J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
NORTH AMERICA

# CATALOG OF ADMINISTRATIVE DATA SETS

Laura Feeney, Jason Bauman, Julia Chabrier,
Geeti Mehra, Kyle Murphy, Michelle Woodford

J-PAL North America, December 2015

povertyactionlab.org/na

**Summary:** This catalog provides information on how to obtain administrative data for a randomized evaluation from specific data sources. It is intended to assist researchers in screening potential data sources and to be a source of preliminary information, rather than a comprehensive data acquisition instruction manual.

We prioritize reviewing data sources that have been used by a researcher in the J-PAL network for a randomized evaluation conducted in the United States. We also include data sources have been investigated by J-PAL affiliates or staff for use in a randomized evaluation, even if they have not yet been used. Data sets that are *not* likely to be useful in a randomized evaluation may also be included. These are included to assist researchers in saving their time by *excluding* these sources from their search.

We encourage readers to share any insights they have into these or other data sets. Please contact Laura Feeney at lfeeney@povertyactionlab.org with any additional information on data sets included in the catalog, or with suggestions for further data sets to include.

Additional resources may be found in J-PAL North America's guide to Using Administrative Data for Randomized Evaluations.

**Acknowledgements:** We thank Mary Ann Bates, Stuart Buck, Amy Finkelstein, Daniel Goroff, Lawrence Katz, Josh McGee, and Heidi Williams for recognizing the need for this type of catalog, and for their ideas and inspiration along the way. Belinda Tang provided excellent research assistance. Alicia Doyon formatted the guide, tables, and figures. This work was made possible by support from the Alfred P. Sloan Foundation and the Laura and John Arnold Foundation. Any errors are our own.

## TABLE OF CONTENTS

# CALIFORNIA HEALTH CARE DATA

## California Office of Statewide Health Planning & Development (OSHPD)

*Clinical care and patient records data from licensed general acute care hospitals, emergency departments, and ambulatory care surgery centers in California. Includes inpatient discharge and outpatient encounter information.*

## ACCESS

OSHPD maintains two privacy levels of data files, and these privacy levels determine the availability of data and access procedures.

**Public Use Files (PUF)** are available to anyone for purchase. Data are anonymized, but include clinical, payer, and facility information. Instructions for ordering these files can be found on the Public Use File Requests page.

**Nonpublic Confidential Use Files** are available to researchers associated with nonprofit educational institutions and to state agencies. OSHPD will not release Social Security number and does not list other identifiers in the data documentation. Eligible researchers must request the data through OSHPD and obtain approval for their research protocol from the state Internal Review Board for the California Health and Human Services Agency, also known as the Committee for the Protection of Human Subjects (CPHS). The request process, including all necessary forms, is detailed here.

OSHPD may link patient data to birth or death certificate data. Any request for such linked data requires additional review by the California Department of Public Health Vital Statistics Advisory Committee.

### Timeline for Access

**Public use** files are generally delivered within 1-5 days.

The timeline for **confidential** data files depends on the complexity and completeness of the request. On average, the time between submission of a request and receipt of data is about six months.

*Lag Time*
Unknown. As of December 2015, data are available through 2014.

**Links**
OSHPD Data Request Center

Overview (from OSHPD)

Data Documentation (from OSHPD)

**Unit of Observation** Encounter

**Personally Identifiable Information Available for Linking?** Yes

**Geography** CA, USA

**Years Available** PUF: 2010 – present
Confidential: 2005 – present, certain files available since 1983

**Free or Paid** Paid. Some data are available free to university researchers

**Frequency of Updates** Unknown. As of December 2015, data are available through 2014.

**Universe** Encounter visit records from licensed general acute care hospitals, emergency departments, and ambulatory care surgery centers in CA. Many facilities that are called ambulatory surgery centers are not required to be licensed as surgical clinics, and do not report data to OSHPD.

## COST

The three most recent years of Public use Patient Discharge Data (PDD), Emergency Data (ED) and Ambulatory Surgery Data (AS) data files are provided on a complimentary basis to eligible university researchers and nonprofit institutions.

PDD confidential data files are $200 per year; ED and AS confidential data files are $200 for six-month data files.

For more information, see OSHPD's pricing policy.

## LINKING

Researchers may request data for all records, or a subset of records (e.g. geographic area, diagnosis). It is unknown whether OSHPD will match records to a list of study participants.

