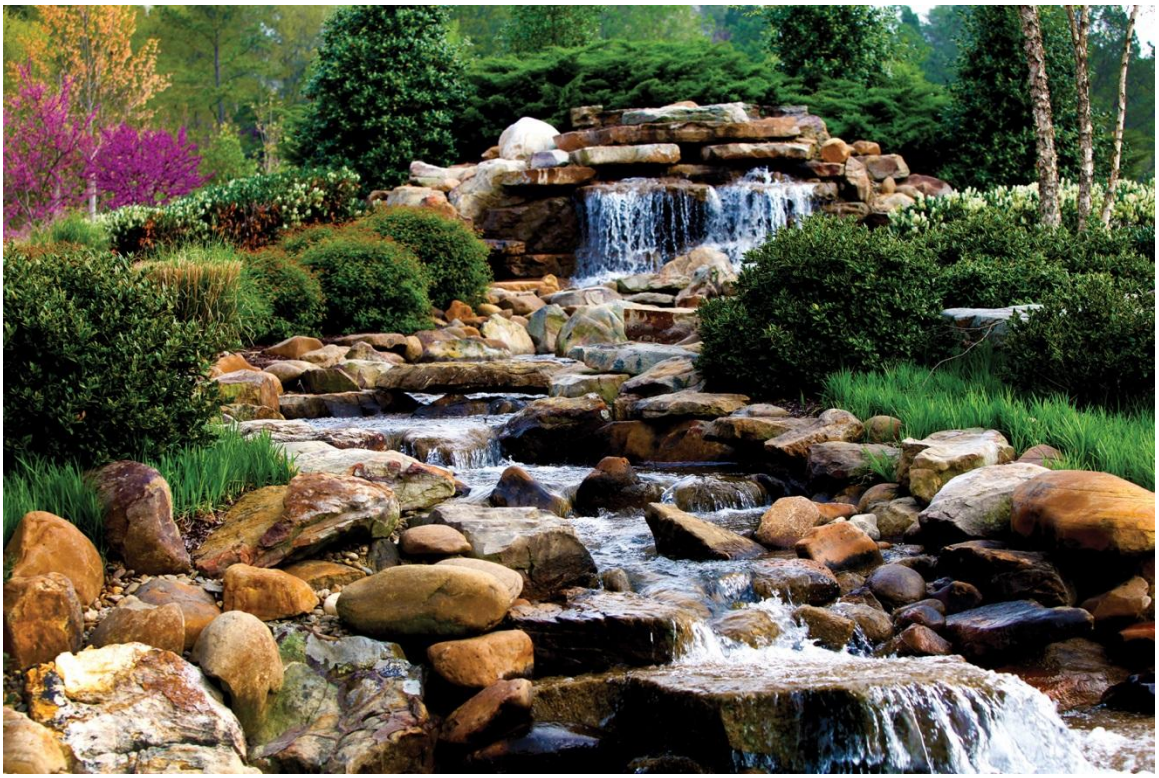
DISCLAIMER

ORNL/SPR-2025/3873
RES PUB ID 234991

VA Determinants of Health Data Curation Documentation FY25-Q3



Hilda Klasky
Kevin Sparks
Alina Peluso
Josh Grant
Joe Tuccillo
Adam Spannaus
Jeremy Logan
Kelly Callaway
Midgie MacFarland
Angela Cunningham
Viswadeep Lebakula
Hope Cook
Heidi Hanson
Susana Martins
Mary J. Bollinger
Matt Boden
Jodie Trafton
Anuj Kapadia

June 2025

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/SPR-2025/3873
RES PUB ID 234991

Computational Sciences and Engineering Division

VA DETERMINANTS OF HEALTH DATA CURATION DOCUMENTATION FY25-Q3

Hilda Klasky
Kevin Sparks
Alina Peluso
Josh Grant
Joe Tuccillo
Adam Spannaus
Jeremy Logan
Kelly Callaway
Midgie MacFarland
Angela Cunningham
Viswadeep Lebakula
Hope Cook
Heidi Hanson
Susana Martins
Mary J. Bollinger
Matt Boden
Jodie Trafton
Anuj Kapadia

June 2025

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

CONTENTS

1. INTRODUCTION AND BACKGROUND

The U.S. Department of Veterans Affairs (VA) places the health and well-being of our nation's veterans as its top priority. VA is dedicated to offering timely access to high-quality, evidence-based mental health care that meets the needs of veterans and supports their reintegration into society. One of our core missions is to prevent suicide among veterans through innovative approaches and resources.

With funding from the VA Office of Mental Health and Suicide Prevention (OMHSP), the Determinants of Health (EDH) project has developed innovative datasets associated with specific health outcomes, a methodology for transforming spatiotemporal data from one spatial reference (e.g., a 1km grid) to another (e.g., US Census Tracts), and capabilities for modeling health outcomes. These datasets represent an enhancement of the Agency for Healthcare Research and Quality (AHRQ), addressing key gaps by introducing finer spatial resolution (Census Tract) and additional geographical covariates into existing data.

The curation and standardization of these datasets is a complex task since they often originate from various sources and are measured at different spatial and temporal resolutions. For example, US Census data products typically use census blocks, block groups, or counties, while data like weather data are available on 1km grids. Some economic data may only be available at the zip code level. In this context, 'standardized' means that all datasets share the same spatial extent (e.g., US Census Tract and/or County), and 'curated' implies a repeatable process with data provenance and the use of appropriate methodologies for covariate conversion.

