that have funded housing (Puerto Rico, U.S. Virgin Islands, Guam, and the Northern Marianas).

Population Public housing and Housing Choice Voucher (HCV) recipients.

Time period coverage 2000 to present.

Periodicity Data are collected daily from the housing agencies and copied to the PIC data warehouse weekly. Extracts are uploaded to the HUD data website and the HUD User portal quarterly.

Lag 1 week to 1 quarter, depending on data source.

Data retention, update, and deletion schedule(s) IMS/PIC is a system of record, and its data are used for funding; all data are retained. Assisted families that have left the program may have their 50058 or MTW-50058 data archived after 5 years.

Data Access

Data acquisition process Quarterly data sets are available from the HUD data website. Requests for specialized data extract requests are variable and require approval from the Privacy Officer. Researchers seeking HUD data must complete a data license application describing the proposed research project

and how they plan to keep the data secure. More information can be found at HUD's research partnerships website.

Are there restrictions on who can access the data? Aggregated tenant data as well as aggregated physical inventory data are available for immediate download at the HUD User website. Nonaggregated tenant and physical inventory may contain PII and require approval from the Privacy Officer.

How long does the data acquisition process take? Aggregated data are available for immediate download. Specialized data set requests may take 2 to 3 weeks to approve and extract.

Is a data-sharing agreement or memorandum of understanding required? Sometimes. It is not required for one-time extraction or data pulled by a researcher using the HUD User website.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Not available.

Access location(s) An encrypted file is shared with approved researchers for access to data that include PII. These data can also be accessed through the Census Bureau's Federal Statistical Research Data Centers (see the Census Bureau Administrative Data Inventory).

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process IMS/PIC uses the head of household's SSN as the primary key for tenant records. PIC-NG, the IMS/PIC modernization effort, will replace the SSN with a unique identifier as the primary key linking information. Physical inventory can be matched by physical address or by a combination of housing authority-designated development/building/unit numbers.[13]

What identifiers are needed for matching? SSN of head of household for tenant data; physical address for housing data.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Aggregated tenant or housing stock data are available immediately — requests for data sets that include PII need to be vetted by HUD's Privacy Office and, if deemed appropriate, will be provided in an encrypted file.

Other restrictions for matching and working with data None.

Cost

Is there a fee associated with accessing the data? No.

Documentation

➤ Department of Housing and Urban Development resources on Inventory Management System (IMS)/PIH Information Center (PIC)
➤ HUD-50058, Family Report, and MTW-50058, Family Report
➤ Census Bureau Administrative Data Inventory

---

[13]Housing authorities are autonomous, not-for-profit corporations that are usually charted under state law. Housing authorities usually work with government agencies to manage public housing. More information can be found at HUD's website on public housing.

Medicaid Analytic eXtract (MAX)

Overview

Content/domain Medicaid and the Children's Health Insurance Program (CHIP) enrollment and claims.[14]

Ownership/authorizing agency Centers for Medicare and Medicaid Services (CMS), Office of Information Products and Data Analytics (OIPDA).

Data Availability

Data element categories The following files are available from MAX:

- Inpatient File (claim file): Diagnoses, procedures, discharge status, length of stay, and payment amount
- Long Term Care File (claim file): Facility type, dates of service, and discharge status; contains diagnosis codes but does not contain procedure codes
- Other Therapy File (claim file): Diagnosis codes, procedure codes, and date of service
- Prescription Drug File: Drugs, supplies, and other items provided by pharmacies
- Personal Summary File (enrollment file): Demographic data (for example, date of birth, gender, race), basis of eligibility, maintenance assistance status, monthly enrollment status, and use summary

Geographic coverage National.

Population Medicaid and CHIP enrollees in all 50 states and Washington, DC.

Time period coverage Varies by state and file. Data for most states are available from 1999 to 2014 as of March 2019. More information on the time coverage of the individual MAX files is available:

- Inpatient File
- Long Term Care File
- Other Therapy Services File
- Prescription Drug File
- Personal Summary File

Periodicity Not available.

Lag The MAX data files include information reported by the states to CMS via the Transformed-Medicaid Statistical Information System (T-MSIS).[15] Refer to documentation on relevant data file (linked below) for file-specific lag information. There is a substantial time lag to getting MAX files. This largely depends

---

[14]The amount of CHIP enrollment info is very limited in most years of MAX. The MAX files — generated primarily for research purposes — are derived from information collected in T-MSIS. T-MSIS may contain more CHIP data.
[15]States report data to CMS through T-MSIS. T-MSIS data must be transformed to be useful for research purposes. More information on the differences between MAX and T-MSIS can be found here: Frequently Asked Questions regarding MAX.

on how quickly states submit the necessary information. As of March 2019, data from 2014 were available, but only 17 states had submitted their information.

Data retention, update, and deletion schedule(s) Not available.

Data Access

Data acquisition process Researchers submit a draft request packet, which includes a description of the research being conducted along with other materials, to the Research Data Assistance Center (ResDAC). Within 4 to 6 weeks, the ResDAC team reviews the packet and returns it to researchers with edits. Researchers then complete edits and return the final signed documents to ResDAC. Upon reception of the final packet, ResDAC submits it to the CMS Privacy Board. After deliberation, CMS notifies researchers of the Privacy Board's decision. If approved, researchers submit the payment for data, at which point the data will be processed. A flowchart detailing the process can be found here: CMS Research Identifiable Request Process & Timeline.

Public data are available for purchase. More information can be found at CMS: Non-Identifiable Data Files.

Are there restrictions on who can access the data? Yes. Only organizations conducting research and receiving approval through the CMS Privacy Board may access the data.

How long does the data acquisition process take? A minimum of 5 to 6 months. Approval from the CMS Privacy Board can take 6 to 8 weeks, and data processing can take 4 to 6 weeks.

Is a data-sharing agreement or memorandum of understanding required? Yes. Limited data sets and identifiable data files require data use agreements.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes. CMS requires all research-identifiable requests to be approved by an IRB, and IRB documentation must be submitted when applying for identifiable data. Requirements for IRB Review and HIPAA Waiver Documentation describes the requirements and presents examples of acceptable IRB documentation.

Access location(s) Researchers approved to access restricted data can access the data physically; data are mailed to the users via external media. Access through the CMS Virtual Research Data Center (VRDC), hosted by the Chronic Conditions Warehouse (CCW), is also possible. Federal Statistical Research Data Centers (FSRDCs) also retain Medicare data for NCHS survey respondents. More information on working on-site can be found here: National Center for Health Statistics: On site at a Federal Statistical RDC.

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process Researchers may request only the minimum data necessary to complete their study project. Please review the requirements for requesting MAX data from CMS.

To link with outside data sources, the linkage must be described in the request. Per Section 10 of the CMS Data Use Agreement (DUA), the researcher may not attempt to link records without prior CMS approval.

What identifiers are needed for matching? Request files used for matching should contain at least one of the following:

- Beneficiary IDs
- Health Insurance Claim (HIC) numbers
- SSN
- MSIS ID
- National Provider Identifiers (2009 and after)

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Not available.

Other restrictions for matching and working with data CMS allows researchers to retain data for 1 year on all data use agreements. After that time, researchers must request an extension to continue working with the data beyond the expiration date or destroy the data.

Any identifiable data files must be reviewed by ResDAC (data access intermediary) and approved by the CMS Privacy Board.

Cost

Is there a fee associated with accessing the data? Yes. The fees vary depending on how data are accessed, either physically or through the VRDC. See more details on cost information here: CMS Fee Information for CMS Research Identifiable Data.

Documentation

- ResDAC: Research Identifiable Files (RIF) Requests
- FAQ on MAX
- Guide to MAX Data
- Details on individual MAX files:
  - Inpatient File
  - Long Term Care File
  - Other Therapy File
  - Prescription Drug File
  - Personal Summary File
- Information on requesting data from the Research Data Assistance Center
- Requirements for Institutional Review Board (IRB) Review and HIPAA Waiver Documentation for RIF DUA Request Submissions
- CMS Research Identifiable Request Process & Timeline
- On site at a Federal Statistical RDC
- Public data available for purchase

Medicare Master Beneficiary Summary File (MBSF)

Overview

Content/domain Medicare beneficiary enrollment information.

Ownership/authorizing agency Centers for Medicare and Medicaid Services (CMS), Office of Information Products and Data Analytics (OIPDA).

Data Availability

Data element categories Demographics (for example, age, gender) and Medicare enrollment.

Geographic coverage National.

Population Medicare beneficiaries.

Time period coverage 1999-2018.

Periodicity Annual.

Lag Annual files have a 1-year lag and are available in February. Quarterly files are available on a 5- to 6-month lag.

Data retention, update, and deletion schedule(s) Data are updated annually.

Data Access

Data acquisition process Researchers submit a draft request packet, which includes a description of the research being conducted along with other materials, to the Research Data Assistance Center (ResDAC). Within 4 to 6 weeks, the ResDAC team reviews the packet and returns it to researchers with edits. Researchers complete edits and return the final signed documents to ResDAC. Upon reception of the final packet, ResDAC submits it to the CMS Privacy Board. After deliberation, CMS notifies researchers of the Privacy Board's decision. If approved, researchers submit the payment for data, at which point the data are processed. A flowchart detailing the process can be found here: CMS Research Identifiable Request Process & Timeline.

Public data are available for purchase. More info can be found here: CMS Non-Identifiable Data Files.

Are there restrictions on who can access the data? Yes. Only organizations conducting research and receiving approval through the CMS Privacy Board may access the data.

How long does the data acquisition process take? A minimum of 5 to 6 months. Approval from the CMS Privacy Board can take 6 to 8 weeks, and data processing can take 4 to 6 weeks.

Is a data-sharing agreement or memorandum of understanding required? Yes. Limited data sets and identifiable data files require data use agreements.

Do researchers need to provide evidence of an IRB review to access these data for research purposes?
Yes. CMS requires all research-identifiable requests be approved by an IRB, and IRB documentation must be submitted when applying for identifiable data. Requirements for Institutional Review Board (IRB) Review and HIPAA Waiver Documentation describes the requirements and presents examples of acceptable IRB documentation. Limited data set requests do not require an IRB review.

Access location(s) Researchers approved to access restricted data can access the data physically; data are mailed to the users via external media. Access through the CMS Virtual Research Data Center (VRDC), hosted by the Chronic Conditions Warehouse (CCW), is also possible. FSRDCs also retain Medicare data for NCHS survey respondents. More information on working on-site can be found here: NCHS: On site at a Federal Statistical RDC.

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process Researchers may request only the minimum data necessary to complete their study project. Please review the requirements for requesting Medicare data from CMS, under "Define and estimate the study cohort."

