

VA EDH Data Curation Documentation FY22-Q2, Rev. 2



Blair Christian
Hilda Klasky
Kevin Sparks
Joe Tuccillo
Deeksha Rastogi
Marcia Branstetter
Matthew Whitehead
Alec Hamaker
Rochelle Watson

March 2022

DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via OSTI.GOV.

Website www.osti.gov

Reports produced before January 1, 1996, may be purchased by members of the public from the following source:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone 703-605-6000 (1-800-553-6847)
TDD 703-487-4639
Fax 703-605-6900
E-mail info@ntis.gov
Website <http://classic.ntis.gov/>

Reports are available to US Department of Energy (DOE) employees, DOE contractors, Energy Technology Data Exchange representatives, and International Nuclear Information System representatives from the following source:

Office of Scientific and Technical Information
PO Box 62
Oak Ridge, TN 37831
Telephone 865-576-8401
Fax 865-576-5728
E-mail reports@osti.gov
Website <https://www.osti.gov/>

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Computational Sciences & Engineering Division

VA EDH DATA CURATION DOCUMENTATION FY22-Q2, REV. 2

Blair Christian
Hilda Klasky
Kevin Sparks
Joe Tuccillo
Deeksha Rastogi
Marcia Branstetter
Matthew Whitehead
Alec Hamaker
Rochelle Watson

March 2022

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CHANGE HISTORY

Rev No.	Change	Performed by:	Date
2	1. A co-author's last name has been corrected from 'Brandstetter' to 'Branstetter'.	Hilda B. Klasky	July, 2022
2	2. Fixed typos in sections 2, 3 and 7.	Hilda B. Klasky	July, 2022

CONTENTS

CHANGE HISTORY	iii
CONTENTS	iii
TABLES	v
1. Introduction.....	1
1.1 Background	1
2. Documentation Overview	1
2.1 Previous Document Releases	1
2.2 Recommended Citation For FY22-Q2 Dataset Package.....	2
3. Contents and Structure.....	2
3.1 Dataset Curation Documentation Standard Format	2
3.2 Dataset Conventions	3
4. Eviction Rates by County	5
4.1 Sponsor	5
4.2 Description	5
4.3 Inclusion.....	5
4.4 Resources	5
4.5 Update Frequency	5
5. American Community Survey Income Inequality Measures Based on Income to Poverty Ratio by Census Block Group	7
5.1 Sponsor	7
5.2 Description	7
5.3 Inclusion.....	7
5.4 Resources	7
5.5 Update Frequency	7
6. Individual-Oriented Social Vulnerability Index, Census Block Groups.....	9
6.1 Sponsor	9
6.2 Description	9
6.3 Inclusion.....	9
6.4 Resources	10
6.5 Update Frequency	10
7. National Instant Criminal Background Check System (NICS)	11
7.1 Sponsor	11
7.2 Description	11
7.3 Inclusion.....	11
7.4 Resources	11
7.5 Update Frequency	11

TABLES

Table 1. Eviction Rates by County (EVICTIONS_CTY).....	6
Table 2. Variable Availability Across Years, EVICTIONS_CTY	6
Table 3. American Community Survey Income Inequality Measures Based on Income to Poverty Ratio by Census Block Group (INCOME_INEQUALITY_IPR_BG)	8
Table 4. Variable availability across years, INCOME_INEQUALITY_IPR_BG.....	8
Table 5. Individual-Oriented Social Vulnerability Index, Census Block Groups (IOSVI_BG).....	10
Table 6. Variable availability across years, IOSVI_BG	10
Table 7. National Instant Criminal Background Check System (NICS) (NICS).....	11
Table 8. Variable availability across years, NICS	12

1. INTRODUCTION

The health and well-being of the Nation's men and women who have served in uniform is the highest priority for the U.S. Department of Veterans Affairs (VA). VA is committed to providing timely access to high-quality, recovery-oriented, evidence-based mental health care that anticipates and responds to Veterans' needs and supports the reintegration of returning Service members into their communities. VA is working to eliminate suicide among all Veterans by developing and implementing innovative suicide prevention approaches and resources.

Health outcomes, such as suicide are typically modeled as a function of genetics and environment, where environment refers to factors beyond medical, e.g., air quality, access to transportation and food, homelessness status, etc. Mental health outcomes for each individual are considered to be associated with multiple stressors that fall under a variety of categories socioeconomic, economic, physical environment. Understanding the relationships between these stressors, covariates and health outcomes, requires curated, standardized data that can be input into the VA's Recovery Engagement and Coordination for Health - Veterans Enhanced Treatment (REACH VET) or other health outcomes model.