The Determinants of Health datasets draw from multiple sources, resulting in variables with varying degrees of availability, patterns of missing data, and methodological considerations across different sources, geographies, and years.

2. DOCUMENTATION OVERVIEW

This documentation report provides researchers with detailed insights into the structure, content, and origins of the datasets included in this release. It specifically pertains to the Fiscal Year 2025, Third Quarter (FY25-Q3) data curation efforts for the Determinants of Health project.

2.1 CHANGE LOG

The datasets included in this delivery and documentation are as follows:

1. **Drive time** * *(Update)*
2. **ERA5 Atmospheric Pressure Data** * *(New)*
3. **Federal Firearms Listings County Counts** *(New)*
4. **House Insecurity** *(New)*
5. **SEDoH - Social and Environmental county-level indicators for inclusion in Suicide and Overdose Next-Generation Predictive Modeling (REACH VET and STORM)** * *(New)*

Note: These datasets are exclusively available for the VA's Determinants of Health project. Please contact your program manager for access.

In addition, updates were made to the metadata tables to correct the following issues:

1. **Correction to column_type in OMHSP.column_metadata:** Beginning in FY25Q2, the column_type field in the OMHSP.column_metadata table erroneously listed table names instead of actual data types. This discrepancy has been corrected to reflect accurate data type information.

2. **Resolution of Issue #59 from the High Priority SDoH Data Items.xlsx received on June 18, 2025 – added_date Null Values.**

   In the OMHSP_FY24Q4.homeless_shelterlist_tract_2024 table, the added_date field contained two null values, despite metadata indicating that nulls were not expected. This inconsistency has now been addressed and resolved in the metadata.

3. Corrected an inconsistency in the number of null rows for the updated_date column in OMHSP_FY24Q4.homeless_shelterlist_tract_2024. One null value was present but did not match the count recorded in OMHSP.column_metadata.

2.2 RECOMMENDED CITATION FOR FY25-Q3 DATA CURATION DOCUMENTATION'S SPONSOR REPORT

Klasky, H.B., Sparks, K., Peluso, A., Grant, J., Tuccillo, J., Spannaus, A., Logan, J., Callaway, K., MacFarland, M., Cunningham, A., Lebakula, V., Cook, H., Hanson, H., Martins, S., Bollinger, M., Boden, M., Trafton, J., & Kapadia, A. (2025). VA Determinants of Health Data Curation Documentation FY25-Q3 (ORNL/SPR-2025/3873 RES PUB ID 234991). Oak Ridge National Laboratory.

Please note that the publication of new technical reports remains subject to an active hold by the Office of Scientific and Technical Information (OSTI). At this time, we have not received guidance on when this restriction will be lifted. We remain committed to disseminating this work through the appropriate channels as soon as clearance is granted and will provide updates as additional information becomes available.

This documentation provides a comprehensive understanding of the data and its sources for the specified period, supporting research and analysis within the Determinants of Health project.

2.3 PREVIOUS DOCUMENT RELEASES

Since the inception of the Determinants of Health project, we have delivered multiple releases of datasets along with data curation documentation sponsor reports. These resources are invaluable for researchers seeking to utilize the Determinants of Health data. Below is a list of the previous releases:

1. EDH Data Curation Documentation delivered in FY21 [1]

- Link to Documentation

2. EDH Data Curation Documentation delivered in FY22-Q1 [2]

- Link to Documentation

- Included Datasets:
    - Social Capital Index (*resolution:* county, 2019, *source:* ORNL)
    - Social Vulnerability Index (*resolution:* census tract, 2018, *source:* Centers for Disease Control, Agency for Toxic Substances and Disease Registry)
    - Area Deprivation Index (*resolution:* block group, 2019, *source:* Neighborhood Atlas, University of Wisconsin)
    - Low Food Access (*resolution:* custom geometry, 2017, *source:* Open Data DC)

3. EDH Data Curation Documentation delivered in FY22-Q2 [3]

- Link to Documentation
- Included Datasets:
    - Eviction Rates (*resolution:* county, 2000-2016, *source:* Eviction Lab)
    - Income Variance (*resolution:* block group, 2019, *source:* American Community Survey)
    - Individual-Oriented Social Vulnerability Index (*alternate name:* IOSVI, *resolution:* block group, 2019, *source:* ORNL, Census Bureau)
    - National Instant Criminal Background Check System (*alternate name:* NICS, *resolution:* state, 2022, *source:* Federal Bureau of Investigation)

4. EDH Data Curation Documentation delivered in FY22-Q3 [4]

- Link to Documentation
- Included Datasets:
    - Veteran Population Status (*resolution:* county, 2020, *source:* American Community Survey)
    - Social Connectedness (*resolution:* county, 2021, *source:* Facebook)
    - Small Area Estimates of Housing Characteristics (*resolution:* block group, 2019, *source:* Census Bureau)
    - Internet Access Services (*resolution:* tract, 2019, *source:* Federal Communications Commission)
    - Medicare Part D Opioid Prescription Rates (*resolution:* county, 2019, *source:* Centers for Medicare & Medicaid Services)
    - High Intensity Drug Trafficking Areas (*alternate name:* HIDTA, *resolution:* county, 2018-21, *source:* Washington/Baltimore High Intensity Drug Trafficking Areas Program)