To link with outside data sources, the linkage must be described in the request. Per Section 10 of the CMS Data Use Agreement (DUA), the researcher may not attempt to link records without prior CMS approval.

What identifiers are needed for matching? Request files used for matching should contain at least one of the following:

- Beneficiary ID
- Health Insurance Claim (HIC) number
- SSN
- Last name or partial SSN or HIC

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Not available.

Other restrictions for matching and working with data CMS allows researchers to retain data for 1 year on all data use agreements. After that time, researchers must request an extension to continue working with the data beyond the expiration date or destroy the data.

The identifiable data file also must be reviewed by ResDAC (data access intermediary) and approved by the CMS Privacy Board.

Cost

Is there a fee associated with accessing the data? Yes. The fees vary depending on how data are accessed, either physically or through the VRDC. More information on cost can be found here: CMS Fee Information for CMS Research Identifiable Data.

Documentation

- ➢ [CMS Research Identifiable Request Process & Timeline](#)
- ➢ [Information on requesting data from the Research Data Assistance Center](#)
- ➢ [On site at a Federal Statistical RDC](#)
- ➢ [Requirements for Institutional Review Board (IRB) Review and HIPAA Waiver Documentation for RIF DUA Request Submissions](#)
- ➢ [Public data available for purchase](#)

National Death Index (NDI)

Overview

Content/domain Fact of death and cause(s) of death records.

Ownership/authorizing agency The Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Division of Vital Statistics, U.S. Department of Health and Human Services. Records are owned by the states and territories of the United States.

Data Availability

Data element categories Date of death, state of death, death certificate number, and 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.

Geographic coverage National (reported at state level).

Population All deaths in the United States that are reported to state vital statistics' offices.

Time period coverage 1979-2016, and a preliminary file for 2017.

Periodicity New death records are added twice a year.

Lag 10 months.

Data retention, update, and deletion schedule(s) Data are updated twice a year.

Data Access

Data acquisition process To access the data, researchers must submit an application form to the CDC. More details on the process and contact information for NDI staff members are available at How to Use the National Death Index: Steps in the Process.

Are there restrictions on who can access the data? Data may be used only for public health and medical research purposes. No administrative uses are permitted.

How long does the data acquisition process take? 2 to 3 months for applications to be reviewed and approved.

Is a data-sharing agreement or memorandum of understanding required? All applicants must have a signed confidentiality agreement, and supplemental confidentiality agreements if there are collaborators and/or financial sponsors.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes, research projects must be approved by an IRB before the application process.

Access location(s) Primary access is through the National Center for Health Statistics, Division of Vital Statistics, National Death Index. Arrangements may be made, under very special circumstances, through several Research Data Centers located at select federal facilities.

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process Researchers must send a compact disc (CD) with identifiers by overnight mail. Use of Secure File Transfer Protocol (SFTP) is also permitted. More detailed information on the process can be found in Chapter 2 of the National Death Index User's Guide.

What identifiers are needed for matching? Request files must have one of the following combinations of identifiers to avoid automatic rejection of their records:

1. First name, last name, and month and year of birth
2. First name, last name, and SSN
3. SSN, date of birth, and sex

Additional requirements are needed to identify a match. Please see the National Death Index User's Guide.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Only state of death, date of death, and the death certificate number are returned. Causes of death are available for an additional fee. No other information is provided.

Other restrictions for matching and working with data Data must be used only for public health and medical research. In part due to recognition of the social determinants of health, applications for the use of data for social policy research may be approved. No administrative uses are permitted.

Any data files with individual identifiers that come from an NDI match must be destroyed after 5 years unless otherwise approved.

More information on restrictions is available on page vii of the National Death Index User's Guide.

Cost

Is there a fee associated with accessing the data? Yes. For matches that include only death records, there is a $350.00 service charge plus $0.15 per record per year searched. For matches to the death codes, fees are $0.21 per record per year searched. These fees are only for UNKNOWN searches.[16] Other fees will apply to other types of searches. More information on fees is available on the NCHS user fees worksheet.

Documentation

➢ National Death Index User's Guide
➢ National Death Index Application Form
➢ National Death Index User Fees
➢ Poverty Action Lab: National Death Index

---

[16]UNKNOWN searches refer to those for which the researcher does not know the vital status of the record and for which different years of death will need to be searched.

National Directory of New Hires (NDNH)

Overview

Content/domain Employment, earnings, and unemployment insurance (UI) benefits.

Ownership/funding agency Office of Child Support Enforcement (OCSE), Administration for Children and Families (ACF), U.S. Department of Health and Human Services (HHS).

Data Availability

Data element categories Data are available in three different files. The New Hire File contains information on all newly hired employees as reported by employers to each State Directory of New Hires (SDNH). The Quarterly Wage (QW) File contains QW information on individual employees from state workforce agency (SWA) and federal agency records. The Unemployment Insurance (UI) File contains UI information on individuals who received or applied for unemployment benefits, as reported by SWAs. See page 2 of A Guide to the National Directory of New Hires for details on data elements contained within each file.[17]

Geographic coverage National.

Population Workers covered by unemployment insurance and federal workers.[18]

Time period coverage The National Directory of New Hires (NDNH) contains approximately 24 months (approximately 8 quarters) of data at any given time.

Periodicity Quarterly.

Lag Varies by data set.

Data retention, update, and deletion schedule(s) NDNH data are deleted on a monthly basis (on the first Saturday after the first Friday of each month). New hire information is submitted within 20 days of hire. The NDNH is updated daily. State workforce agencies transmit QW data to the NDNH within 4 months of the end of each calendar quarter. Federal agencies transmit QW data to the NDNH within 1 month after the end of a calendar quarter.

Data Access

Data acquisition process A research request must be submitted by a sponsoring federal or state agency to OCSE for approval. It must include a justification that the data will be used "to conduct research found by the Secretary of Health and Human Services (HHS) to be likely to contribute to achieving the purposes of part A or part D of the Social Security Act."[19] The purposes in part A relate to Temporary

---

[17]The interpretation of the variables and how the data are entered are uniform. For select states, OCSE assesses the collection of the data during an annual quality assurance audit.

[18]Currently, independent contractors are not required to report employment and wage information to the state directories of new hires. When NDNH receives the data, the names and SSNs are verified, and edits are implemented, if needed.

[19]Federal Parent Locator Service, "A Guide to the National Directory of New Hires" (Administration for Children and Families, Office of Child Support Enforcement, March 10, 2017), p. 5.

Assistance to Needy Families (TANF), and the purposes in part D relate to child support. If the request is approved, this authorizes OCSE to retrieve personal identifiers to begin retaining the data for research purposes (to prevent it from being deleted). Meanwhile, a memorandum of understanding (MOU) between the requesting federal agency (including its research partners) and OCSE is developed and will include detailed information about the specifications for the input (including PII and data to be passed through to the output) and output (matched) data. Once the MOU is executed, the data are released to the requesting agency.

Are there restrictions on who can access the data? Yes. Most often, OCSE transmits deidentified data to the federal or state agency that requests the match. These files can then be used by researchers who are assisting with conducting the research with the agency.

How long does the data acquisition process take? 6 to 9 months.

Is a data-sharing agreement or memorandum of understanding required? Yes, an MOU is required.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? No. However, IRB materials should be available upon request.

Access location(s) Data can be accessed in a few different ways. For example, data-requesting agencies store the matched data from OCSE on their servers and provide their research partners access to the relevant folders to analyze the data via a secure virtual portal. Another method for accessing the data involves the requesting agency saving copies of the matched data to a stand-alone laptop that research partners can use in a clean room environment. Disaggregated versions of the data cannot leave the requesting agency's server.

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process For more information, see the [National Directory of New Hires Research Request Form.](#)

What identifiers are needed for matching? Most matches are done by name and SSN.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Only the data elements listed on page 14 of the [National Directory of New Hires Research Request Form](#) will be returned.

Other restrictions for matching and working with data OCSE does not return personal identifiers on the output data and prohibits reidentification.[20]

Cost

Is there a fee associated with accessing the data? Yes. **"**OCSE uses a standard methodology to calculate fees based on three components: (1) Access (a fee which is split evenly among

---

[20]Public-use NDNH data without personal identifiers may be released for research purposes if the DHHS Secretary deems that those purposes would contribute to the purposes of the TANF Program and Child Support Enforcement. More information on the purpose of the TANF program can be found in [Block Grants to States for Temporary Assistance for Needy Families: Purpose](#). The NDNH is a restricted-use data set. OCSE cannot disclose NDNH information if the law does not authorize an agency to receive specified NDNH information or if the comparison being requested does not meet statutory restrictions.

NDNH users), (2) Frequency of matches, or (3) User-specific costs related to performing the match."[21]

Documentation

➢ Administration for Children and Families, Office of Child Support Enforcement: A Guide to the National Directory of New Hires
➢ Administration for Children and Families, Office of Child Support Enforcement: National Directory of New Hires: Guide for Data Submission
➢ National Directory of New Hires Research Request form
➢ Information for technical support liaisons: OCSE National Directory of New Hires Contacts
➢ TANF Purpose
➢ FY 2016 Preliminary Report to Congress (see page 101)
➢ ACF Administrative and Survey Data Resources Compendium[22]

---

[21]Federal Parent Locator Service, p. 8.
[22]ACF expects to release a public version of this report in 2019.

National Student Clearinghouse StudentTracker

Overview

Content/domain Student enrollment and graduation at postsecondary institutions.

Ownership/authorizing agency National Student Clearinghouse (NSC).

Data Availability

Data element categories The StudentTracker data include:

- Student demographics
- Time periods of college enrollment at different institutions, including level of enrollment (full-time, half-time, etc.)
- Degrees received from different colleges, including date of receipt and (in some cases) type of degree received (associate, certificate, bachelor, master, etc.)

Geographic coverage National.

Population Students at Title IV, degree-granting institutions. This is approximately 98 percent of all students at public and private institutions (about 84 percent of such institutions report data to the National Student Clearinghouse).

There is less coverage of:

- Enrollments in certain geographic regions (for example, Louisiana, West Virginia) and private/nonprofit colleges
- Degrees from private/nonprofit institutions

Time period coverage 1993 to present.[23]

Periodicity The data are term-based, but colleges report data to the NSC, and the NSC updates data throughout the year.[24] Most institutions report to the NSC three to four times per term, most often around "events" (for example, start of the term, close of the add/drop period, end of the term).

Lag Approximately 45 days after an event as described above, though this varies by college.

Data retention, update, and deletion schedule(s) Data are retained for the duration of the institution's participation with the NSC.

---

[23]Data from some institutions may cover broader time periods for some data types. In rare circumstances, some degree records range back to the 1950s.

[24]Term-based is defined as the time period in which classes at an educational institution are in session.