When linking to outside data sources is necessary, OSHPD says, "Details of how the linkage can occur without the identifiers being released to the researcher need to be discussed. For example, you could track "Patient A" but not actually know who this person is."

### Personally Identifiable Data Available for Linking

OSHPD confidential data contain Social Security numbers. Other available identifiers are unknown.

### Linking to Outside Data Sources

Researchers must request permission from OSHPD for linking patient-level data across years/data sets. OSHPD requires this information even if the researcher is just linking within/between OSHPD data sets.

OSHPD records linked with California's Vital Statistics Birth Statistical Master File, Birth Cohort File, and Death Statistical Master Files are also available for request through OSHPD.

## DATA CONTENTS

The three main files provided by OSHPD are Patient Discharge Data (PDD), Emergency Data (ED) and Ambulatory Surgery Data (AS). For a full list of variables, see the OHSPD's Data Request Center, which hosts an excel file of the Patient Data Master Variable Grid, a Nonpublic PDD Variable Availability Table, and a Nonpublic ED/AS Variable Availability Table. Full data dictionaries are available on OSHPD's Data Documentation page.

### Partial List of Variables

Patient discharge data, emergency department data, patient demographics, diagnosis codes, MS-DRG, ICD-9, ICD-10, admission dates, discharge dates, hospital identifiers, payer information, clinical information, ambulatory surgery center data, outcomes related to Coronary Artery Bypass Graft (CABG)

### Subjects/Tags

Health; healthcare; health utilization; encounters; acute care hospitals; emergency departments; ambulatory care surgery

*Randomized Evaluations Using This Data Set*
Unknown.

# CHICAGO PUBLIC SCHOOLS DATA

## Chicago Public Schools (CPS) Office of Performance

*Student- and staff-level information collected by CPS, including Illinois State Achievement Test (ISAT) scores, gender, and race.*

## ACCESS

Identified data are available to researchers by application. Researchers must submit an "External Data Request Form for Research or Program Evaluation Data" to the CPS Office of Accountability. Applications will be reviewed by the CPS Research Review Board (RRB). As part of the application process and as required by the Federal Educational Rights and Privacy Act (FERPA), researchers must obtain prior written informed consent from students (or the parents of students less than 18 years of age) and staff for whom they are requesting identifiable data. From publicly available documentation, it is not clear whether it is possible to request a de-identified data set and obtain a waiver of the consent process.

In order to obtain access to student- or staff-level data, researchers must also sign a Data Security Agreement, which identifies requirements for the storage, use, maintenance, protection, dissemination, and destruction of the data.

**Links**
CPS Research Resources

CPS Research Study and Data Policy

**Unit of Observation** Individual

**Personally Identifiable Information Available for Linking?** Yes

**Geography** Chicago, IL, USA

**Years Available** Unknown

**Free or Paid** Paid

**Frequency of Updates** Unknown

**Universe** Chicago Public Schools students and staff

### Timeline for Access

The review of an application for data usually takes one month. The RRB meets monthly to evaluate all requests to conduct research, and the meeting schedule is available here.

The amount of time required to receive the data files after approval, and the amount of time for which researchers can use these files, is not known.

*Lag Time*
Unknown.

## COST

There is a $50 application processing fee to request CPS data. The CPS Office of Performance may also establish fees to charge researchers for review and evaluation of proposals and the compilation of data. Details regarding these fees are not known.

## LINKING

The linking process for CPS data files is not known. From publicly available documentation, it is unknown whether CPS would link their data to a pre-defined study sample. It is unknown whether CPS would be willing and able to send a de-identified version of the data set for research purposes.

### Personally Identifiable Information Available for Linking

Unknown.

## DATA CONTENTS

### Partial List of Variables

ISAT scores, gender, race, age, birth year, measures of special education needs, US Census tract of residence

### Subjects/Tags

Students; Chicago; test scores; education; school

*Randomized Evaluations Using This Data Set*

Aizer, Anna, and Joseph J. Doyle Jr. March 2015. "Juvenile Incarceration, Human Capital and Future Crime: Evidence from Randomly-Assigned Judges." *The Quarterly Journal of Economics* doi: 10.1093/qje/qjv003.