5. EDH Data Curation Documentation delivered in FY22-Q4 [5]

- Link to Documentation
- Included Datasets:
    - Occupational Employment and Wage Statistics (*alternate name:* Mental Health Care Professionals per capita, *resolution:* state, 2021, *source:* Bureau of Labor Statistics)
    - National Survey on Drug Use and Health (*alternate name:* NSDUH, *resolution:* state, 2019, *source:* Substance Abuse and Mental Health Services Administration)
    - National Mental Health Services Survey (*alternate name:* N-MHSS, *resolution:* state, 2018, *source:* Substance Abuse and Mental Health Data Archive)

6. EDH Data Curation Documentation delivered in FY23-Q1 [6]

- Link to Documentation
- Included Datasets:
    - State and Local Policies (Naloxone laws, *resolution:* state, 2017, *source:* Rand) (Good Samaritan laws, *resolution:* state, 2018, *source:* Rand)
    - Area Deprivation Index (*resolution:* block group, 2020, *source:* University of Wisconsin)
    - Opioid Mortality Rate (*resolution:* county, 2014-2018, *source:* OEPS, University of Chicago)
    - Opioid Prescribing Rate (*resolution:* county, 2019, *source:* OEPS, University of Chicago)

7. EDH Data Curation Documentation delivered in FY23-Q2 [7]

- Link to Documentation
- Included Datasets:
    - Total Household Income (*resolution:* county, 2016-2021, *source:* American Community Survey)
    - Medicare Part D Opioid Prescription Rates (update, *resolution:* county, 2013-2020, *source:* Centers for Medicare & Medicaid Services)
    - Poverty (*resolution:* county, 2016-2021, *source:* American Community Survey)
    - Rural Urban Continuum Codes (*resolution:* county, 2013, *source:* Census Bureau, Department of Agriculture)
    - Civil Engagement Atlas (*resolution:* county, 2022, *source:* Social Capital Atlas)

- Cohesiveness Atlas (*resolution:* county, 2022, *source:* Social Capital Atlas)
- Economic Connectedness Atlas (*resolution:* county, 2022, *source:* Social Capital Atlas)
- Local Unemployment (*resolution:* county, 2018-2021, *source:* Bureau of Labor Statistics)

8. EDH Data Curation Documentation delivered in FY23-Q3 [8]

- Link to Documentation
- Included Datasets:
  - Population Weighted Average Elevation (*resolution:* county, 2020, *source:* United States Geological Survey, Jim VanDerslice)
  - Education (*resolution:* county, 2016-2021, *source:* US Census Bureau, American Community Survey)
  - Eviction Rates (update, *resolution:* county, 2016-2021, *source:* The Eviction Lab, Princeton University)
  - Food Insecurity (*resolution:* county, 2010-2021, *source:* Feeding America, US Hunger Relief Organization)

9. EDH Data Curation Documentation delivered in FY23-Q4 [9]

- Link to Documentation
- Included Datasets:
  - National Instant Criminal Background Check System NICS, (*resolution:* state, 2021-2023, *source:* US Federal Bureau of Investigation)
  - Internet Access Services (*resolution:* Census tract, 2021-2022, *source:* US Federal Communications Commission (FCC))

10. EDH Data Curation Documentation delivered in FY24-Q1 [10]

- Link to Documentation
- Included Datasets:
  - ORNL Daily Surface Weather Summaries - Daymet, (*resolution:* county, 2017-2021, *source:* Daymet)
  - Veterans Service Organizations (VSO) (*resolution:* state, 2010-2022, *source:* VSO)
  - Veterans Service Organizations (VSO) (*resolution:* county, 2010-2022, *source:* VSO)
  - Veterans Service Organizations (VSO) (*resolution:* zip code, 2010-2022, *source:* VSO)

11. EDH Data Curation Documentation delivered in FY24-Q2 [11]

- Link to Documentation
- Included Datasets:
    - HUD USPS Zip Code Crosswalk Files (*resolution:* ZIP-to-tract, 2023, *source:* HUD)
    - HUD USPS Zip Code Crosswalk Files (*resolution:* ZIP-to-county, 2023, *source:* HUD)
    - Social Capital Index 2019 (*resolution:* county, 2019 (updated), *source:* research paper https://www.sciencedirect.com/science/article/pii/S0143622823002990)
    - High Intensity Drug Trafficking Areas (HIDTA) (*resolution:* state-level, 2018 - 2021 (re-delivered), *source:* Washington/Baltimore High Intensity Drug Trafficking Areas Program)
    - High Intensity Drug Trafficking Areas (HIDTA) (*resolution:* county-level, 2018 - 2021 (re-delivered), *source:* Washington/Baltimore High Intensity Drug Trafficking Areas Program)

12. EDH Data Curation Documentation delivered in FY24-Q3 [13]

- Link to Documentation
- Included Datasets:
    - High Intensity Drug Trafficking Areas (HIDTA) (*resolution:* state and county, 2022 - 2023, *source:* Washington/Baltimore High Intensity Drug Trafficking Areas Program)
    - Drive-time sample at selected lat/long points in the state of Tennessee (*resolution:* point-level, 2024, *source:* VA and ORNL)