Data Access

Data acquisition process Researchers contact the NSC with a description of their project and how they are planning to use the data.

Are there restrictions on who can access the data? Yes. Researchers must be affiliated with an approved organization or institution.

How long does the data acquisition process take? The NSC responds to initial requests within 3 business days.

Is a data-sharing agreement or memorandum of understanding required? Yes. All data exchange takes place after the negotiation and execution of an agreement between the NSC and the researcher.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? No, IRB review is not always required.

Access location(s) Data exchange takes place through the National Student Clearinghouse's secure File Transfer Protocol (FTP) server.

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process Researchers submit a finder file that contains the identifiers for their sample. The National Student Clearinghouse then uses this file to match with its records.

What identifiers are needed for matching? First name, last name, and date of birth are required for matching. It is also possible to match on SSN, but consent from the student is required to do so. Middle name or initial is an optional element.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Restrictions vary by type of researcher and research agenda.

Other restrictions for matching and working with data Some students' records are "blocked" (under the Family Educational Rights and Privacy Act [FERPA]), meaning that they can be acquired only if active consent is given by the student.[25] The consent form must explicitly mention (1) requesting and receiving records from the NSC and (2) using SSNs to match and retrieve those records.

Cost

Is there a fee associated with accessing the data? Yes. The fee varies depending on who is accessing the data and for what purpose. Usually, researchers are charged based on the number of students for whom they are requesting data. There is a minimum fee of $425 per file to cover operational administrative costs. There is also a $500 setup fee. More information on fees can be found in Appendix B of The National Student as an Integral Part of the National Postsecondary Data Infrastructure.

---

[25]See Family Educational Rights and Privacy Act (FERPA).

Documentation

➢ Poverty Action Lab: National Student Clearinghouse StudentTracker Data
➢ National Student Clearinghouse: Reading the StudentTracker Detail Report
➢ The National Student Clearinghouse as an Integral Part of the National Postsecondary Data Infrastructure
➢ General information on the Family Educational Rights and Privacy Act (FERPA)

Social Security Administration (SSA)

- Master Beneficiary Record (MBR)
- Master Earnings (ME)
- Numerical Identification (NUMIDENT)
- Supplemental Security Record (SSR)
- Death Master File (DMF)

Overview

Content/domain Applicants, recipients, and beneficiaries of Supplemental Security Income (SSI) and Old Age, Survivor's, and Disability Insurance (OASDI).

Ownership/authorizing agency U.S. Social Security Administration.

Data Availability

Data element categories SSA maintains a [system of records](#) — or a compilation of various types of administrative data files that include personal identifiers — and the files commonly used for research include:

- MBR: Information on individuals applying for and receiving OASDI benefits; includes detailed benefit payment information, summary earnings data, as well as SSNs for the primary worker and the beneficiary
- ME: Individuals' lifetime records of wages and self-employment earnings, including annual earnings used for OASDI contributions
- NUMIDENT: Information on individuals applying for SSNs; includes name, date and place of birth, parents' names, and date of death
- SSR: Information on individuals applying for and receiving SSI benefits; includes SSI benefit amounts received, date of claim, citizenship status, income, resources, eligibility code, payment code, and payment amount
- DMF: Information on deceased individuals, including SSN, first name, middle name, surname, date of birth, and date of death

Geographic coverage National.

Population

- MBR: Recipients of OASDI
- ME: U.S.-based wage earners[26]
- NUMIDENT: All individuals with an SSN

---

[26]This includes individuals with a W-2, as well as those who claimed self-employment income on Form 1040 to the Internal Revenue Service (IRS).

- SSR: Recipient of SSI benefits
- DMF: All individuals with an SSN

Time period coverage

- MBR: Varies by data element
- ME: Varies by data element — for example, annual total wages are available from 1978 to present, and annual earnings used for OASDI contributions are available from 1951 to present
- NUMIDENT: Cumulative file built from updates received since original delivery in 1998
- SSR: Varies by data element — all variables are available from at least 1978 to present
- DMF: If available, since 1936

Periodicity

- MBR: Annually
- ME: Annually
- NUMIDENT: Quarterly
- SSR: Annually
- DMF: Weekly

Lag

- MBR: 1 month
- ME: 18 months
- NUMIDENT: 1 month
- SSR: 1 month
- DMF: 1 month

Data retention, update, and deletion schedule(s) Files are retained indefinitely.

Data Access

Data acquisition process Federal, state, or local agencies, tribal organizations, or private entities wishing to enter into a data exchange agreement with the SSA must complete a data exchange request form.

Are there restrictions on who can access the data? Yes. Identified data are typically available only to researchers within the SSA, although there are a few exceptions. For example, the SSA provides PII to the Office of Management and Budget and the Treasury Department.

The SSA also engages in research data exchange agreements in which external researchers work collaboratively with SSA employees. Only SSA employees are permitted to work with the person-level data in these situations. SSA employees will provide aggregate data to outside researchers involved with the project with approval of the SSA's Office of General Counsel.

How long does the data acquisition process take? Approximately 12 to 15 months (for any of the five abovementioned data files).

Is a data-sharing agreement or memorandum of understanding required? Yes.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes.

Access location(s) From the SSA or through the Census Bureau's Federal Statistical Research Data Centers.

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process Not available.

What identifiers are needed for matching? Dates of birth, first and last names, and SSNs.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Yes, but these restrictions may vary by data source. For example, 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.

Other restrictions for matching and working with data 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. Additional information can be found under Additional Statistical Linkages and Services.

Cost

Is there a fee associated with accessing the data? Yes. Costs are based on the volume of records for each data pull (for example, there are separate costs for MBR data and SSR data). Some factors that impact the cost are computer run time, staff time, and resources. Users also pay an agreement startup fee and an annual administrative fee.

Documentation

- ➢ Poverty Action Lab: Social Security Administration Data
- ➢ Social Security Administration: Uses of Administrative Data at the Social Security Administration
- ➢ Social Security Administration: Data Exchange Request Form (DXRF) Request for Information from SSA
- ➢ Social Security Administration: Requesting an Electronic Data Exchange with SSA
- ➢ Social Security Administration's Master Earnings File: Background Information
- ➢ Information on how to request the Death Master File

Temporary Assistance for Needy Families (TANF) — Federal

Overview

Content/domain Demographic, public assistance, and work activity data.

Ownership/funding agency Office of Family Assistance (OFA), Administration for Children and Families (ACF), U.S. Department of Health and Human Services (HHS).

Data Availability

Data element categories Personal identifiers (SSNs, date of birth, case number), demographics (family type, race, ethnicity, etc.), geography (county, state, zip code), benefits received (child care, education, job training, etc.), case status, income, and work activities. See Final TANF & SSP-MOE Data Reporting System Transmission File Layouts and Edits for additional information on the federally-reported TANF data from states and US territories, including the types of data submitted, how these data are organized, and some of the validation checks performed on the data.

Geographic coverage U.S. national coverage, but reported at the state and county levels.

Population TANF and Separate State Programs (SSP) cash assistance families and recipients. Currently 21 states and territories submitted data for only a sample of their families. As of Fiscal Year 2017, 69 of 74 tribal TANF cash assistance programs submitted data on families. While there are many similarities between the Tribal TANF program and the state TANF program, there are also differences in what data are collected, how the data are submitted, and how the data are made available.

Time period coverage October 2008 - September 2018.

Periodicity Collected monthly and reported quarterly.

Lag Data for the fiscal year are available approximately 9 months after the end of the fiscal year.

Data retention, update, and delete schedule(s) States, tribes, and territories are required by law to submit data within 90 days following the end of each quarter. Data are updated (and overwritten) on an ad hoc basis throughout the fiscal year. OFA maintains only the most recent version of data submitted.

Data Access

Data acquisition process Public-use data covering the fiscal year period are available upon request, typically 9 months after the period ends. Data prior to Fiscal Year 2010 are available at TANF Administrative Data: Sample Data Available to the Public; more recent data are usually saved on DVDs, password-protected, and sent to approved researchers.

Requests for restricted-use data that include PII must be submitted to OFA for review. Approved researchers typically need to sign a data-sharing agreement before data can be released. Data can also be accessed at Census Bureau's Federal Statistical Research Data Centers (see page 58).

Are there restrictions on who can access the data? Yes. Only OFA-approved researchers may access restricted-use TANF data.

How long does the data acquisition process take? Requests for unidentifiable data that aren't publicly available typically take 1 to 2 weeks. Restricted-use data requests may take 2 to 4 months.

Is a data-sharing agreement or memorandum of understanding required? Typically, a data-sharing agreement is required for access to restricted-use data. No such agreement is in place for access to public-use versions of the data.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Data-sharing agreements for restricted-use data may require evidence of an IRB review.

Access location(s) Public-use data are available via the Office of the Assistant Secretary for Planning and Evaluation's website, or OFA will send the data to researchers by mail. OFA transmits data with PII to approved researchers via a method that meets FIPS 140-2 security standards. TANF data are also available via the Census Bureau's Federal Statistical Research Data Centers.

Matching and Working with Data

Can a research sample be matched to the data source? OFA typically provides all data covering the requested time period for researchers to conduct the match.

Matching process NA.

What identifiers are needed for matching? SSN, case number, date of birth, and/or state FIPS code.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Additional restrictions may apply and are contingent upon the data-sharing agreement negotiated between OFA and the research entity.

Other restrictions for matching and working with data NA.

Cost

Is there a fee associated with accessing the data? Typically, no.

Documentation

➢ Final TANF & SSP-MOE Data Reporting System Transmission Files Layouts and Edits
➢ Aggregate data reports on TANF and SSP-MOE caseloads, work participation rates, and recipient characteristics: Data Publications
➢ Bibliography of studies using TANF-linked administrative data
➢ TANF sample data available to the public
➢ ACF Administrative and Survey Data Resources Compendium[27]

---

[27]ACF will be releasing a compendium of administrative and survey data resources to the public at a future date.

Part 2: State Administrative Data Sources

This section features data sources with state-level coverage. States collect and maintain several types of administrative data, and in many cases, state-level data are the original sources of federal- and national-level data. Because of this, the process for obtaining access to up-to-date state data may be easier, especially since state data also tend to be updated more frequently than national- or federal-level data.

State data also have limitations. For example, state data capture information on activities occurring within a given state (not across states), so, for example, if a person obtains a job in another state, his or her employment and earnings information would not be captured in the original state's unemployment insurance data. This is especially relevant when considering whether or not to match a sample of individuals whose activities are likely to occur out of state. Additionally, if a research study includes sites or sample members in various states, researchers would have to complete a separate data acquisition process for each individual state. This can be time consuming and require additional resources.