# MEDICARE DATA

## Center for Medicare and Medicaid Services (CMS)

*Claims data from Medicare Parts A and B, prescription drug data from Part D, beneficiary information, and cost reports.*

## ACCESS

There are three levels of Medicare data, and the level of data requested determines the availability and access procedures. Data are available to any class of researcher, though some data requests require review and approval by CMS, a formal Data Use Agreement (DUA), and approval from the CMS Privacy Board.

**Non-Identifiable Data Files** are public use data and are available to anyone for purchase on the PUF/Non-Identifiable Data Files page of the CMS website.

**Limited Data Sets** require a DUA, but are not subject to Privacy Board review. Instructions for requesting a Limited Data Set are here.

**Research Identifiable Files** require a DUA and are subject to Privacy Board review. The Research Data Assistance Center (ResDAC) at the University of Minnesota is the intermediary for processing and filing requests for these data. The request process, including all necessary forms and tips for completing them, is detailed here.

### Timeline for Access

The timeline for a research identifiable file request varies, but ResDAC recommends planning 3-4 months between a draft application for data and receipt of the data. See the Data Request Timeline for details on the stages of a request.

Data must be destroyed upon reaching the expiration date of the DUA. Researchers must request an extension to continue working with the data beyond that time.

**Links**
List of data files (from CMS)

List of files (from ResDAC)

General documentation (ResDAC)

Data dictionaries

**Unit of Observation** Individual beneficiary; Medicare claim

**Personally Identifiable Information Available for Linking?** Yes

**Geography** USA

**Years Available** Varies by file; most available 1999-present. See File Availability.

**Free or Paid** Paid

**Frequency of Updates** Annual with an approximately one year lag (e.g., CY 2014 expected in Dec 2015). Some files are released quarterly.

**Universe** All Medicare Fee for Service recipients in the US: over 65 million beneficiaries.

### *Lag Time*

Files are updated annually, and are available on approximately a one year lag, with updates usually available in December. Additionally, some files are updated quarterly, and available on a 5-6 month lag. For a list of available file-years and upcoming file availability, see File Availability.

## COST

CMS generally charges for data by the file-year, rather than by the individual record or number of records. See Pricing Information for CMS Data Files. The ResDAC Assistance Desk can provide a ballpark cost estimate, or a formal cost estimate with the completion of the Cost Estimate Request form submitted to resdac@umn.edu.

Limited Data Sets and Research Identifiable Files may also be available for "reuse" at a lower cost if a particular file is already available to a researcher within the same research organization. Contact your university department's data manager for more information on which files may be available. NBER affiliates should also contact the NBER.

## LINKING

Researchers may request data for the full set of Medicare beneficiaries, on a random sample of beneficiaries, or define their cohort. There are two options for defining a cohort:

**Option One:** Researchers may limit the cohort by sex, age, date of death, race, residence, or Medicare status.

**Option Two:** Researchers may send a finder file with a list of individuals for whom records are requested. This file will be linked using an exact match on one of the variables to be linked on (see below). For more information, see Submission of Medicare Data Finder and Crosswalk Files.

### Personally Identifiable Information Available for Linking

According to ResDAC, finder files must contain one or more of the following:

- Beneficiary IDs received from a previous data shipment from CMS
- Health Insurance Claim numbers
- Social Security numbers
- RES_ID / State Code - Identifies resident in the national repository
- Unique Physician Identification Number

CMS files also contain name and date of birth, but it is unknown whether CMS will match based on these variables.

### Linking to Outside Data Sources

Researchers must request permission prior to matching CMS data with any external data sources, or with CMS files not listed in the initial Data Use Agreement.

## DATA CONTENTS

### Partial List of Variables

Beneficiary demographic information; claim payment amount; type of claim; claim procedure and diagnostic codes; ICD-9; ICD-10; location of claim; Drug plan characteristics: copay, coinsurance, type of donut-hole coverage; Prescription drug event characteristics: drug NDC11, drug cost, drug OOP cost, drug benefit phase of claim; prescriber NPI; prescription filled location

### Subjects/Tags

Medicare; health; health utilization; healthcare

*Randomized Evaluations Using this Data Set*
Unknown.

# NATIONAL DEATH INDEX

## Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Division of Vital Statistics

*Death record information from state vital statistics offices. Includes date and cause of death.*

## ACCESS

Identified data are available to investigators for statistical purposes in medical and health research. Researchers must submit a data request application form to NCHS to access the data. As part of the application process, researchers must obtain approval from an Institutional Review Board (IRB), and sign a confidentiality agreement which requires the researchers to specify how they will securely store the data.