13. EDH Data Curation Documentation delivered in FY24-Q4 [14]

- Link to Documentation
- Included Datasets:
    - Multi-Exposure Environmental Index (MEEI) (*resolution:* census tract level, 2024, *source:* Link to paper)
    - Homeless Shelter List Data (*resolution:* census tract level data, 2024, *source:* ShelterList.com)

14. EDH Data Curation Documentation delivered in FY25-Q1 [15]

- Link to Documentation
- Included Datasets:

- Medicare Part D Opioid Prescribing Rates (*resolution:* county level, 2013-2022, *source:* https://data.cms.gov/)
- Rural-Urban Continuum Codes (*resolution:* county level, 2023, *source:* https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx)
- Drive Time Data (*resolution:* county level, 202, *source:* VA Locations: https://www.va.gov/directory/guide/home.asp , OpenStreetMap: https://www.va.gov/directory/guide/home.asp)
- Food Pantry List (*resolution:* census tract level, 2024, *source:* https://www.foodpantries.org/)
- Food Bank List (*resolution:* census tract level, 2024, *source:* https://www.freefood.org/)
- Homeless Shelter List Data, version 2, (*resolution:* census tract level, 2024, *source:* https://www.shelterlist.com/page/about)

15. EDH Data Curation Documentation delivered in FY25-Q2 [16]
- Link to Documentation (https://www.osti.gov/biblio/xxx)

- Included Datasets:
  - High Intensity Drug Trafficking Areas (HIDTA) Annual Aggregated Counts: Two datasets (*resolution:* county and state level, 2018-2023, *source:* https://www.hidta.org/resources/pmp-data-sharing/)
  - Occupational Employment and Wage Statistics (OEWS) - Mental Health Professionals Per Capita (*resolution:* state level, 2021-2023, *source:* https://www.bls.gov/oes/tables.htm)
  - Bureau of Labor Statistics (BLS) - Unemployment Data (*resolution:* state level, 2018-2023, *source:* https://www.bls.gov/lau/tables.htm)
  - Opioid Dispensing Rate Maps: Two datasets (*resolution:* county and state level, 2019-2023, *source:* https://www.cdc.gov/overdose-prevention/data-research/facts-stats/opioid-dispensing-rate-maps.html)

Please note that the URL for the FY25-Q3 documentation's URL will be provided next delivery.

This comprehensive list allows researchers to access previous releases for reference and analysis, enhancing the utility of the Determinants of Health project's data curation documentation.

3. CONTENTS AND STRUCTURE

3.1 DATASET CURATION DOCUMENTATION STANDARD FORMAT

Each data source description adheres to a standardized format with the following fields:

1. **Source**: The name of the organization that provided the raw data (e.g., Health Resources and Services Administration [HRSA] for the Area Health Resources Files [AHRF]). Note: Prior to the FY23Q4 release, we referred to the source organization as the "sponsor."

2. **Description**: A brief, general description of the data.

    - *Inclusion in the datasets*: Lists the determinants of health domains to which the data source has contributed variables. Includes additional information relevant to the dataset.

3. **Resources**: Links to original data source documentation, data download sites, and other pertinent information.

4. **Update Frequency**: Indicates how often each dataset will be updated.

5. **Variable Definitions and Specifications (in tabular format)**:

    - *Variable name (column name)*

    - *Variable label (optional, if different from the variable or column name)*

    - *Source table (optional, if multiple data tables were available from the original data source)*

    - *Numerator (for derived variables; optional)*

    - *Denominator (for derived variables) or original variable (when renamed for the EDH dataset; optional)*

    - *Total_rows*: Indicates the number of rows in each column within each dataset (Starting in FY23Q2).

    - *Null_rows*: Specifies the count of null rows for each column in each dataset (Starting in FY23Q2).

6. **Variable Availability Across Years (in tabular format)**:

    - *Variable name (column name)*

    - *Data year availability (e.g., 2009 to 2018)*

This standardized format ensures consistency and ease of reference in the curation documentation for each data source.

3.2 DATASET CONVENTIONS

The variables within the Determinants of Health datasets are derived from various data sources through one of two methods:

1. Direct extraction from the original data source: When the data was readily available from the source, we renamed the original variables to ensure clarity and consistency across years, aligning them with the naming conventions of the Determinants of Health data files.

2. Derivation using data from the original data source: In certain cases, we needed to calculate percentages or rates for inclusion in the data files. We provide the numerators and denominators for these variables, along with their respective sources, in the data source descriptions.

To ensure the Determinants of Health datasets serve as a consistent and user-friendly resource for researchers, we adhered to the following conventions:

- **Standardization of all datasets table names:** we follow the convention of <year(s)>, for example: hidta_county_2018_2021.

- **Variable assignment to annual datasets:** Variables appear in the annual datasets corresponding to (1) the single year represented by the original data source (e.g., US Area Deprivation Index 2020) or (2) the final year in a period represented by the data (e.g., American Community Survey data aggregated over 2012 to 2016 is included in the 2016 dataset).

- **Variable availability:** Variable availability varies across data years. Following each data source description in this report, you will find a table that outlines the availability of each variable in the annual datasets. When a variable is not available, we indicate it with 'NA' (not available) or simply '-'.

- **Variable naming:** With the exception of geographic identifier variables — typically named 'fips' — all other variable names use descriptive titles, which are generally retained from the original source column names.