Data information was compiled separately for unemployment insurance, Temporary Assistance for Needy Families (TANF), the Supplemental Nutrition Assistance Program (SNAP), State Vital Statistics, and Statewide Longitudinal Data Systems (SLDS) since data content across states is similar, though the acquisition and matching procedures are widely variable. For example, with few exceptions, states capture the same information on employment and earnings, but the process for obtaining access to these data tends to be contingent on state-level policies and regulations. Also, Early Childhood Integrated Data Systems (ECIDS) are a relatively new source of information, and states vary substantially in terms of readiness to support collection and sharing of these data. The state example represented herein currently considers matching requests for research, though many more states are still developing these systems or have developed systems but do not currently consider these requests.

The information in this section is intended primarily to document data availability. It is important to connect with individual states for more specific guidance on access and matching procedures.

Glossary — State Data Sources

Overview

Content/domain A general description of the domains included in the data source — for example, demographic, public assistance, employment, or health data.

Ownership/authorizing agency The organization or agency that has the authority to provide access to the data for research purposes. This may not be the organization or agency that originally collected the data.

Data Availability

Data element categories The types or categories of data available — for example, personally identifiable information (PII), geographic information, or quarterly wages.

Geographic coverage Geographic location(s) the data cover — for example, state- or county-based data.

Population Who is covered by the data source — for example, all Temporary Assistance for Needy Families cash assistance recipients or all individuals with an assigned Social Security Number (SSN).

Time period coverage The time period (years or months) for which data from each data set were available as of fall 2018. If data are collected only for parts of a year, a description may be provided.

Periodicity How often the data are collected and/or in what time interval the data elements are aggregated — for example, annually or quarterly; Supplemental Nutrition Assistance Program payments are reported monthly.

Lag The amount of time between when the data are collected and when they are available to researchers (including other state or federal agencies and contractors). For example, if wage records for Quarter 1, 2014, are available to researchers in Quarter 3, 2014, the lag is two quarters.

Data retention, update, and deletion schedule(s) The time period for which data are retained before purging or archiving occurs, how often the data are updated, and/or on what schedule data are deleted (if applicable) — for example, data may be retained for 24 months before purging and updated on a quarterly basis.

Data Access

Data acquisition process Description of how researchers request access to the data source, including a summary of the process to obtain approval and access — for example, a research request might need to be submitted for approval, evidence of approval by an institutional review board (IRB) or study participant consent (and/or waiver of confidentiality) might be required, or a background check might be conducted.

Are there restrictions on who can access the data? Yes/no answer to whether there are any restrictions on who can apply for access to the data for research purposes. If yes, a summary of the restrictions may be provided — for example, all third-party researchers might be eligible; access may be limited to (local, state, or federal) government contractors or research partners; or those wanting to access the data might need to sign a nondisclosure agreement.

How long does the data acquisition process take? A (rough) timeframe for how long the data acquisition process usually takes from the point when researchers first apply for access until the point when they are able to work with the data. If applicable, a timeframe may be given for each step in the acquisition process — for example, researchers might need to obtain Special Sworn Status before data can be accessed, a process that can take 4 to 6 weeks.

Is a data-sharing agreement or memorandum of understanding required? Yes/no answer to whether a data-sharing agreement of memorandum of understanding is required to access the data center. If yes, a description of the requirements may be provided.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes/no answer to whether evidence of an IRB review is required to permit access to the data. If yes, a description of the requirements may be provided.

Access location(s) The physical or virtual locations where approved researchers may access the data, including any data centers that house the data source. If there are physical locations, a summary of how many there are and where they are located may be provided. If the data can be transmitted securely to researchers (for example, via Secure File Transfer Protocol, or SFTP), this information may be provided.

Matching and Working with Data

Can a research sample be matched to the data source? Yes/no answer to whether researchers can submit a file with personal identifying information to the authorizing agency — for example, researchers may be able to send a list of SSNs for 1,000 study sample members, and that list is used to match to the data source.

Matching process Summary of the process by which researchers initiate a match to the data source after they have been granted approval to work with the data. This may include a list of the steps in the matching process with links to any online resources that describe the steps, as appropriate — for example, how researchers submit a data file and how the data are linked.

What identifiers are needed for matching? A list of the person-level identifiers required to link/match a research sample to the data available through the data center (for example, SSNs). If a data center allows for fuzzy matching, a list of the fields — such as name, date of birth, and phone number — used in the algorithm may be provided.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Yes/no answer to whether there are any restrictions on the data file that is returned to researchers. If yes, a list of the restrictions may be provided. For example, researchers may be allowed to work only with deidentified data or aggregate data.

Other restrictions for matching and working with data Summary of the limitations around accessing, matching, or using the data. These could be limitations around specific data sets or elements — for example, perhaps researchers must work with the data on-site, or researchers are given access only to deidentified data.

Cost

Is there a fee associated with accessing the data? Yes/no answer to whether researchers are charged a fee to access the data. If yes, an overview of the fee structure may be provided.

Documentation

Links to any publicly available documentation. These sources could include how to apply for access to the data, standard formats or record layouts of input and output files, examples of research publications that feature analysis of matched data, or other relevant information.

Early Childhood Integrated Data Systems (ECIDS)

Overview

An ECIDS is a data system that houses and links detailed information from a variety of early learning services and programs — including data related to children and families served by the programs, members of the workforce, and the characteristics of the program or services — offered across the state. Several states have implemented an ECIDS. For this Compendium, information was collected on Pennsylvania's ECIDS, which currently considers matching requests, to better understand the kind of data collected and the process for obtaining access to these data for research purposes. Pennsylvania's ECIDS is called Pennsylvania's Enterprise to Link Information for Children Across Networks (PELICAN). When possible, information that is generally applicable across existing ECIDS is included.

Content/domain Integrated early childhood program data (child care, education, health, and social services).

Ownership/authorizing agency State education, early childhood, social services, and health agencies.

Data Availability

Data element categories The exact data elements vary by state. For PELICAN, the data system includes demographic data and data on comprehensive care, education, and workforce services received by children and/or their family members.

Geographic coverage State level.

Population Generally, children participating in early education and health social service programs. Some ECIDS contain data from programs that have information on the families of participating children and participating service programs.

Time period coverage Varies by state. For PELICAN, coverage varies by information system, with the earliest starting in 2002.

Periodicity Varies by state and — in some cases — by service program.

Lag Varies by state. PELICAN's lag times range from 1 to 3 (or more) months.

Data retention, update, and deletion schedule(s) Varies by state. For PELICAN, data in the different systems can be updated daily depending on programmatic needs and requirements. As of fall 2018, the systems did not have a formal archival process; most data still exist from the initial implementation. Some systems have implemented purges of correspondence with the possibility to recreate if necessary.

Data access

Data acquisition process Varies by state. Public-use data can often be found on the state ECIDS websites. For restricted data, some states have a formal data request process; others are currently developing one.

For PELICAN, individuals/entities requesting data must complete a data request form and submit it to the Office of Child Development and Early Learning (OCDEL). Depending on the type and granularity of the data being requested, a data-sharing agreement may be required. The agreement may require Department of Health Services (DHS) legal and/or OCDEL executive-level review/approval.

Are there restrictions on who can access the data? Varies by state. In some states, it is not yet possible to release data to researchers because the data request policy is still under development.

PELICAN's information management systems are confidential, and only authorized persons will have access to the records, per federal and state confidentiality, privacy, and security laws. Identified data are available to researchers only with DHS legal approval. Aggregated data may need to be suppressed depending on the type of data requested and the geographic breakdowns.

How long does the data acquisition process take? Varies by state. For PELICAN, the process may take between 3 weeks and 3 months.

Is a data-sharing agreement or memorandum of understanding required? Varies by state. For PELICAN, yes.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Varies by state. For PELICAN, yes.

Access location(s) Varies by state, but ECIDS can often be accessed from state websites (public use) or by a secure file transmission (restricted use).

Matching and Working with Data

Can a research sample be matched to the data source? Varies by state. For PELICAN, yes, although OCDEL programs do not require individuals to provide an SSN, which may affect the match rate.

Matching process Varies by state. For PELICAN, OCDEL will typically provide a data file; it is up to the requesting/receiving entity to identify/perform a match.

What identifiers are needed for matching? Varies by state. For PELICAN, identifiers used for matching depend on if the participant is an individual or a service provider. Individuals can be matched using name, date of birth, and other demographic information. Service providers can be matched using name, location, or federal tax identification number.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Varies by state.

Other restrictions for matching and working with data Varies by state. For PELICAN, data files with personally identifiable information require an agreement to be signed by the requestor and at least OCDEL, with the possibility of departmental legal counsel approval. Any required data-sharing agreements may outline other restrictions and/or requirements for each individual request.

Cost

Is there a fee associated with accessing the data? Varies by state. For PELICAN, there is typically no fee.

Documentation

- The Early Childhood Data Collaborative's 2018 State of State Early Childhood Data Systems
- The Early Childhood Data Collaborative
- What Is an ECIDS Overview
- Pennsylvania Early Learning Dashboards

Statewide Longitudinal Data Systems (SLDS)

Overview

SLDS store and maintain longitudinal K-12 education data, and many also include early childhood, postsecondary, and/or workforce data. Every state has an SLDS, but the completeness, content, and coverage vary. Most of the information presented here draws from the SLDS Project, including some information from a few specific states' profiles to provide an overview of the kinds of data stored in SLDS, the range of data acquisition processes in place, and for states where matching is possible, the kinds of mechanisms used to link data across different agencies and systems. The information below is not comprehensive, but rather is meant to give an idea of the potential factors involved with acquiring and matching to SLDS data. More information on each state's SLDS can be found in the individual state profiles or by contacting the state directly.

Content/domain Education data (mainly K-12).

Ownership/authorizing agency State educational agencies, Institute of Education Sciences (IES).

Data Availability

Data element categories The content of SLDS varies by state, and the systems often include a range of different data types. Many states link their K-12 education data with early childhood education, postsecondary, and/or workforce data. Some states also link data with career and technical education, adult education, and social service data. Some common data elements include:

- Student, school, and district identifiers
- Student demographics
- Enrollment and attendance
- Completion and withdrawal
- Behavioral data
- State and local assessment data
- Courses, grades, and credit attainment
- College Board and ACT test data
- Staff data, including teacher data

More information on common data elements can be found on page 5 of A Guide to Using State Longitudinal Data for Applied Research. Information on the content of specific state systems can be found in Appendix B of the same guide.

Geographic coverage State level.

Population Varies by state. For example:

- California: The California Longitudinal Pupil Achievement Data System (CALPADS) holds data on every K-12 student in the California public education system.[28]
- Florida: The Florida Statewide Longitudinal Data System (FSLDS) has data on students and staff members from every P-12[29] education institution in Florida, as well as higher-education data records from every public community college and state university in Florida.[30]

Time period coverage Varies by state.