The request process, including links to necessary forms, is detailed in Chapter 1 of the NDI User's Guide.

### Timeline for Access

The review of an application for data usually takes 2-3 months. Data are typically received 2 weeks after the application is approved.

After 5 years, any data that has identifiable information that came from the NDI match must be destroyed, and NCHS must be notified (unless the researcher specifies a justification for keeping the data longer in their initial application).

#### *Lag Time*

Death records are added annually, approximately 12 months after the end of the calendar year. NCHS is piloting an "Early Release" program in which researchers can request data sooner, but state records may not be complete by that time.

**Links**
CDC National Death Index

National Death Index User's Guide

**Unit of Observation** Individual

**Personally Identifiable Information Available for Linking?** Yes

**Geography** USA

**Years Available** 1979 – Present

**Free or Paid** Paid

**Frequency of Updates** Annual with an approximately one year lag (e.g., CY 2014 expected in Dec 2015).

**Universe** All deaths in the US that are on file in state vital statistics' offices

## COST

The fees for routine NDI searches consist of a $350.00 service charge plus $0.15 per user record for each year of death searched. For example, 1,000 records searched against 10 years would cost $350 + ($0.15 x 1,000 x 10) or $1,850. Fees for the "NDI Plus" service are slightly higher. More information on user fees is available on NCHS' user fees worksheet.

## LINKING

Researchers must send a CD containing a file(s) of individuals for whom records are requested to NCHS by overnight delivery. Further details regarding preparing records to be sent to NCHS can be found in Chapter 2 of the NDI User's Guide.

### Personally Identifiable Information Available for Linking

NCHS requires that the researcher's finder file has at least one of these three combinations of variables in order to search for matches in the NDI data set:

1. First name, last name, month and year of birth
2. First name, last name, Social Security number (SSN)
3. SSN, date of birth, sex

NCHS counts records as "matches" if they meet one of the following seven criteria:

1. Exact match of SSN
2. Exact month and +/- 1 year of birth, first and last name
3. Exact month and +/- 1 year of birth, first and middle initials, last name
4. Exact month and day of birth, first and last name
5. Exact month and day of birth, first and middle initials, last name
6. Exact month and year of birth, first name, father's surname
7. If the subject is female: Exact month and year of birth, first name, last name (on user's record), and father's surname (on NDI record)

## DATA CONTENTS

Variable lists and data documentation for each file are available in the NDI User's Guide.

### Partial List of Variables

Date of death, state of death, death certificate number, cause of death codes in ICD-9 or ICD-10 format

### Subjects/Tags

Vital statistics; death; mortality

*Randomized Evaluations Using this Data Set*

Jacob, Brian A., Jens Ludwig, and Douglas L. Miller. 2013. "The Effects of Housing and Neighborhood Conditions on Child Mortality." *Journal of Health Economics* 32(1): 195-206. J-PAL Evaluation Page.

*Other Research Using this Data Set*

Use of the NDI in Health Research (CDC)

# NEW YORK CITY VITAL STATISTICS DATA

## New York City Department of Health and Mental Hygiene

*Individual-level data from birth and death certificates in New York City.*

## ACCESS

Data are available to any class of researcher. For non-aggregated data, researchers must submit a request to the New York City Office of Vital Statistics (OVS). If requesting linkages or matching, researchers must submit an additional form with information about the data sets to be matched, variables for identification, as well as their data set for matching. Identifiers, or a de-identified case ID may be included in the returned data set, but researchers must specifically request those variables in Section C of the data usage application. Otherwise, the OVS will perform the match, then return the data set with all identifiers stripped, including those provided to conduct the match. As part of the application process, researchers must have IRB approval and complete a Data Use and Non-Disclosure Agreement.

### Timeline for Access

The OVS strives to provide customized data and vital record copies to approved researchers within eight weeks of application submission. However, the timeline is subject to analyst workload, priority internal requests, and the resolution of any questions or concerns about the Data Use Agreement.

Upon the project's completion or within three years, any data that has identifiable information that came from the New York City Department of Health and Mental Hygiene (DOHMH) or the National Death Index must be destroyed, and DOHMH must be notified.