- **Missing values:** In the datasets, we use a blank to denote missing values, with one exception being the provider ratio variables from the County Health Rankings (CHR) data. These have negative values for counties where the number of providers is zero, a detail further explained in the CHR data description.

For comprehensive information about each data source, please refer to the subsequent sections of this report.

3.3 DATASET VERSIONING

In terms of dataset versioning, we utilize the Microsoft SQL Server database system to provide these datasets to be consistent with the VA's CDW work environment. Each dataset is stored in a dedicated table within a schema in the database. The quarterly releases are organized under distinct schema names within the database, such as OMHSP_FY22Q1, OMHSP_FY22Q2, OMHSP_FY22Q3, OMHSP_FY22Q4, OMHSP_FY23Q1, and so forth. These schema names facilitate distinguishing between releases when we deliver the same dataset, albeit updated, from one release to the next. Our approach to naming schemas with fiscal year and quarter identifiers (e.g., FY25Q1) was designed to ensure precise tracking of datasets over time. It provides an unambiguous record of when datasets were delivered, which is critical for project management and accountability.

3.4 METADATA

As stated above, the metadata tables in our project have been renamed starting in FY25Q2 to enhance clarity and consistency in our data curation process. The following two paragraphs outline previous changes made in prior fiscal years, providing context for the evolution of our metadata structure before detailing the current organization of metadata tables.

Starting from FY23Q1, the ORNL team provided an updated metadata table, the original name of this table was SEDH_meta_table, and it was located in the OMHSP schema.

Continuing our efforts to simplify and improve our data curation documentation, in the FY24Q3 delivery, the ORNL team performed a reorganization of the metadata table. The original SEDH_meta_table, located in the OMHSP schema, has been refactored into two tables: OMHSP.SEDH_table_metadata and OMHSP.SEDH_column_metadata. The division of the columns is as follows:

3.4.1 OMHSP.table_metadata Table

This table contains the following columns:

4. **table_id**: The ID of the table that this column belongs to, using the schema name and table name to facilitate identification of the source.

5. **schema_name**: Quarterly release schema names in the database (e.g., OMHSP_FY22Q4, OMHSP_FY23Q1, OMHSP_FY23Q2, OMHSP_FY23Q3, and so on).

6. **table_name**: The table name as it appears in the MS SQL Server database.

7. **table_name_description**: A description of the table name.

8. **availability_across_years**: The years for which data are available.

9. **data_source**: The name of the source organization that provided the raw data (starting in FY23Q4).

10. **data_source_description**: Description of the source organization (starting in FY23Q4).

11. **data_source_url**: URL of the source organization (starting in FY23Q4).

12. **spatial_resolution**: Spatial resolution or geography (e.g., state, county, block group, census tract, and zip code) (starting in FY23Q4).

13. **determinant**: Using the ontology from: Dang, Yifang, et al. “Systematic Design and Evaluation of Social Determinants of Health Ontology (SDoHO).” arXiv preprint arXiv:2212.01941 (2022). Current options used are: health care, neighborhood, social and community context, economic stability, food, and education.

14. **source_attribute**: Two option values: derivative (datasets produced from other datasets by applying a model and creating an index value) and authoritative (datasets that have not been modified other than ensuring the inclusion of required geographic administrative boundary identifiers such as FIPS codes) (started in FY24Q2).

15. **dimension**: Two option values: social and environmental (started in FY24Q2).

16. **osti_id**: All our reports are publicly available at the U.S. Department of Energy Office of Scientific and Technical Information (osti) at osti.gov. The osti_id is the unique identifier assigned to each report (started in FY24Q2).

17. **ornl_res_pub_id**: All our reports are available at ORNL in the Resolution Publication System; this column provides this unique identifier (started in FY24Q2).

18. **edh_project_exclusive**: This flag indicates whether the dataset can be shared outside the VA OMHSP EDH project. Datasets (i.e., tables and their columns) with this flag set to 'Y' should not be shared outside the OMHSP EDH project (Column added in FY24Q3).

3.4.2 OMHSP.column_metadata Table

This table contains the following columns:

19. **column_id**: A sequential number.

20. **table_id**: The ID of the table that this column belongs to, using the schema name and table name to facilitate identification of the source.

21. **column_name**: Column names within each dataset as they appear in the MS SQL Server table.

22. **column_name_description**: Descriptions of each column name.

23. **column_type**: The column type in the MS SQL Server table.

24. **column_length**: The column length in the MS SQL Server table.

25. **total_rows**: The number of rows in each column in each dataset (starting in FY23Q2).

26. **null_rows**: The number of null rows for each column in each dataset (starting in FY23Q2).

With each new quarterly release, the metadata table will be updated with new information in the aforementioned columns for each delivered dataset.

3.4.3 OMHSP.reports Table

Starting from FY24Q2, the ORNL team provides an updated OMHSP.SEDH_reports table which includes not only the metadata related to sponsor reports but also the PDF content of the sponsor reports. This table contains the following columns:

- **schema**: Quarterly release schema names in the database (e.g., OMHSP_FY22Q4, OMHSP_FY23Q1, and so on).

- **osti_id**: All our reports are publicly available at the U.S. Department of Energy Office of Scientific and Technical Information at osti.gov. The osti_id is the unique identifier assigned to each report.