Periodicity Varies by state. For example, in Colorado, local education agencies are required to submit their data records six times during the academic year.[31]

Lag Varies by state.

Data retention, update, and deletion schedule(s) Varies by state.

Data Access

Data acquisition process Many but not all states use a formal research request process. Profiles and evaluations of each state's SLDS can be found in the SLDS Project's State Profiles. Those interested in finding state-specific information on the acquisition process for researchers beyond the examples given should refer to the "Research Accommodations" section of each state's page. Examples include:

- Virginia: Researchers interested in data in the Virginia Longitudinal Data System (VLDS) must work with a sponsoring agency to develop research questions consistent with the agency's research agenda. After receiving preliminary approval, researchers must submit a formal research request to the sponsoring agency. This request must provide a description of the project, a nondisclosure agreement signed by each researcher on the project, and the select data fields needed for the research. The agreement must be approved and signed by each agency from which data are being requested. Once this process is complete, researchers will be provided access to execute the research query. After the research is conducted, researchers must submit the findings to the sponsoring agency for review.[32]
- Kentucky: The Kentucky Longitudinal Data System (KLDS) has public-use data files. If the specific data being requested are not available through standard reports or the public-use files, interested researchers must fill out a Kentucky Center for Statistics (KYStats) data request form requesting either individual-level data or aggregate data. Deidentified data sets may be provided by the Kentucky Center for Education and Workforce Statistics (KCEWS) if it is determined that publicly accessible data files do not fully address the research question identified by the requesting party and KCEWS has reviewed the risk of reidentification.[33]

Are there restrictions on who can access the data? Varies by state.

How long does the data acquisition process take? 4 to 8 months, on average, but it can take as long as 1 to 2 years.

---

[28]SLDS Project: California profile.
[29]P-12 is defined as Preschool to 12th Grade
[30]SLDS Project: Florida profile.
[31]SLDS Project: Colorado profile.
[32]SLDS Project: Virginia profile. More information on the research process can be found under "Research Process" at the Insights VLDS page.
[33]SLDS Project: Kentucky profile.

- Kentucky: Data request and agreement can take up to 4 months to complete.

Is a data-sharing agreement or memorandum of understanding required? Yes, in some states.

- Kentucky: Yes. A Data and Information Sharing Agreement is required for any individual-level, deidentified data requests to ensure data security and privacy in accordance with state and federal laws.
- Virginia: Yes. When researchers submit a formal research request, each researcher on the project will need to submit a nondisclosure agreement along with data fields they'd like to receive. If data are requested from multiple agencies, each agency will need to approve and sign the agreement.
- Colorado: Yes. All requirements outlined in the Colorado Department of Education's (CDE) Research Data Sharing Agreement Template must be complied with. More information can be found at CDE Data Requests under PII Data Requests.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Some states require an IRB review of all research projects.

Access location(s) Varies by state.

Matching and Working with Data

Can a research sample be matched to the data source? Varies by state.

- Kentucky: Most likely yes. Data from interested researchers can be matched with data in the Kentucky Longitudinal Data System (KLDS) to measure program outcomes or help answer other research questions.
- Colorado: Yes. Researchers interested in linking must provide detailed information to CDE outlining the data being linked and other sources of data.

Matching process Varies by state.

- Kentucky: Interested researchers must notify the Kentucky Center for Statistics (KYStats) that they are interested in matching data to KLDS when they submit their data request. A staff member in KYStats will contact the researcher so the data can be sent to KYStats via a secure FTP server.
- Arkansas: To begin the matching process, interested researchers must contact the Arkansas Research Center (ARC) for more information.

What identifiers are needed for matching? Varies by state.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Varies by state. States balance the sharing of student data against the legal requirements under the Family Educational Rights and Privacy Act (FERPA). Below are examples of the types of data returned to researchers in two states:

- Kentucky: Kentucky state law states that no personally identifiable information can be released. Deidentified, individual-level data are available to researchers.[34]

---

[34]See Kentucky Center for Statistics Data Request and FAQs for more information.

- Arkansas: The Arkansas Research Center returns deidentified data sets to individual researchers, partner agencies, and graduate students who are conducting education-related research with regard to students in Arkansas.[35]

Other restrictions for matching and working with data States are not required to share their data; requests to access data are up to each state's discretion.

Cost

Is there a fee associated with accessing the data? Some states charge fees for accessing their data. The majority of states do not have fees.

Documentation

➢ National Center for Education Evaluation and Regional Assistance: A Guide to Using State Longitudinal Data for Applied Research
➢ Forum Guide to Supporting Data Access for Researchers: A Local Education Agency Perspective
➢ Forum Guide to Supporting Data Access for Researchers: A State Education Agency Perspective
➢ State SLDS profiles from the SLDS Project

---

[35]SLDS Project: Arkansas profile.

State Unemployment Insurance Wage and Benefits

Overview

State unemployment insurance (UI) data systems store information related to employment status, earnings, and UI benefits for all workers covered by UI. States report these data to the United States Department of Labor, Employment, and Training Administration, who produce quarterly aggregate reports such as data summaries, program performance, and information related to benefits and claims. State departments of labor or employment agencies are often the agencies authorized to manage UI systems and to administer UI benefit programs. UI systems can vary significantly from state to state. To give interested researchers an idea of the kinds of processes in place, New York State is profiled below. The information here is not comprehensive, but rather is meant to give an idea of the potential factors involved with acquiring and matching to state UI data. Researchers requiring more information should contact the labor or employment agencies of the states they want to work with. Other states have different parameters around data availability, data access, and costs, as well as different processes for matching and working with data.

Content/domain Employment, earnings, and unemployment insurance benefits of all workers covered by UI.

Ownership/authorizing agency U.S. Department of Labor and state employment departments.

Data Availability

Data element categories Vary by state.

For UI wage data: Nearly all, if not all, states have total earnings and calendar quarter. Some states share employer IDs, other states share only pseudo employer IDs, and some states have neither. A few states share hours worked. Some states share North American Industry Classification System (NAICS) codes that define what industry an individual's employer is part of.

For UI benefits data: Some states are able to share claims, claim and benefit amounts, and/or calendar quarter. As an example, the New York State Department of Labor (NYSDOL) shares Wage Record information (employee name, SSN, quarterly wages, etc.), Quarterly Census of Employment and Wages (UI account number, total wages, taxable wages, etc.) information, and UI benefits (initial claims, continued claims, combined wage claims, etc.) information.

Geographic coverage State level.

Population All employees covered by UI earnings. It is estimated that UI records cover approximately 90 percent of jobs, though the rate might be lower for low-wage workers.[36] For example, state administrative records will not have information on self-employment, jobs in the informal sector, or jobs

---

[36]Robert Kornfeld and Howard S. Bloom, "Measuring Program Impacts on Earnings and Employment: Do Unemployment Insurance Wage Reports from Employers Agree with Surveys of Individuals?," *Journal of Labor Economics* 17, 1 (1999): 168-197.

with the federal government. Furthermore, state administrative records do not have information on employment outside of the state. See National Directory of New Hires section in Part I for information pertaining to the differences in coverage between state and federal UI data.

Time period coverage Variable; states archive their data over variable time scales.

Periodicity Quarterly.

Lag Usually 6 months but could be longer; some states have quicker turnaround times due to electronic reporting.

Data retention, update, and deletion schedule(s) Vary by state. NYSDOL has data back to calendar year 2000, and new quarters of data are available 4 months after the end of a given calendar quarter.

Data Access

Data acquisition process Varies by state; some states use a formal research request process. To obtain data from NYSDOL, an application must be submitted that contains the following information: the purpose for which data are being requested, identification of all parties who may receive the information, how long the data will be needed/retained, if the data will be merged with other data, and confidentiality safeguards.

Are there restrictions on who can access the data? Yes. For example, most states require individuals store and work with their data in highly secure environments that may be available only at professional research organizations and universities. For New York, any federal, state, or local government agency or their agents or contractors may request data.

How long does the data acquisition process take? Varies from months to years. The initial request tends to be the most time consuming.

For New York, the data request process varies depending on the complexity of the request and how long it takes for the signed agreement to be returned to NYSDOL. NYSDOL reviews all applications that are received within 20 days.

Is a data-sharing agreement or memorandum of understanding required? Yes, in most or all states.

For New York, NYSDOL will prepare a data-sharing agreement between NYSDOL and other New York state agencies and a contract between NYSDOL and non-state government agencies. For data disclosures to contractors of the other government agency, NYSDOL will form an agreement with the government agency and the contractor.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Varies by state. Some states require evidence of signed informed consent forms and in some cases require researchers to collect signed state waivers authorizing the release of UI wage and benefits data for research purposes.

Access location(s) Variable. Some states' data are available in the Census Bureau's Federal Statistical Research Data Center. NYSDOL shares confidential data files with researchers through a secure file transfer.

Matching and Working with Data

Can a research sample be matched to the data source? Yes.

Matching process In some states, direct matches at the individual level are done by SSN. For NYSDOL, if interested researchers have their applications approved, the researcher will provide identifying information (SSN or employer ID) to identify the records requested. NYSDOL staff members will work with the researcher to arrange a secure file transfer. The PII is required to be encrypted in transit and at rest (in storage) using the Pretty Good Privacy (PGP) encryption standard.

What identifiers are needed for matching? Most matches are done by SSN. NYSDOL requires SSNs for matching.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Earnings data are easier to obtain than employer-level (ES-202) data.

Other restrictions for matching and working with data Varies by state. For New York, all safeguards are specified in the memorandum of understanding (MOU). Along with stipulations specified in the MOU, all individuals with access to confidential data must sign a nondisclosure agreement annually and take part in the UI Confidentiality Training developed by NYSDOL. Other restrictions are outlined in the NYSDOL data-sharing FAQ.

Cost

Is there a fee associated with accessing the data? Varies by state, but it is typical for states to charge a fee. These costs can vary based on a number of factors, including the number of matches and the size of the research sample. In some cases, there is no cost or the fee is negligible.

NYSDOL charges $1,000 to set up each agreement. There is an additional hourly fee to produce the data files, and this usually ranges from $80 to $100 but can vary depending on the complexity of the request.

Documentation

➢ Availability of online documentation varies by state
➢ NYSDOL presentation on Data Sharing Under Labor Law 537

State Vital Statistics

Overview

Vital statistics or records document information pertaining to life (or vital) events that are tracked or managed by a government authority. For states, these vital events often pertain to state-registered births, births of state residents that occur outside of that particular state, in-state deaths, and deaths of state residents that occur outside of that particular state. Vital records data may also include events related to marriage and divorce and issuing those respective certificates. Legal authority for the registration of these events belong to the 50 states, 2 cities, and 5 territories.[37] These entities are responsible for maintaining the collection and storage of vital records and for issuing the necessary certificates (for example, death, birth, and marriage). The National Center for Health Statistics receives these data from states and stores them in the National Vital Statistics System. California's vital records system is profiled below to give researchers interested in linking data an idea of the processes and requirements in place. The information below is not comprehensive, but rather is meant to give an idea of the potential factors involved with acquiring and matching to state vital records data.