Environmental Determinants of Health (EDH) as defined by the World Health Organization (WHO) is clean air, stable climate, adequate water, sanitation and hygiene, safe use of chemicals, protection from radiation, healthy and safe workplaces, sound agricultural practices, health-supportive cities and built environments, and a preserved nature are all prerequisites for good health.

1.1 BACKGROUND

With funding from the VA Office of Mental Health and Suicide Prevention (OMHSP), the EDH project has developed novel datasets associated with select health outcomes, a methodology for converting spatiotemporal data from one spatial reference (such as a 1km grid) to another (such as US Census Tracts), and health outcomes modeling capabilities. The datasets are an advancement to the Agency for Healthcare Research and Quality (AHRQ) Social Determinants of Health (SDoH) covariates as key gaps are addressed, a finer spatial resolution (Census Tract), and environmental covariates are included. The process of curating and standardizing these datasets is non-trivial, as they are often measured at different spatial and temporal resolutions and have different spatial and temporal granularities. For example, the US Census data products typically use census blocks, block groups, or counties, whereas air pollutants from the EPA and weather data are available on 1km grids, and some economic data may be available only at a zip code level. In this context, standardized refers to the datasets all being at the same scale of spatial extent (e.g., US Census Tract and/or County), and curated refers both to a process that is repeatable, has data provenance, and which uses appropriate methodologies for converting covariates. The data contained in the EDH dataset are drawn from multiple sources, and variables may have differing degrees of availability, patterns of missing data, and methodological considerations across sources, geographies, and years.

2. DOCUMENTATION OVERVIEW

This data source documentation report provides researchers with information on the structure and contents of the datasets, as well as descriptions of the data sources utilized to populate the data files. This document covers FY22-Q2 dataset curation documentation revisions.

2.1 PREVIOUS DOCUMENT RELEASES

Among previous releases are the following. Please consult the following source for FY21 dataset curation documentation: Christian, J.B., Branstetter, M, Klasky, H.B., Tuccillo, J., Sparks, K., Rastogi, D.,

Watson, R., Yoon, H-J., Kim, Y., VA EDH Data Curation Documentation - FY 2021, ORNL/SPR-2021/2366 - Pub ID 170648. 2021.

For the EDH Data Curation Documentation delivered on FY22-Q1, which contains the following datasets:

- Social Capital Index Dataset (2019 - updated)
- Social Vulnerability Index Dataset (2018)
- Block Group Area Deprivation Index Dataset for Washington, DC (2019)
- Low Food Access Area Dataset for Washington, DC (2017)

please refer to: Christian, J.B., Klasky, H.B., Sparks, K., Peluso, A., Tuccillo, J., Devineni, P., and Watson, R. VA EDH Data Curation Documentation - FY22-Q1, ORNL/SPR-2022/2316- Pub ID 172755. 2022.

2.2 RECOMMENDED CITATION FOR FY22-Q2 DATASET PACKAGE

Christian, J.B., Klasky, H.B., Sparks, K., Peluso, A., Tuccillo, J., Rastogi, D., Branstetter, M., Whitehead, M., Hamaker, A., and Watson, R., VA EDH Data Curation Documentation - FY22-Q2, ORNL/SPR-2022/2391 - Pub ID 174092. 2022.

3. CONTENTS AND STRUCTURE

The data in the EDH datasets were gathered from a variety of publicly available data sources. In FY22-Q2 release the data sources included follow:

- Eviction Rates (by county)
- Income Inequality (American Community Survey Income Inequality Measures Based on Income to Poverty Ratio by Census Block Group)
- Individual-Oriented Social Vulnerability Index (IOSVI), Census Block Groups, and
- National Instant Criminal Background Check System (NICS), Lethal Means Access.