**Links**
Overview

Application Process for Identifiable Vital Statistics Data

**Unit of Observation** Individual death or birth

**Personally Identifiable Information Available for Linking?** Yes

**Geography** New York City, NY, USA

**Years Available** 1996 – present. Some variables available 1978 – present.

**Free or Paid** Free

**Frequency of Updates** Annual with an approximately one year lag

**Universe** All birth and death records in New York City since 1996

### *Lag Time*

Files are updated annually, and are available on approximately a one year lag, following the release of the annual Summary of NYC Vital Statistics.

## COST

Access to records is free.

## LINKING

The Office of Vital Statistics (OVS) provides data matching services for approved research. To request a linkage, DOHMH requires researchers to submit an additional application form. In the application, researchers must specify the vital data type and year which they are requesting (e.g., "All birth records from 2001."). Additionally, researchers must provide a list of databases with which the OVS data will be linked, the anticipated protocol for matching the data to those data sets, and a list of variables on which those matches will be made. Researchers then submit their data to the OVS in SAS, CSV, or another format which

provides data element formats and nomenclature. All identifiable data must be submitted by a secure method approved by the DOHMH. DOHMH will process the match and return the data set to the researcher.

Unless specifically listed and justified in the original data request, OVS will strip all identifiers before returning the data set, including those provided to conduct the match.

### Personally Identifiable Information Available for Linking

Full name and full date of birth (DOB) are the minimum required data elements for a match. A description of the linking process is available here. Other available variables include:

Birth Records

1. Case ID
2. Infant's full name
3. Mother's full name, maiden name, DOB, Social Security number (SSN) (last 4 digits may be used for matching purposes), and address
4. Father's full name, DOB, and SSN
5. Medical record number on birth certificate

Death Records

1. New York City Office of the Chief Medical Examiner case ID
2. Deceased full name, DOB, SSN
3. Decedent's full name, alias, gender DOB, SSN, address
4. National Death Index certificate numbers and year of certificate

### Linking to Outside Data Sources

If a link/match is to be made to New York State Department of Health (SDOH) administrative databases (e.g., the Statewide Planning and Research Cooperative System (SPARCS)), matching and linkage may be performed by SDOH Office of Quality and Patient Safety (OQPS) with DOHMH permission.

## DATA CONTENTS

### Partial List of Variables

*For birth records*: prenatal care, parents' age at birth, maternal morbidity, parents' demographics, type of place of birth, insurance coverage, characteristics of labor and delivery, congenital anomalies

*For death records:* cause of death, age at death, deceased's demographics, hospice care, and place of death

### Subjects/Tags

Natility; mortality; births; deaths

*Randomized Evaluations Using This Data Set*

Gelber, Alexander, Adam Issen and Judd B. Kessler. 2014. "The Effects of Youth Employment - Evidence from New York City Summer Youth Employment Program Lotteries." Working Paper, December 2014. J-PAL Evaluation Page.

# NEW YORK STATE INCARCERATION DATA

## New York State Department of Corrections and Community Supervision (DOCCS)

*Information on individuals over the age of 18 who are or have been incarcerated in a New York State prison.*

## ACCESS

Data are available by request to professional researchers and doctoral students. Researchers must contact the DOCCS Division of Program Planning, Research, and Evaluation. Researchers must submit an application for data access which includes the title of the study, contact information for all research staff, an endorsement by a recognized research organization, evidence of IRB approval, and justification for the research. In addition, DOCCS requires researchers to submit a detailed research design document including the DOCCS resources and personnel that may be needed for the study, the criteria and procedures for selection of subject or records for the research, and a data security plan. If the request is approved, researchers send datasets to DOCCS to be matched with the incarceration data. DOCCS will perform the match, remove any personally identifiable information, and return de-identified data to the researcher. DOCCS will not release identified information.

### Timeline for Access

Data are updated and reported annually. Some variables are updated more frequently.

The timeline for approving a research request depends on the complexity of the proposal, the variables requested, and the number of DOCCS offices which need to approve the request. Research proposals are reviewed in February, May, and October.

After approval, the timeline for data access and matching depends on the complexity of the request and DOCCS staff capacity.

*Lag Time*
Unknown.