- **ornl_res_pub_id**: All our reports are available at ORNL at the Resolution Publication System; this column provides this unique identifier.

- **reference_report**: This column contains the reference of the report in APA format.

- **report_url**: This column provides the osti.gov URL link.

- **pdf_file_name**: The PDF format file name follows this naming convention: OMHSP_[database schema used for versioning, which is also the quarterly delivery]_[osti_id].

- **pdf_content**: The report content in blob format.

Starting from FY25Q2, to enhance the accessibility and usability of our reports, the ORNL team has updated the reports table to include dedicated sections that align with the structure of our documentation. These additions ensure that each key component of the report is clearly defined and readily available for reference. The following columns have been introduced to systematically capture different sections of the report:

- **introduction**: A summary of the report's purpose, scope, and key objectives.
- **document_overview**: A high-level description of the report's contents, including its structure and previous releases.
- **contents_and_structure**: A detailed outline of the report's sections, conventions, versioning, metadata tables, an introduction to FIPS codes, guidance on how to join cohorts to our datasets, and error-checking procedures.
- **appendix**: Additional supporting error checking materials, such as tables, figures, or references, that complement the main report.
- **change_log**: A record of modifications made to the report in the current release, including updates, corrections, and revisions.

By implementing these updates, we are aligning our documentation structure with our VA sponsor's requests, ensuring clarity, consistency, and ease of access to critical report sections. These enhancements reflect our ongoing commitment to improving data organization and delivering comprehensive, well-structured documentation.

Please note that the report_url column will be updated in the VA's CDW transmit database as soon as it becomes available on the Office of Scientific and Technical Information website (osti.gov) of the US Department of Energy, typically four weeks after each quarterly release.

3.5 FIPS AS GEOGRAPHIC IDENTIFIERS AND PRIMARY KEYS

At ORNL, we utilize the Federal Information Processing Standards (FIPS) as geographic identifiers and primary keys in each dataset or table for this project. FIPS codes are publicly recognized standards developed by the National Institute of Standards and Technology (NIST) for computer systems and non-military applications, particularly for standardizing codes of geographical areas. FIPS specifications encompass various geographical areas:

- FIPS 10-4 for country and region codes
- FIPS 5-2 for state codes
- FIPS 6-4 for county codes

These codes are unique within their respective geographic entities. For example, FIPS state codes are unique within a country, and FIPS county codes are unique within a state. Since counties nest within states, a complete county FIPS code combines the state and county identifiers. For instance, if multiple counties end with "001," the state FIPS code is added to make each county FIPS code distinct (e.g.,

01001, 02001, 04001), where the first two digits indicate the state, and the last three digits represent the county.

Although NIST initiated the replacement of FIPS with the Geographical Name Information System (GNIS) Feature ID in 2002, many federal organizations in the United States, including the US Census Bureau, continued to use FIPS due to its broader coverage and precision in identifying geographic entities, especially smaller areas with uncertain natural boundaries. The US Census Bureau maintains a comprehensive hierarchy of census geographic entities for reference.

As the primary key in all datasets for this project, we consistently use the column "FIPS" to ensure unique data identification, regardless of the source FIPS granularity. We specify the FIPS granularity, such as region, state, county, census division, tracks, group blocks, etc., in the metadata table and reports' descriptions. Users are presumed to be familiar with joining datasets using FIPS columns at different geographic levels.

It's worth noting that only a few datasets since the inception of this project do not include a FIPS column. These exceptions are the following:

1. The National Mental Health Services Survey (table: national_mental_health_services_survey), delivered in FY22Q4.

2. The Veterans Service Organizations (VSO) 2010-2022, by zip code, delivered in FY24Q1.

3. HUD USPS Zip Code Crosswalk Files, ZIP-to-tract for 2023, delivered in FY24Q2.

These datasets were provided upon special request from the sponsor.

3.6 MAPPING ZIP CODES TO FIPS CODES FOR COUNTIES: OUR METHODS

When realigning spatial data to different boundaries that do not perfectly match or nest within the original spatial units, some data loss is inevitable. This occurs because the spatial distribution of data at higher resolutions than the native unit is often unknown. For example, certain zip code boundaries overlap with multiple county boundaries. When attempting to map zip code-level data to counties, there are situations where data must be reassigned to two or more counties with limited knowledge of how to allocate it accurately. Various methods exist to mitigate the degree of data loss, each with its strengths and weaknesses based on the data's nature. For social data, one effective approach is to allocate data based on population distribution or addresses within those boundaries to reduce misallocation.

Visual examples are provided below to illustrate this challenge:

Example: - *Left*: Zip Code 27826 - *Right*: County FIPS 37177

Figure 1. Example: - Left: Zip Code 27826 - Right: County FIPS 37177

3.7 HOW TO LINK COHORTS TO OUR DATASETS

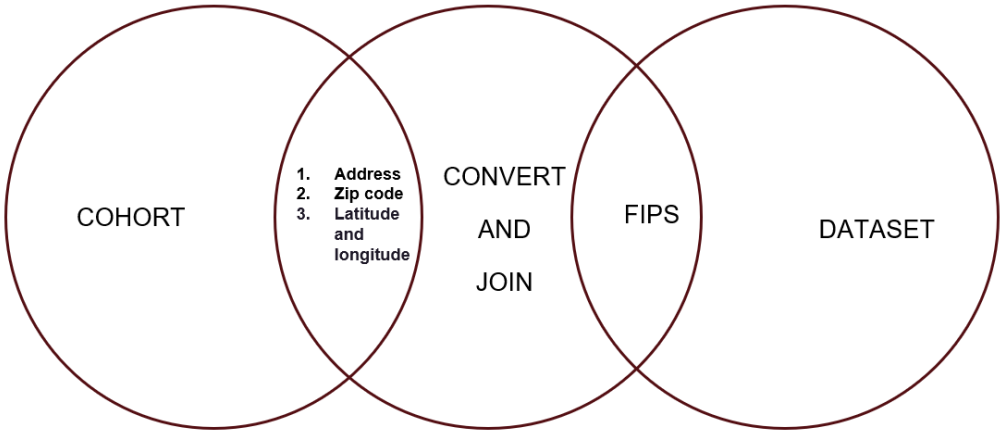


Figure 2. Joining geoids to our datasets

The above image focuses on the practical aspect of utilizing our datasets, specifically highlighting the process of joining them using FIPS codes as the primary key. On the left side of the slide, we have a visual representation of the cohort data, which may include various location identifiers such as addresses, zip codes, or latitude and longitude coordinates. These location identifiers serve as the basis for mapping, essentially translating them into FIPS codes, which are standardized geographic identifiers used in our datasets.

The circle on the left symbolizes this mapping process, where the location information from the cohort is transformed into FIPS codes for compatibility with our datasets. It's important to note that there are several methods to perform this conversion and join process. Different tools and techniques may be employed based on the specific requirements and characteristics of the datasets and cohort.

By effectively joining the cohort data with our datasets using FIPS codes, we can integrate and analyze information from various sources, enhancing the depth and breadth of our insights. This process of intersection and integration facilitates comprehensive analysis and decision-making, enabling us to leverage the full potential of our datasets in addressing research questions and informing strategic initiatives. As we proceed, it's crucial to prioritize data integrity and accuracy throughout the conversion and join process, ensuring reliable and meaningful outcomes from our analyses.

One method for converting and joining a cohort to the datasets. This approach involves using addresses and/or zip codes from the cohort data and mapping them to FIPS codes, which serve as the common identifier in our datasets. To facilitate this mapping process, we rely on crosswalk tables provided by the US Housing and Urban Development's Office of Policy Development and Research.

These crosswalk tables offer a reference point for associating zip codes with corresponding FIPS codes, enabling integration with our datasets. However, it's important to acknowledge that this process is not without its limitations. In approximately 45% of cases, zip codes cannot be perfectly mapped to FIPS codes at the county level. This imperfection underscores the challenges inherent in geographic data integration and highlights the need for careful consideration and validation when performing these conversions.

Despite its imperfections, leveraging crosswalk tables remains a valuable approach for linking cohort data to our datasets, providing a foundational step in the analysis and interpretation of determinants of health. As we navigate through this process, it's essential to remain mindful of these limitations and explore alternative methods for data integration where necessary, ensuring the accuracy and reliability of our analyses.

The second method for utilizing our datasets, which involves leveraging latitude and longitude coordinates for conversion and joining purposes. SQL Server offers support for two spatial data types: Geometry and Geography. These data types enable a more precise conversion of latitude and longitude coordinates to FIPS codes. Unlike the first method which relies on crosswalk tables, this approach provides a higher level of accuracy in mapping locations to FIPS codes.

However, it's important to note that implementing this method requires a higher level of expertise and experience in working with Geometry and Geography files within SQL Server. Users must possess a deeper understanding of spatial data manipulation techniques and SQL Server functionalities to effectively execute this conversion process. Despite the complexity involved, leveraging latitude and longitude coordinates through SQL Server's spatial data types offers the advantage of increased precision and accuracy in data integration.

Organizations with skilled personnel and advanced technical capabilities may opt for this method to ensure the highest level of spatial data accuracy in their analyses. As with any advanced technique, thorough testing and validation are essential to verify the integrity of the converted data and ensure its suitability for analysis and decision-making purposes.

3.8 ERROR CHECKING

Beginning with the FY23Q1 release, the ORNL team will additionally give succinct information regarding error checking activities in order to provide formal evidence that the datasets supplied have been thoroughly error checked. Our data profiling process is described in our project's overview manuscript [12]:

“Following standard data and software development methodologies, data profiling is performed in four different work environments: 1) a team-shared work environment for selection, extraction, and refinement of raw data (development); 2) an ORNL intranet work environment focused on quality assurance testing (QA-Intra); 3) an ORNL Knowledge Discovery Infrastructure (KDI) secure work environment that stores highly sensitive data and ensures its security (QA-KDI). And finally, 4) a production environment housed within the KDI environment and accessible to our VA sponsors, (Production). We carried out test iterations in each of the four work environments as the datasets moved through them to confirm data integrity and system compatibility.

All datasets were error-checked using a data profiling strategy that includes at least two reviewers and the following test groups:

1. evaluating missingness: i.e. determining the amount of missing data by randomly checking for them;

2. compiling descriptive statistics, such as the number of rows, columns, and types of variable data;

3. appending checksums to a subset of the columns on both the source and destination copies to ensure consistency;

4. consistently representing the social and physical environment using FIPS codes as geographic administrative boundaries and confirming that the FIPS codes correspond to the geographic administrative boundaries of the original data;

5. manually comparing the first, last, and five additional randomly selected rows for consistency between the source and target datasets.