3.1 DATASET CURATION DOCUMENTATION STANDARD FORMAT

Each data source description follows a standard format with the next fields:

- Sponsor (name of the organization that provided the raw data, e.g., Health Resources and Services Administration [HRSA] for the Area Health Resources Files [AHRF])
- Description (brief, general description of the data)
- Inclusion in the EDH datasets
- Lists the SDOH domains to which the data source has contributed variables

- Includes additional information about the data source relevant to the EDH dataset
- Resources (links to original data source documentation, data download sites, and other relevant information)
- Update frequency: how often is each dataset going to be updated.
- EDH variable definitions and specifications (in tabular format)
 - Variable name
 - Variable label
 - Source table, if multiple data tables were available from the original data source
 - Numerator (for derived variables)
 - Denominator (for derived variables) or original variable (when renamed for the EDH dataset)
- EDH variable availability across years (in tabular format)
 - Variable name
 - Variable label
 - Data year availability (e.g. 2009 to 2018)

3.2 DATASET CONVENTIONS

In regard to the datasets' versioning, Microsoft SQL Server database system is used to supply the datasets. A table is used to hold each dataset. The following schema names for the quarterly releases are (or will be) included in the database: OMHSP_FY22Q1, OMHSP_FY22Q2, OMHSP_FY22Q3, OMHSP_FY22Q4, OMHSP_FY23Q1,..., and so on. These will aid in distinguishing between releases when we deliver the same data set from one release to the next.

Variables in the EDH dataset were created from these several data sources in one of two ways:

- Drawn directly from the original data source. When the data was available from the data source as needed, we renamed the original variables for clarity and consistency across years, and to fit the naming conventions of the SDOH beta data files.
- Derived using data from the original data source. For some data sources, it was necessary to calculate percentages or rates for inclusion in the beta data files. The numerators and denominators for the variables and their sources are shown following each data source description.

The following conventions were followed in constructing the EDH datasets to provide researchers with a consistent and easy-to-use resource:

- Variable assignment to annual datasets. Variables appear in the annual datasets that correspond with (1) the single year represented by the original data source (e.g., Nursing Home Compare data for facilities in 2016 appears in the 2016 county dataset), or (2) the last year in a period represented by the data (e.g., American Community Survey data aggregated over 2012 to 2016 is in the 2016 dataset).
- Variable availability. The availability of each variable changes across data years. Following each data source description in this report is a table showing the availability of each variable in the annual datasets.
- Variable naming. Except for the geographic ID variables, all variable names begin with a data source acronym followed by an underscore and a descriptive title.

- Missing values. The datasets use a blank to denote a missing value, almost exclusively. The one exception is the provider ratio variables from the County Health Rankings (CHR) data, which have negative values for counties where the number of providers is zero. This is described further in the description of the CHR data.

Detailed information about each data source is included in the following sections of this report.

4. EVICTION RATES BY COUNTY

4.1 SPONSOR

Eviction Lab

4.2 DESCRIPTION

Eviction rates 2000 - 2016 for United States Counties.

4.3 INCLUSION

Eviction rates 2000 - 2016 for United States Counties with selected socioeconomic characteristics.

This research uses data from The Eviction Lab at Princeton University, a project directed by Matthew Desmond and designed by Ashley Gromis, Lavar Edmonds, James Hendrickson, Katie Krywokulski, Lillian Leung, and Adam Porton. The Eviction Lab is funded by the JPB, Gates, and Ford Foundations as well as the Chan Zuckerberg Initiative. More information is found at evictionlab.org.

From Eviction Lab: “The data we collected is comprised of formal eviction records from 48 states and the District of Columbia. Eviction records include information related to an eviction court case, such as defendant and plaintiff names, the defendant's address, monetary judgment information, and an outcome for the case. We combined these records with demographic information from the Census to paint a better picture of the areas in which these evictions are happening.”

The data we collected is comprised of formal eviction records from 48 states and the District of Columbia. Eviction records include information related to an eviction court case, such as defendant and plaintiff names, the defendant's address, monetary judgment information, and an outcome for the case. We combined these records with demographic information from the Census to paint a better picture of the areas in which these evictions are happening.

Data from Eviction Lab were acquired from Amazon Web Services using the R `aws.S3` library. Eviction Lab is an ongoing effort, and as such all United States counties have not been validated. Missing (unvalidated) states include ID, MD, MS, NH, NJ, WV, and DC.

4.4 RESOURCES

Eviction Lab: <https://evictionlab.org/>

Methods and Validation FAQ: <https://evictionlab.org/methods/#validation>

Data Dictionary: https://evictionlab.org/docs/DATA_DICTIONARY.txt

4.5 UPDATE FREQUENCY

Every fiscal year, or as requested by the sponsor, this dataset will be updated and distributed. Minimal quarterly updates may be necessary to correct minor data inaccuracies.