Content/domain Vital events (for example, births, deaths, and fetal deaths).

Ownership/authorizing agency State health and human services, or public health agencies. California's data are managed by the California Department of Public Health.

Data Availability

Data element categories State vital records may contain birth, death, marriage, and divorce data. California vital records contain the following data elements:

- Birth data contain all state-registered births, and births to California residents that occurred out of state.
- Birth cohort data contain data for all live births and the infants who died in the first year of life.
- Death data contain in-state California deaths and deaths of California residents that occurred in other states or jurisdictions.
- Fetal death data contain data obtained from fetal death certificates registered in California.

Geographic coverage State level.

Population Varies by state. California birth files include all state-registered birth and births of California residents that occurred out of state. California death files include in-state deaths and deaths of California residents that occurred out of state.

Time period coverage Varies by state and by data file.

California has birth data for 1978 to 2016 and death data for 1960 to 2016 (note: fetal death data are more limited). More information can be found under Data Dictionaries on the vital records data

---

[37]The two cities are Washington, DC, and New York City. The five territories are Puerto Rico, the Virgin Islands, Guam, American Samoa, and the Commonwealth of the Northern Mariana Islands.

applications webpage and under [Vital Records Data File Types](#) on the data types and limitations webpage.

Periodicity Varies by state. Birth and death files are released annually. For California, most data are updated annually, but some death files can be requested on a weekly, monthly, or quarterly basis.

Lag Varies by state.

Data retention, update, and deletion schedule(s) Varies by state.

Data Access

Data acquisition process The process varies by state and the type of file researchers want to access. For the public-use birth and death files, downloadable files can often be found on state agency websites. For confidential data, interested researchers may have to complete a data request application.

For California, applications differ based on the researcher's affiliation with an entity (i.e., government or California local health departments) and the use of the data (i.e., for research or public). Researchers who are not affiliated with a state or federal government entity must complete a [research application](#) in which they describe what type of data files are being requested, the years of request, and the frequency of the data such as annual or others if they are available. Other required information includes but is not limited to the names of the researchers, a project description, and a [signed Information Privacy and Security Requirements form](#) outlining data security requirements for each organization in which vital records will be stored or accessed. Additional applications may also be required.

Are there restrictions on who can access the data? Varies by state.

How long does the data acquisition process take? Varies by state. For California, data requests can take on average 2 to 4 months to process once the data request application has been completed.

Is a data-sharing agreement or memorandum of understanding required? Varies by state. For California, researchers seeking vital records must complete the [Information Privacy and Security Requirements form](#).

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Varies by state.

Access location(s) Varies by state. Public-use data can be found on state agency websites.

Matching and Working with Data

Can a research sample be matched to the data source? Varies by state. For California, yes.

Matching process Varies by state. Researchers interested in linking to California vital statistics data must list each external data source during the application process. These vital records can be retained only for the duration of the approved study time frame and must be destroyed or returned to the California Department of Public Health.

What identifiers are needed for matching? Varies by state.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Varies by state. For California, researchers must justify each requested variable. California statute dictates that only the "minimum data necessary" can be disclosed. Justifications need to specify

how the data will be used to achieve the goals of the proposed study. Other states may have similar or additional restrictions. Look under Data Dictionary for more information.

Other restrictions for matching and working with data Varies by state.

Cost

Is there a fee associated with accessing the data? Varies by state. For California, data files are subject to cost recovery pursuant, and the cost depends on the complexity of the data request. An invoice is provided for the total cost of the data file after the application is received and required before the data files can be released to the researcher.

Documentation

➢ California Vital Records Data and Statistics

Supplementary Nutrition Assistance Program (SNAP) and Temporary Assistance for Needy Families (TANF) — State or Local

Overview

Content/domain SNAP and/or TANF (cash assistance) benefits.

Ownership/funding agency State or local human service agencies.

Data Availability

Data element categories

- SNAP: Many states and localities provide data that includes monthly benefit amounts received and information on sanctions for the SNAP case unit. Case-level data are typically easier to obtain than individual-level data.
- TANF: Many states and localities provide data that includes benefit amounts, benefit type, month benefits received, and information on sanctions for the TANF case unit. Some states provide case composition data that indicate who is on which TANF case and TANF case type (e.g., one-parent, two-parent, child-only). Some states will provide individual-level data on activities (often welfare-to-work activities), emptions from welfare-to-work requirements, and number of months counting toward federal or state time limits.

The Family Self Sufficiency Data Center provides sample TANF data. Some of the core data elements included are an individual and a case identifier, a geographic indicator, basic demographics such as age or gender, and whether the person receives other benefits like SNAP.

Geographic coverage State- or local-level.[38] It may be possible to acquire more granular geographic data such as county-level data, but this will vary by state. In some states, state human service agencies will have data for all cases across the state. In other states, these data might need to be acquired at the local (usually county but sometimes city) level, depending on the data fields requested. This is especially true in states where the TANF is locally-administered.

For example, in California, TANF data are maintained at the local level. As another example, in New York State, state human services agency records will typically include TANF benefits data for all cases and work participation data for nearly all counties, but will not contain work activity data for New York City, which must be obtained locally. In many other states, all TANF benefit data and work activity data will be available at the state level.

Population

- SNAP: SNAP recipients.

---

[38]Local may mean county, city, or office level.

- TANF: TANF recipients. Some eligibility requirements for TANF vary by state, but all applicants to TANF programs must be either pregnant or responsible for a child under 19 years of age. The majority of cases are part of single-parent households, though some two-parent households participate.

Some states also operate cash assistance similar to TANF, known as Separate State Programs (SSPs) or Solely State Funded Programs (SSFPs). These programs are exclusively state-funded and include populations who may be eligible for TANF in a state that lacked an SSP or an SSFP. Data on SSP and SSFP recipients may also be available upon request.

Time period coverage Varies by state, but often several years of data are available. It is usually possible to collect a minimum of 5 quarters of historical data at a time.

Periodicity Monthly.

Lag Varies by state. Typically, researchers allow for 3 to 6 months after the reporting period.

Data retention, update, and deletion schedule(s) Varies by state. Typically, state agencies archive data after 6 quarters, and archived data may cost more to acquire. States with more modernized data systems and management practices may maintain more data.

Data Access

Data acquisition process Varies by state. Some states or counties use a formal research request process. To obtain data from the Department of Public Social Services (DPSS) in Los Angeles County, for instance, a formal research request must be submitted listing the following information: indication of all parties and institutions conducting the research and having access to the data, background of the research project, type of research to be conducted, funding sources, benefits of the proposed research to Los Angeles County, project timeframes, and description of data requested. This initial research application then becomes the basis of a formal data sharing agreement. The data sharing agreement further specifies all this information (e.g., description of specific data elements or fields requested), and adds data security provisions, data deletion requirements, and signed confidentiality statements for all researchers accessing data. Once a data sharing agreement or memorandum of understanding is in place, researchers continue to work with county staff to further refine the data request elements and specifications and obtain a test file of results.

Are there restrictions on who can access the data? Varies by state and purpose of data request.

How long does the data acquisition process take? Months (typically, less than 1 year).

Is a data-sharing agreement or memorandum of understanding required? Yes, data-sharing agreements are typically required. In Los Angeles County, for example, this data-sharing agreement is formed out of the initial research request (described above). Some counties may share data with other social service programs that are under the same agency without a formal agreement in place.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Varies by state.

Access location(s) Varies by state. Some state data are available in the Census Bureau's Federal Statistical Research Data Centers. In the future, there may be a national repository of reported TANF data, but at this time most data need to be collected state by state.

Matching and Working with Data

Can a research sample be matched to the data source? Typically, yes.

Matching process Matching is usually at the case level (not the individual level), though it is not uncommon to have personal identifiers available. For TANF data, there are usually multiple people on a case (for example, the head of household and minor children), and some cases include only children.

What identifiers are needed for matching?

- SNAP: Usually can match by SSN or SNAP case number.
- TANF: Usually can match by TANF case number and/or SSN.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Varies by state, but identifiers are typically returned to researchers.

Other restrictions for matching and working with data Varies by state. Research personnel may be required to sign a nondisclosure agreement and take data security training. For TANF data from Los Angeles County, researchers must sign both a confidentiality agreement and a conflict of interest form.

Cost

Is there a fee associated with accessing the data? Varies by state, but some states may charge a fee. These costs can vary based on a number of factors, including the number of matches and the size of the research sample. In some cases, there is no cost, or the fee is negligible. Los Angeles County does not typically charge a fee specifically for data requests, but may receive funds as part of a larger research project or initiative.

Documentation

- Los Angeles County Department of Public Social Services Formal Research Request Information.
- Availability of online documentation varies by state and often depends on the contractor that states are working with to store and process the data.
- It is not uncommon for data dictionaries and layouts to be unavailable or only be available for an earlier period when data systems were first established or underwent their last major upgrade. This information may not be publicly available. In these instances, it may be helpful to request a user manual used by front-line staff, if one is available.

Part 3: Administrative Data Centers

This section includes information about administrative data housed in select federal and university-based data centers. Researchers interested in accessing multiple data sources for a single project may find data centers appealing, since these centers collect and retain data from multiple sources, including state social/human services programs. The geographic coverage, population, time period covered, and frequency of updates for each data set within the data centers are widely variable. In addition, some data sets can be accessed, linked, and analyzed only in these data centers.

Still, obtaining access to data centers can take considerable time and effort (relative to many state and national data sources, for example). Across the data centers featured in this Compendium, all have a formal data request process — including a research justification component — for access to restricted-use data. In some cases, this may involve obtaining Special Sworn Status (for example, to access data at the Census Bureau's Federal Statistical Research Data Centers). In addition, data coverage, availability, and completeness are variable — contingent upon the source providing the data to the center and — in some cases — the type of organization requesting the data.

Glossary — Data Centers

Overview

Content/domain A general description of the domains included in the data source — for example, demographic, public assistance, employment, or health data.

Ownership/authorizing agency The organization or agency that manages the data center and has the authority to provide access to the data for research purposes (not the organization or agency that provides that data housed at the data center).

Data sets housed at the data center A list of the data sets and sources housed at the data center as of fall 2018. If this list is available online, a link to the source may be provided.