**Links**
Department of Corrections and Community Supervision

**Unit of Observation** Individual incarceration episode

**Personally Identifiable Information Available for Linking?** Yes

**Geography** NY, USA

**Years Available** Unknown

**Free or Paid** Free

**Frequency of Updates** Annual. Lag time is unknown. Some data are updated more frequently.

**Universe** All individuals incarcerated in a New York State prison except youthful offenders (i.e. 18 or younger at the time of the offense), those who have had their convictions reversed by a court, and certain previously incarcerated non-violent offenders who are covered by a special provision which removes information on incarceration episodes for relatively minor crimes.

## COST

Access to records is free.

## LINKING

Researchers must submit a list of individuals for whom records are requested to DOCCS. The specific procedure and required variables for linking are evaluated on a case-by-case basis. DOCCS will strip personally identifiable information from the resulting file. It is unknown whether DOCCS will permit the use of a de-identified "Study ID" for linking with additional data sets.

### Personally Identifiable Information Available for Linking

- First and last name
- Date of birth
- Social Security number
- DOCCS Department Identification number (DIN)

## DATA CONTENTS

### Partial List of Variables

Date of admission, date of release, community supervision participation

### Subjects/Tags

Crime; incarceration, criminal justice, New York State

*Randomized Evaluations Using This Data Set*

Gelber, Alexander, Adam Issen and Judd B. Kessler. 2014. "The Effects of Youth Employment - Evidence from New York City Summer Youth Employment Program Lotteries." Working Paper, December 2014. J-PAL evaluation page.

# SOCIAL SECURITY ADMINISTRATION DATA

## Social Security Administration (SSA)

*Earnings and benefits data from the Social Security Administration. Includes information on applications for Social Security numbers; annual earnings; and receipt of old age, survivor and disability insurance (OASDI) and Supplemental Security Income (SSI).*

## ACCESS

Disclosure of identified data outside of the Social Security Administration is limited by various levels of legislation and policy. With a few exceptions, identified data are available only to SSA researchers. SSA may also disclose identified data to a federal, state, or Congressional support agency for research, evaluation, or statistical studies, and other exceptions may be made for epidemiological research. When SSA researchers collaborate with non-SSA co-investigators, only the SSA affiliate may access the identified data.

The agency also produces public-use microdata files (non-identified, and non-linkable to individuals) that are available to outside researchers.

### Timeline for Access

Unknown.

**Links**
SSA data documentation

Master Earnings File documentation

Requesting a data exchange

Epidemiological research

**Unit of Observation** Individual, claim

**Personally Identifiable Information Available for Linking?** Yes

**Geography** USA

**Years Available** Depends on variable; at least 1978-present

**Free or Paid** Paid

**Frequency of Updates** Unknown

**Universe** Depends on file. Includes all SSN applicants, OASDI applicants, and SSI recipients. Includes rejected applicants.

## COST

Fees are charged on a cost-reimbursable basis. Further estimates or cost algorithms are not known.

## LINKING

The linking process for SSA files is unknown.

### Personally Identifiable Information Available for Linking

The SSA does not list specific requirements for linking to individuals' outcomes. Previous researchers have used name and date of birth to match records, achieving an 85% match rate on a sample of 75,000 individuals (Baicker et al, 2014). The SSA data set includes at least the following identifiers:

- First and last name
- Date of birth
- Social Security number

### Linking to Outside Data Sources

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

## DATA CONTENTS

### Partial List of Variables

*Numident file*: name, date and place of birth, parents' names, date of death.

*Master earnings file*: wages, self-employment earnings, annual total wages (1978 to present), annual self-employment earnings, annual earnings used for OASDI contributions (1951 to present), report year

*Master beneficiary record:* primary worker's SSN, beneficiary's own SSN, benefit application date, disposition of application (approved, denied, etc), benefit entitlement date, type of benefit, amount of benefit

*Supplemental security record:* date of claim, citizenship status, income, resources, eligibility code, payment code, and payment amount

### Subjects/Tags

Social services, income, disability, social security, Social Security Disability Insurance, SSDI, social security number, SSN, OASDI

*Randomized Evaluations Using this Data Set*

Baicker, Katherine, Amy Finkelstein, Jae Song*, and Sarah Taubman. 2014. "The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment." American Economic Review, 104(5): 322-28. J-PAL Evaluation Page.

Data and techniques from this evaluation are described in the online appendix.