Table 1. Eviction Rates by County (EVICTIONS_CTY)

variable name	variable label
FIPS	County FIPS code.
year	Year of the record.
evictions	Number of eviction judgments in which renters were ordered to leave in a given area and year.
eviction_filings	All eviction cases filed in an area, including multiple cases filed against the same address in the same year.
eviction_rate	Share of renter-occupied households that received an eviction judgment in which renters were ordered to leave.
eviction_filing_rate	Ratio of number of evictions filed to renter-occupied homes.
low_flag	Unknown.
imputed	Flag for imputation of count of renter-occupied households and number of renters.
subbed	Flag for substitution of sources count of renter-occupied households/number of renters was derived from a source other than Eviction Lab.

Table 2. Variable Availability Across Years, EVICTIONS_CTY

variable name	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
FIPS	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
year	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
evictions	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
eviction_filings	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
eviction_rate	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
eviction_filing_rate	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
low_flag	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
imputed	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
subbed	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

5. AMERICAN COMMUNITY SURVEY INCOME INEQUALITY MEASURES BASED ON INCOME TO POVERTY RATIO BY CENSUS BLOCK GROUP

5.1 SPONSOR

United States Census Bureau

5.2 DESCRIPTION

Income inequality measures and descriptive statistics based on the income to poverty ratio (IPR) for the last 12 months, including the Gini index, majority IPR class, and population shares by IPR class. Contrasted with the American Community Survey's default (quintile-based) representation of income inequality, the IPR is adjusted for differences in living arrangement status (i.e., families with children, couple households, non-family/living alone).

5.3 INCLUSION

Data was acquired from the American Community Survey 2015 - 2019 5-year Summary File and is available for all populated census block groups in the United States, including Puerto Rico. Population share by IPR class were computed by normalizing count estimates by the total population for whom poverty status is determined (all adults not in group quarters living arrangements and all related children based on householder income). The Gini index and majority IPR class were computed based on shares of each IPR class.

5.4 RESOURCES

The source table utilized for this dataset is C17021: Ratio of Income to Poverty Level in the Past 12 Months.

For more information on the ACS:

- Main:
<https://www.census.gov/programs-surveys/acs>
- Data:
https://www2.census.gov/programs-surveys/acs/summary_file/2018/data/?#
(<https://www.census.gov/topics/income-poverty/income-inequality/data/data-tables/acs-data-tables.html>)[<https://www.census.gov/topics/income-poverty/income-inequality/data/data-tables/acs-data-tables.html>]
- Measurement of Poverty Status:
(<https://www.census.gov/topics/income-poverty/poverty/about.html>)[<https://www.census.gov/topics/income-poverty/poverty/about.html>]

5.5 UPDATE FREQUENCY

Every fiscal year, or as requested by the sponsor, this dataset will be updated and distributed. Minimal quarterly updates may be necessary to correct minor data inaccuracies.

Table 3. American Community Survey Income Inequality Measures Based on Income to Poverty Ratio by Census Block Group (INCOME_INEQUALITY_IPR_BG)

variable name	variable label	source table
FIPS	US Census Block Group FIPS code	NA
gini	Gini index of income inequality	C17021: Ratio of Income to Poverty Level in the Past 12 Months
maj	Majority income to poverty ratio class	C17021: Ratio of Income to Poverty Level in the Past 12 Months
incpov_L50	Share of population with income less than 50% of poverty threshold	NA
incpov_50_99	Share of population with income 50% - 99% of poverty threshold	NA
incpov_100_124	Share of population with income 100% - 124% of poverty threshold	NA
incpov_125_149	Share of population with income 125% - 149% of poverty threshold	NA
incpov_150_184	Share of population with income 150% - 184% of poverty threshold	NA
incpov_185_199	Share of population with income 185% - 199% of poverty threshold	NA
incpov_GE200	Share of population with income 200% of poverty threshold or greater	NA

Table 4. Variable availability across years, INCOME_INEQUALITY_IPR_BG

variable name	2019
FIPS	X
gini	X
maj	X
incpov_L50	X
incpov_50_99	X
incpov_100_124	X
incpov_125_149	X
incpov_150_184	X
incpov_185_199	X
incpov_GE200	X

6. INDIVIDUAL-ORIENTED SOCIAL VULNERABILITY INDEX, CENSUS BLOCK GROUPS

6.1 SPONSOR

United States Census Bureau

6.2 DESCRIPTION

Our team at ORL computed an individually-oriented social vulnerability index (IOSVI) adapted from the Centers for Disease Control and Prevention (CDC) SVI. Contrasted with the CDC SVI, IOSVI takes a bottom-up approach, first by computing an individual vulnerability index (IVI) from the American Community Survey's (ACS) Public-Use Microdata Sample (PUMS), then computing the SVI based on aggregate IVI characteristics at the census block group level (social areas of roughly 600 - 3000 people) to compute the SVI. Block group-level IVI estimates were generated using synthetic population produced from the American Community Survey and Public-Use Microdata Sample via the UrbanPop project at Oak Ridge National Laboratory.