Access to Data Center

Data acquisition process Description of how researchers request access to the data sets in the data center, including a summary of the process to obtain approval and access — for example, a research request might need to be submitted for approval, evidence of approval by an institutional review board (IRB) or study participant consent (and/or waiver of confidentiality) might be required, or a background check might need to be conducted.

Are there restrictions on who can access the data center? Yes/no answer to whether there are any restrictions on who can apply for access to the data for research purposes. If yes, a summary of the restrictions may be provided — for example, all researchers might be eligible; access may be limited to only (local, state, or federal) government contractors or research partners; or those wanting to access the data might need to sign a nondisclosure agreement.

How long does the data acquisition process take? A (rough) timeframe for how long the data acquisition process usually takes from the point when researchers first apply for access until the point when they are able to work with the data. If applicable, a timeframe may be given for each step in the acquisition process — for example, researchers might need to obtain Special Sworn Status before data can be accessed, a process that can take 4 to 6 weeks.

Is a data-sharing agreement or memorandum of understanding required? Yes/no answer to whether a data-sharing agreement of memorandum of understanding is required to access the data center. If yes, a description of the requirements may be provided.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes/no answer to whether evidence of an IRB review is required to permit access to the data center. If yes, a description of the requirements may be provided.

Access location(s) The physical or virtual locations where researchers access the data housed by the data center. If there are physical locations, a summary of how many there are and where they are located may be provided.

Matching and Working with Data

Can a research sample be matched to the data sets housed at the data center? Yes/no answer to whether researchers can submit a file with personal identifying information to the authorizing agency — for example, researchers may be able to send a list of SSNs for 1,000 study sample members, and that list is used to link the data source.

Matching process Summary of the process by which researchers initiate a match to the data center after they have been granted approval to work with the data. This includes a list of the steps in the matching process with links to any online resources that describe the steps, as appropriate — for example, how researchers submit a data file and how the data are linked.

What identifiers are needed for matching? A list of the person-level identifiers required to link/match a research sample to the data available through the data center — for example, SSNs. If a data center allows for fuzzy matching, a list of the fields — such as name, date of birth, and phone number — used in the algorithm is provided.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Yes/no answer to whether there are any restrictions on the data file that is returned to researchers. If yes, a list of the restrictions may be provided — for example, approved researchers may be allowed to work only with deidentified data or aggregate data.

Other restrictions for matching and working with data Summary of the limitations around accessing, matching, or using the data from the data center. These could be limitations pertaining to the data center as a whole or limitations around specific data sets or elements — for example, researchers might be required to work with the data from the data center on-site, or researchers might be given access only to deidentified data.

Cost

Is there a fee associated with accessing the data in the data center? Yes/no answer to whether researchers are charged a fee to access the data in the data center. If yes, an overview of the fee structure may be provided.

Documentation

Links to publicly available documentation on the data center. These sources could include documentation on how to apply for access to the data center, the data sets housed at the data center, input and output record layouts or codebooks, and any other relevant information.

Administrative Data Research Facility (ADRF)

Overview

Content/domain Business, employment, human services, corrections.[39]

Ownership/authorizing agency [The Coleridge Initiative at New York University.](#)

Data sets housed at the data center The ADRF contains a variety of data, including the Census Bureau's American Community Survey (ACS) and Longitudinal Employer-Household Dynamics (LEHD) data, Temporary Assistance for Needy Families (TANF) data from the Illinois Department of Human Services, Quarterly Census of Employment and Wages and Unemployment Insurance wage records from the Illinois and Missouri Departments of Employment Security, and corrections data from the Illinois Department of Corrections. Where applicable, individual person and business data sets can be linked across agencies and state lines using common identifiers (such as SSN, name, and employer identification number).

Access to Data Center

Data acquisition process All data access is project-based. To access data, users submit a project proposal that is reviewed by the relevant data owner(s).

Are there restrictions on who can access the data center? Yes. Only approved researchers on approved projects can access ADRF data.

How long does the data acquisition process take? Varies by project and data source, but typically a few months.

Is a data-sharing agreement or memorandum of understanding required? Yes.

---

[39]The ADRF is an analytical computing environment built to flexibly host a variety of data sets based on user needs. The ADRF is a secure, [FedRAMP](#)-authorized environment to store and share confidential data. It was initially established as a data facility that would inform the [Commission on Evidence-Based Policymaking](#). The approach has three dimensions:
1. Technical: A secure environment — the ADRF — within which data providers can place and share their data across agency and jurisdictional lines. The ADRF operates within the Amazon Web Services (AWS) government cloud and has been listed on [FedRAMP Marketplace](#) as a FedRAMP-authorized vendor. It has also received Authorization to Operate (ATO) from the U.S. Census Bureau.
2. Operational: It provides users with administrative tools that permit disparate data to be found, documented, and used — through discovery, stewardship, and collaboration modules.
3. Practical: It creates value for participating organizations by training the workforce through Applied Data Analytics (ADA) training programs. The focus of the programs is to ensure that data use both is consistent with the agency mission and creates demonstrable value to the agency.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes, and evidence of any IRB exemptions is acceptable.

Access location(s) Data in ADRF are accessed via a secure, cloud-based environment that is FedRAMP certified.

Matching and Working with Data

Can a research sample be matched to the data sets housed at the data center? Yes.

Matching process Varies by project and data source. A standardized approach to hashing the data in conjunction with the U.S. Census Bureau has been developed. In that protocol, New York University and the data provider will agree on a hash algorithm to hash all values — the Hash-based Message Authentication Code (HMAC) algorithm.[40]

What identifiers are needed for matching? Varies by project and data source.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Yes. There is a disclosure review process, determined by the data owner, that must be passed before information can be exported from the system. For example, any data point a researcher wants to export that uses the Illinois Department of Security data must be an aggregation of at least 10 individuals (in addition to some other restrictions).

Other restrictions for matching and working with data Varies by project and data source.

Cost

Is there a fee associated with accessing the data in the data center? Yes, the cost varies by project and data source.

Documentation

➢ Description of the Administrative Data Research Facility
➢ The Coleridge Initiative at New York University
➢ Information on the Commission on Evidence-Based Policymaking
➢ Information on FedRAMP and the FedRAMP Marketplace
➢ More information on the available data sets: the Census Bureau's American Community Survey (ACS) and Longitudinal Employer-Household Dynamics (LEHD) data

---

[40]In an HMAC, "salt" is used to create an encryption key that is then used to encrypt a "message" (the value being hashed), which is then hashed (using SHA256 in our implementation). Using the salt to seed an encryption algorithm allows the additional information from our salt to be more uniformly and opaquely distributed throughout the value that is then hashed, rather than arbitrarily placing it on one end or the other or interspersing portions of it algorithmically. Details available on request.

Census Bureau's Federal Statistical Research Data Centers (FSRDC)

Overview

Content/domain Business, household, demographic, and health data.

Ownership/authorizing agency U.S. Census Bureau.

Data Availability

Data sets housed at the data center The Census FSRDCs contain nonpublic data from:

- Census Bureau — 51 data sets are housed at the FSRDCs. These data come from federal, state, or third-party sources, and the sets include data from the Department of Housing and Urban Development (HUD), unemployment insurance (UI) data from all states, the District of Columbia, and Puerto Rico, and credit bureau data from Experian. A list of available data sets is available at Census Bureau Administrative Data Inventory.
- National Center for Health Statistics (NCHS). Most of the NCHS restricted data are available to approved researchers through the Federal Statistics Research Data Center (FSRDC) network.[41]
- Agency for Healthcare Research and Quality.[42]
- Bureau of Labor Statistics (BLS).[43]

Access to Data Center

Data acquisition process Researchers must follow a proposal process to access the data housed at the FSRDCs. More information on the FSRDC proposal process can be found at the FSRDC Available Data website and the application portal for restricted data can be found at the Restricted Data for Federal Statistics.

This process varies somewhat depending on which specific data sets the researcher is applying to access. Researchers apply directly to the agency from which they are requesting data (for example, the Bureau of Labor Statistics). They must also contact the FSRDC where they will access the data. The steps to apply for access to Census Bureau restricted-use data can be found on the Census Bureau's How to Apply webpage, and the proposal process for working with data sets not housed at the Census Bureau can be found on the Bureau of Labor Statistics' Restricted Data Access webpage (for data sets from the Bureau of Labor Statistics) and on the National Center for Health Statistics: FSRDC webpage (for data sets from the National Center for Health Statistics and Agency for Healthcare Research and Quality).

---

[41] On site at a Federal Statistical RDC.

[42] Restricted Data Files Available at the Federal Statistical Research Data Centers.

[43] BLS data access through the Federal Statistical Research Data Centers (FSRDCs) is currently unavailable. A financial agreement between the BLS and the Census Bureau is not yet in place to authorize BLS data access through FSRDCs for Fiscal Year 2018. BLS will still review proposals to access data at FSRDCs, but approved researchers will be subject to a delay for an uncertain time period. The option for applying and being approved for data access at the BLS National Office in Washington, DC, instead of at FSRDCs, is still available.

Are there restrictions on who can access the data center? Yes. All FSRDC researchers must obtain Special Sworn Status from the Census Bureau to begin work on their projects. This process includes completing a Moderate Background Investigation and interview conducted by the Office of Personnel Management.

How long does the data acquisition process take? Varies by agency, predominantly due to proposal development activities and reviews. It can take around 1 year for projects using data from the Census Bureau and may take less time for projects using data from the Bureau of Labor Statistics, the National Center for Health Statistics, or the Agency for Healthcare Research and Quality.

Is a data-sharing agreement or memorandum of understanding required? Yes. Proposal forms can be found at the How to Apply website.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? Yes, as required by their research institution.

Access location(s) There are 29 data centers nationwide. Most are on college campuses. All data processing and analysis must be done at one of the FSRDC locations.

Matching and Working with Data

Can a research sample be matched to the data sets housed at the data center? Yes.

Matching process Census Bureau staff members append Protected Identification Keys (PIKs) to researcher-provided data sets. All PII on the researcher-provided data set(s) is then deleted, and the PIK'd file is made available for research in the secure FSRDC computing environment. Children of sample members can be identified through a KidLink file (based on the Social Security Administration's Numerical Identification [NUMIDENT] database or tax records).

What identifiers are needed for matching? SSN and other identifiers can be used.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Researchers receive a deidentified data file for analysis, and only aggregated results may be removed from the secure computing environment. Researchers may NOT remove results on their own. The aggregated results must be reviewed for identity disclosure risk and approved by Census Bureau staff members, who then send the results to the researcher via email.

Other restrictions for matching and working with data The data access requirements vary by data set and state. For example, access to data containing Federal Tax Information (FTI) requires an additional review by the IRS after the Census Bureau completes its review. In terms of state-level data, states can choose whether to allow researchers access to their data. Some states allow access, other states require an additional review after the Census Bureau review, and other states do not allow access at all.