*SSA researcher

# SOUTH CAROLINA HEALTH UTILIZATION DATA

## South Carolina Revenue and Fiscal Affairs Office (RFA)

*Clinical care and patient records data from licensed general acute care hospitals, emergency departments, and ambulatory care surgery centers in South Carolina. Includes inpatient discharge and outpatient encounter information.*

## ACCESS

Encounter-Level data elements are available for general public release subject to an application and a data use agreement. Restricted data elements include exact dates or times, patient ages, and zip codes. Researchers must receive approval by the researcher's Institutional Review Board (IRB), approval for release through South Carolina's Data Oversight Council (DOC), and must complete an application and confidentiality contract.

Personally identifiable information such as patient name, address, and Social Security number are considered confidential, and thus releasable only if a mandate has been established by statutory law, or never releasable. However, these data elements may be used for statistical linking performed by RFA.

### Timeline for Access

The time needed to access these data will vary depending on the security-level of the data and the data use agreements that need to be put into place. The linking process has been standardized and can usually be done quickly, within 1-2 weeks.

**Links**
SC Data Oversight Council

**Unit of Observation** Encounter

**Personally Identifiable Information Available for Linking?** Yes

**Geography** SC, USA

**Years Available** 1996-present

**Free or Paid** Paid

**Frequency of Updates** Monthly with an approximately 6-month lag

**Universe** Encounter visit records from licensed general acute care hospitals, emergency departments, and ambulatory care surgery centers in South Carolina.

### *Lag Time*

RFA receives billing data on a monthly basis from hospitals throughout the state. The lag time for these data is six months for complete data, and four months for partially processed data.

## COST

The RFA charges for the release of data based on a cost recovery basis (i.e. to cover the cost of staff time). That cost is approximately $1.25 per 1,000 (encounter-level) records.

## LINKING

Researchers may send RFA a finder file with a list of individuals for whom records are requested. Researchers may include a de-identified "Study ID" in the data set and request that RFA include the Study ID in the linked data. RFA will use these identifiers and a probabilistic matching strategy to link study participants to existing records in the state's database.

### Personally Identifiable Information Available for Linking

- Social Security number
- First name, middle initial, last name
- Date of birth

### Linking to Outside Data Sources

RFA maintains an integrated data system, including information about clients' use of programs and services of various state, private and non-profit entities. This includes data on legal and safety services; social services; behavioral health; child care; education; disease registries; Medicare; Medicaid; state employee health services; free clinics; community centers and homelessness. While RFA maintains these data and is able to link these data sets, each originating agency maintains control of its own data and dictates how they may be released. Many require additional, separate data use agreements, DOC approval, and IRB approval. The system enables researchers to analyze the use of services and crossover by clients among these entities.

Researchers who are interested in linking health utilization data to other data sets must first request permission from the DOC.

## DATA CONTENTS

Lists of variables are available online: Encounter-Level, Restricted, Confidential, Never Releasable

### Partial List of Variables

Patient demographics, diagnosis codes, procedure codes, admission and discharge dates, hospital identifiers, charges for services by revenue center, length of stay, patient disposition

### Subjects/Tags

Health; healthcare; health utilization; all-payer claims, encounters; acute care hospitals; emergency departments; ambulatory care surgery

*Randomized Evaluations Using This Data Set*
Unknown.

*Other Documentation*
Health and Demographics, South Carolina Revenue and Fiscal Affairs Office

The Circle of Love: South Carolina's Integrated Data System (note: Effective July 1, 2014, the Office of Research and Statistics became part of an independent agency operating as the Revenue and Fiscal Affairs Office.)

# TRANSUNION CONSUMER CREDIT REPORTS

## TransUnion

*Comprehensive consumer credit and borrowing data gleaned from public records, collection agencies, and trade lines, such as credit card balances, auto loans and mortgages. Includes credit balances, credit scores, unpaid bills, and bankruptcy.*

## ACCESS

Data are available to researchers for purchase. Researchers must contact an account manager from TransUnion to request data and to negotiate details of the data transfer. A data sharing agreement is required.

For certain types of data, TransUnion may require that the researcher's organization is certified as a "trusted financial entity" that can receive credit data. This may require an in-person site visit by TransUnion staff.

If all personally identifiable information (PII) is removed from records prior to delivery to researchers, no hard or soft inquiries will be posted to consumers' files as a result of extracting their credit report data.

### Timeline for Access

Unknown.