6.3 INCLUSION

IOSVI was computed on the pooled PUMS sample from the ACS 2015 - 2019 5-Year Estimates for all Public-Use Microdata Areas (PUMAs) in the United States and Puerto Rico.

IVI was constructed from the PUMS by 1) separating vulnerability indicators at the census tract level into four thematic domains used to compute the CDC SVI (Socioeconomic Status, Household Composition and Disability, Minority Status and Language, Housing Type and Transportation); 2) tabulating the number of vulnerability indicators describing every unique person-level record in the pooled PUMS for the United States; 3) standardizing the indicator counts by domain with percentile ranking (to control for uneven numbers of indicators per domain); 4) combining the domain scores into a final index based on the mean for each individual profile.

Modifications of the CDC vulnerability indicators for IVI included A) removing per-capita income, an area-level variable; B) expanding single-parent households to include any single-adult guardian relationship; C) expanding limited English ability to include any member of a linguistically-isolated household according to the PUMS.

SVI was constructed by 1) linking the IVI scores for unique individual profiles back to each PUMS response 2) joining IVI scores to UrbanPop block group-level synthetic population estimates based on PUMS household serial number and person order 3) for each block group, computing the cumulative proportion of the population at each unique IVI level in ascending order 4) computing SVI as the difference in area under curve (AUC) between a) the cumulative distribution of a hypothetical population consisting entirely of people with IVI = 0 and b) the estimated cumulative distribution of IVI scores. For each block group, the final SVI is a Monte Carlo estimate (average) of indices across 30 realizations of the synthetic population produced by UrbanPop.

Using the IOSVI approach, it is possible to compute a "high-low" index break based on the observed cumulative distribution of IVI scores for the entire United States. In this case, the break value occurs at SVI ~ 0.51.

6.4 RESOURCES

- American Community Survey: <https://www.census.gov/programs-surveys/acs>
- UrbanPop: <https://acsdatacommunity.prb.org/m/2021-acs-conference-files/161/>
- CDC Social Vulnerability Index:
https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html

6.5 UPDATE FREQUENCY

Every fiscal year, or as requested by the sponsor, this dataset will be updated and distributed. Minimal quarterly updates may be necessary to correct minor data inaccuracies.

Table 5. Individual-Oriented Social Vulnerability Index, Census Block Groups (IOSVI_BG)

variable name	variable label
FIPS	Block Group FIPS code.
svi	Social Vulnerability Index score.
se	Monte Carlo Error.

Table 6. Variable availability across years, IOSVI_BG

variable name	2019
FIPS	X
svi	X
se	X

7. NATIONAL INSTANT CRIMINAL BACKGROUND CHECK SYSTEM (NICS)

7.1 SPONSOR

Federal Bureau of Investigation

7.2 DESCRIPTION

Record of the number of permits and firearm transactions from 1998 to present by state and month.

7.3 INCLUSION

The numbers were aggregated by year.

7.4 RESOURCES

- National Instant Criminal Background Check System (NICS):
<https://raw.githubusercontent.com/BuzzFeedNews/nics-firearm-background-checks/master/data/nics-firearm-background-checks.csv>
- <https://www.fbi.gov/services/cjis/nics>

7.5 UPDATE FREQUENCY

Every fiscal year, or as requested by the sponsor, this dataset will be updated and distributed. Minimal quarterly updates may be necessary to correct minor data inaccuracies.

Table 7. National Instant Criminal Background Check System (NICS) (NICS)

variable name	variable label
FIPS	State level FIPS.
state	State of residence of persons on which background checks are conducted.
permit	Number of permits approved.
handguns	Number of handgun transactions.
totals	Total number of firearm transactions, including handguns, long guns, private sales, returns to seller, firearm rentals.
year_month	The year and month of data collection.

Table 8. Variable availability across years, NICS

variable name	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
FIPS	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
year	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
state	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
perm it	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
hand guns	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
total s	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X