Cost

Is there a fee associated with accessing the data in the data center? Yes. Each participating agency determines fees for making its data available for research. The Census Bureau does not charge researchers for accessing data. See detailed information at BLS Restricted Data Access: Fees and Invoicing and NCHS: Fees and Invoicing.

FSRDC locations may charge researchers for facility access. Contact your local Research Data Center for more information.

Documentation

➢ Federal Statistical Research Data Centers
➢ FSRDC Research Proposal Guidelines
➢ Census Bureau Proposal Registration Form used to apply for access to the FSRDC for a given project
➢ Lists of available restricted data sets from the Census Bureau housed at the FSRDCs
➢ Additional steps in the proposal process for working with non-Census Bureau housed data sets can be found here: Bureau of Labor Statistics and National Center for Health Statistics

Inter-university Consortium for Political and Social Research (ICPSR)

Overview

Content/domain Social and behavioral sciences, education, aging, criminal justice, substance abuse, terrorism, and other fields.[44]

Ownership/authorizing agency ICPSR.

Data sets housed at the data center ICPSR houses restricted-use and public-use files from over 10,500 studies. The full list is available from ICPSR: Find & Analyze Data.

Access to Data Center

Data acquisition process To download data or analyze data from ICPSR's website, users need to log in (authenticate) and agree to terms of use. Users may log in to ICPSR using a Facebook or Google account. They may also use an ICPSR MyData account if they prefer not to use third-party authentication. The website also contains additional information on how ICPSR handles data access and authentication.

Public-use data: Requesters can access public-use data by downloading the files directly from the home page of a public-use study. Requesters affiliated with an ICPSR member institution can access ICPRS members-only data for free. Requesters who are not affiliated with an ICPSR member institution need to pay a fee of $550.

Restricted-use data: Access is granted to these files following an application process during which researchers agree to follow strict legal and electronic requirements for maintaining data confidentiality. Requesters should review ICPSR's information on accessing restricted data.

Are there restrictions on who can access the data center? The vast majority of ICPSR data holdings are public-use files with no restrictions on access beyond ICPSR's standard terms of use. However, in some cases ICPSR provides access for vetted researchers and sponsor-supervised students to restricted-use data versions that retain confidential or sensitive data.

Researchers can access restricted-use data (RUD) by showing the data are needed for legitimate research purposes and by signing and adhering to the Restricted Data Use Agreement (RDUA). Some RUD investigators (as specified in the RDUA) are required to have a terminal degree and hold a faculty appointment or research position at the institution that will be party to the RDUA. For some RUD, access is limited to researchers employed by an organization possessing a current NIH Multiple Project Assurances (MPA) Certification Number. Access for individuals employed by organizations that do not have an MPA Certification Number may be obtained by providing additional information as requested in the application.[45]

---

[44]About ICPSR.
[45]Restricted-Use Data Management at ICPSR.

How long does the data acquisition process take?

Public-use data: Data that are freely available can be obtained immediately via direct downloads.

Restricted-use data: The process depends on the access level of the data collection as well as the completeness of a requester's application. Typically, access to restricted-use data is granted within 5 to 7 business days.

Is a data-sharing agreement or memorandum of understanding required? Yes. Data users accessing a study from ICPSR will be required to agree to the terms of use prior to receiving the files, and a copy of the terms of use will be included with their download. Those planning to use ICPSR data are encouraged to review ICPSR's Terms of Use.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? To access some of ICPSR's restricted-use data, applicants provide documentation of IRB approval or exemption for the proposed research project. Requesters are strongly encouraged to review the general application requirements for access to any restricted-use data.

Access location(s) For public-use data, data users typically acquire the files via web downloads from the study home pages.

ICPSR offers several methods of restricted-use data access: Secure Dissemination, Virtual Data Enclave, and Physical Data Enclave. For select data sets, ICPSR also offers Restricted Survey Documentation and Analysis (SDA). Requesters can typically start their restricted-use data application process via the home pages of any associated study.

Matching and Working with Data

Can a research sample be matched to the data sets housed at the data center? Yes, in some instances and when doing so does not compromise the confidentiality of the research subjects. Some restricted data sets cannot be merged with extant data sources.

Matching process There is no formal process in place at this time, since most data sets housed at ICPSR do not have PII.

What identifiers are needed for matching? Not available.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? Not available.

Other restrictions for matching and working with data Not available.

Cost

Is there a fee associated with accessing the data in the data center? Yes. Requesters who are affiliated with an ICPSR member institution can access ICPSR member-only data for free. Requesters who are not affiliated with an ICPSR member institution need to pay a fee of $550.

The majority of restricted data at ICPSR are federally funded and do not have an access fee, but some studies do have fees. If a study has an access fee, this information will be displayed within the application.

Documentation

- ➤ Information on accessing restricted data from ICPSR
- ➤ List of evaluations with available restricted-use and public-use files
- ➤ About ICPSR
- ➤ ICPSR's Terms of Use

National Center for Health Statistics' Research Data Centers (NCHS RDCs)

Overview

Content/domain Health.

Ownership/authorizing agency National Center for Health Statistics.

Data Availability

Data sets housed at the data center The NCHS RDCs contain both public-use and restricted-access data sets. The restricted-access data sets include data from several health-related surveys. The RDCs also host some restricted data from the Department of Health and Human Services.

The public-use data sets housed include: National Health and Nutrition Examination Survey (NHANES), National Health Care Surveys, National Vital Statistics System (NVSS), National Survey of Family Growth (NSFG), National Health Interview Survey (NHIS), National Immunization Survey (NIS), Longitudinal Studies of Aging (LSOA), State and Local Area Integrated Telephone Survey (SLAITS), the Compressed Mortality File, and the Stochastic Population Analysis for Complex Events (SPACE) Program.

The restricted-use datasets housed include: Geographic codes for all NCHS Surveys, NHANES, National Health Care Surveys, NHIS, NSFG, SLAITS, NVSS, National Maternal and Infant Health Survey, Drug Involved Mortality Restricted Variables, and Redacted Death Certificate Literal Text File.

The NCHS also links some NCHS survey data with data from the National Death Index (NDI), the Centers for Medicare and Medicaid Services (CMS), the United States Renal Data System (USRDS), the Social Security Administration (SSA), and the Department of Housing and Urban Development (HUD). More information about NCHS's data linkage activities can be found at NCHS Data Linkage Activities.

Geographic coverage The data sets available through the RDC vary by geographic coverage.

Population The data sets available through the RDC vary by which populations are covered.

Time period coverage The data sets available through the RDC vary by time period covered.

Periodicity The variety of data sets available through the RDC vary by periodicity.

Lag The variety of data sets available through the RDC vary by lag time.

Access to Data Center

Data acquisition process Researchers must submit the RDC Research Proposal to apply for access to the data at the RDCs. The proposal must explain why the researcher needs access to the restricted data, which data elements are being requested, and how the results will be used. All proposals are reviewed by a committee, and the committee returns a decision of approved, denied, or revise and resubmit.

Are there restrictions on who can access the data? All researchers must complete a confidentiality orientation and score 100 percent on a test and submit two confidentiality forms (that is, the data access agreement and designated agent agreement).

How long does the data access process take? The application review process can take 8 to 12 weeks. A researcher who wants to access data at a Census RDC must obtain Special Sworn Status (SSS) from the Census Bureau, which can take an additional 4 to 6 weeks.

Is a data-sharing agreement or memorandum of understanding required? Researchers are expected to use the data as indicated in their approved proposal. In addition to the proposal, a data access agreement and a designated agent agreement must be completed, signed, and notarized.

Do researchers need to provide evidence of an IRB review to access these data for research purposes? No.

Access location(s) There are three NCHS RDCs: in Hyattsville, Maryland; Atlanta, Georgia; and Washington, DC.

Most data sets available through the NCHS RDC can also be accessed at the Census Bureau's Federal Statistical Research Data Centers.[46]

Matching and Working with Data

Can a research sample be matched to the data sets housed at the data center? Yes. External databases proposed to be matched to restricted-use data must be described in the proposal.

Matching process? If matching external data to restricted-use data is approved in the proposal, then NCHS RDC staff members will conduct the match for the researcher.

Researchers must not include any variables that would help them reidentify individuals in their match files.[47]

What identifiers are needed for matching? Only indirect identifiers (for example, geographic location) are available for matching.

Are there restrictions on the types of data that can be returned to researchers (for example, personal identifiers)? PII is not returned to researchers. Final tables needed for publication are returned to researchers after they clear a disclosure review. Exploratory analyses and large volumes of tables are not returned to researchers.

Other restrictions for matching and working with data? Exact matching using direct identifiers is not allowed. Researchers should be familiar with the rules for conducting research on-site at the RDC for accessing and analyzing restricted-use data. Researchers should also be familiar with the analytical and publication guidelines. External databases for matching must be nationally representative.

All analysis must be done on-site at the RDC. RDC staff members will review all output, and after approval, output will be emailed to the researcher. The output review process can take up to 3 weeks.

---

[46]CDC/National Center for Health Statistics: On site at a Federal Statistical RDC
[47]CDC/National Center for Health Statistics: Preparing for proposal submission

Cost

Is there a fee associated with accessing the data in the data center? Yes, there are fees for data access. While the fees vary by project, NCHS estimates the average cost of fees for a project to be around $3,000.[48]

---

[48]CDC/National Center for Health Statistics: Fees and Invoicing

Documentation

- ➢ General information on NCHS Research Data Center
- ➢ The RDC's available public-use and restricted-access data sets
- ➢ Information about NCHS' data linkage activities
- ➢ Frequently Asked Questions about NCHS Data Linkage
- ➢ Information related to NCHS Fees and Invoicing
- ➢ RDC Research Proposal form

Index

In this section, MDRC has categorized the sources into the following domains: employment and earnings, health and child care, education, housing, and benefit data. Sources can be sorted into more than one category. Only administrative data sources have been sorted into domains. Administrative data centers cover all five domains.

Benefits

Employment and Earnings

Education

Health and Child Care

Administrative Data Research Facility, 56
Census Bureau's FSRDC, 58
Child Care and Development Fund, 7
Early Childhood Integrated Data System, 37
Inter-university Consortium for Political and Social Research, 61
Medicaid Analytic eXtract, 14
Medicare Master Beneficiary Summary File, 17
National Center for Health Statistics' RDC, 64
National Death Index, 20
State Vital Statistics, 47

Housing

Administrative Data Research Facility, 56
Census Bureau's FSRDC, 58
Eviction Lab, 10
Housing and Urban Development IMS PIC, 12
Inter-university Consortium for Political and Social Research, 61