**Links**
www.transunion.com

**Unit of Observation** Individual and trade-level (balances and payments for specific accounts)

**Personally Identifiable Information Available for Linking?** Yes

**Geography** USA

**Years Available** Unknown

**Free or Paid** Paid

**Frequency of Updates** Unknown

**Universe** US residents with a credit score (may be cut at a minimum score threshold)

## COST

Cost depends on the number of records requested as well as the types of variables and level of detail requested. Cost seems to be negotiable, but significant (i.e. thousands of dollars per file). Price algorithms are not publicly available.

There are two types of files: standard attributes files with aggregate data and "raw" files with finer data and date. Raw files are much more expensive. Researchers may also request custom sets of attributes.

## LINKING

Researchers may send a finder file with a list of individuals for whom records are requested. Researchers may include a de-identified "Study ID" in the data set and request that TransUnion include the Study ID in the linked data. TransUnion will match these individuals and send back a file stripped of PII. To our knowledge, TransUnion is not willing to transmit an identified data set.

### Personally Identifiable Information Available for Linking

TransUnion will link on any of the following:

- Name
- Address
- Social Security number

TransUnion will not use date of birth or other data elements for linking, to our knowledge.

## DATA CONTENTS

### Partial List of Variables

Credit balances, credit card balance, credit limit, credit score, reason for credit score, unpaid bills, debts sent to collection agencies, tax liens, judgments, bankruptcy, medical bills

### Subjects/Tags

Credit; consumption; proxy for consumption; debt; bankruptcy; financial; finance; consumer finance

*Randomized Evaluations Using this Data Set*

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group, "The Oregon Health Insurance Experiment: Evidence from the First Year", Quarterly Journal of Economics, 2012 Aug; 127(3): 1057-1106. J-PAL Evaluation Page.

*Other Research Using this Data Set*

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo. "The Economic Consequences of Hospital Admissions for Individuals with Health Insurance." (in progress).

*Other Documentation*

Julia Brown, Lucia Goin, Nora Gregory, Katherine Hoffmann, and Kim Smith. "Evaluating Financial Products and Services in the US." December 2015. (See pages 53-56)

# US DEPARTMENT OF VETERANS AFFAIRS DATA

## US Department of Veterans Affairs (VA)

*Clinical care and patient records data from the Department of Veterans Affairs. Includes information on beneficiaries, inpatient and outpatient claims, radiology and laboratory tests and results, cost information, and Medicare and Medicaid Services (CMS) data.*

## ACCESS

The VA restricts release and use of data containing Protected Health Information (PHI) for research use to VA employees with approved VA research protocols. For non-VA researchers, the options for using these data are to collaborate with a VA researcher, or obtain HIPAA Authorization from each individual veteran to use their identified data. HIPAA Authorization is similar to, but distinct from, the concept of informed consent. For more information, see Using Administrative Data for Randomized Evaluations.

Publicly available documentation is limited. For researchers with a VA-approved project, a video and PDF handout describe the data request process.

The Veterans Benefits Administration (VBA) provides free access to aggregated summary statistics on weekly claims processed.

### Timeline for Access

Unknown.

**Links**
Overview of data sets, research uses, access instructions

**Unit of Observation** Individual beneficiary; claim

**Personally Identifiable Information Available for Linking?** Yes

**Geography** USA

**Years Available** Unknown

**Free or Paid** Unknown

**Frequency of Updates** Unknown

**Universe** US veterans who utilize VA medical services

## COST

Unknown.

## LINKING

VA data contain Social Security numbers. Options and processes for linking and other available identifiers are unknown.

## DATA CONTENTS

### Partial List of Variables

Demographic information; type of claim; claim procedure and diagnostic codes; ICD-9; ICD-10; CPT4; location of claim; lab test data: test; cost; code; results; radiology test data; prescriptions; vital statistics: date of death; Medicare data; Medicaid data

### Subjects/Tags

Health; health utilization

*Randomized Evaluations Using This Data Set*

Doyle, Joseph J., Ewer, Steven M., and Todd H Wagner*. 2010. "Returns to Physician Human Capital: Evidence from Patients Randomized to Physicians Teams." Journal of Health Economics, 29(6): 866-882. J-PAL Evaluation Page.

*VA researcher

*Other Documentation*

Sohn, M.-W., Arnold, N., Maynard, C., Hynes, D.M., 2006. "Accuracy and completeness of mortality data in the Department of Veterans Affairs." Population Health Metrics 4, b14.