When datasets are developed at ORNL, which we call ‘derivative’, ORNL will provide extra error-checking utilizing a combination of statistical methodologies based on each dataset’s properties, in addition to the data profiling methodology described above.” [12]

The error-checking results for FY25Q3 follows:

			Development		QA-Intra		QA-KDI (VIEWS)		Production (Transmit)		
Dataset Name	Rows	Columns	Passes	Fails	Passes2	Fails3	Passes4	Fails5	Passes6	Fails7	Error ratio
[OMHSP].[table_metadata]	76	15	4	1	5	0	5	0	4	1	0.11
[OMHSP].[column_metadata]	1009	8	4	1	5	0	5	0	5	0	0.05
[OMHSP].[reports]	15	12	4	1	5	0	5	0	4	1	0.11
[OMHSP_FY25Q3].[atf_federal_firearms_listings_county_2025]	9628	3	5	0	5	0	5	0	5	0	0
[OMHSP_FY25Q3].[atm_press_metrics_county_2017_2020]	4537680	6	5	0	5	0	5	0	5	0	0
[OMHSP_FY25Q3].[housinginsecurity_county_2024]	3144	7	4	1	5	0	5	0	5	0	0.05
[OMHSP_FY25Q3].[SEDoH_county_2018_2019]	6466	13	5	0	5	0	5	0	5	0	0
[OMHSP_FY25Q3].[VADriveTimeData_County_June2025]	13129801	11	5	0	5	0	5	0	5	0	0

Appendix A presents descriptive statistics of error-checking results.

4. DRIVE-TIME

4.1 DATA SOURCE

VA and ORNL

4.2 DESCRIPTION

The Drive-Time dataset provides calculated driving distances and travel times from origin addresses to the nearest U.S. Department of Veterans Affairs (VA) facilities, including healthcare locations.

The FY25Q3 data incorporates three updated inputs: patient tables, facility tables, and the OpenStreetMap (OSM) road network. The team successfully mapped a greater number of patients to FIPS codes by using an updated geocoding method. Rather than relying on text-based address information, the revised approach uses the geographic coordinates of the origin location and performs a spatial join with county boundary data. This method improves mapping accuracy and replaces the previous address-based technique.

Since the FY25Q1 update, county-level Federal Information Processing Standard (FIPS) codes have been included to support geographic referencing.

The SourceTable column defined below has been added since FY25Q3.

Notes on Naming Conventions:
Variables joined from external sources retain their original naming (e.g., camelCase) in alignment with the VA's Corporate Data Warehouse (CDW). Newly created variables follow internal conventions using lowercase letters with underscores (e.g., driving_time_minutes).

Method for Identifying the Nearest Facility:
The process begins by identifying the three nearest VA facilities using straight-line (Euclidean) distance between geographic coordinates. Next, routing algorithms are applied using the most recent version of the OpenStreetMap (OSM) road network—updated by the ORNL team with each dataset release. These algorithms estimate driving times to each facility, and the one with the shortest estimated time is selected. The other two are excluded from the final dataset.

This dataset is restricted to use within the VA's Environmental Determinants of Health (EDH) initiative. For access or additional details, please contact your program manager.

4.3 INCLUSION

- **Year**: 2025
- **Geographical unit:** FIPS at county level of the origin address. Continental USA.
- **Data included**: Selected FIPS at county level of the origin addresses.

4.4 RESOURCES

Link to VA Locations

Link to OpenStreetMap

4.5 UPDATE FREQUENCY

The dataset will be reviewed and updated quarterly.

Table 1. Drive-Time (DRIVETIME)

variable name	variable label	total_rows	null_rows
fips	County fips code of the origin address.	13129801	0
State	State of the origin address.	13129801	0
County	County name of the origin address.	13129801	741239
StreetAddress1	Origin address.	13129801	0
City	City of the origin address.	13129801	0
Zip	Zip code of the origin address.	13129801	1664
SourceTable	Indicates the originating electronic health record (EHR) system for the patient data—either VistA or Cerner.	13129801	0
drive_time(min)	Drive time in minutes.	13129801	0
drive_distance(km)	Drive distance in kilometers.	13129801	0
FacilityID	The VA's Facility identification number.	13129801	0

Table 2. Variable availability across years, (DRIVETIME)

variable name	2025
fips	X
State	X
County	X
StreetAddress1	X
City	X
Zip	X
SourceTable	X
drive_time(min)	X
drive_distance(km)	X
FacilityID	X

5. ATMOSPHERIC PRESSURE DATA

5.1 DATA SOURCE

National Oceanic and Atmospheric Administration (NOAA) ERA5

5.2 DESCRIPTION

ERA5 is the latest climate reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), delivered through the Copernicus Climate Change Service (C3S) and distributed via the U.S. National Oceanic and Atmospheric Administration (NOAA). It provides comprehensive climate variables covering the atmosphere, land, and ocean, and has replaced the earlier ERA-Interim dataset.

This dataset includes ERA5 Table 2: Surface and Single Level Parameter — specifically, surface pressure values provided as daily minimum, maximum, and mean. All atmospheric pressure values are reported in Pascals.

To generate county-level estimates of atmospheric pressure, the ORNL team processed the data using Python as follows:

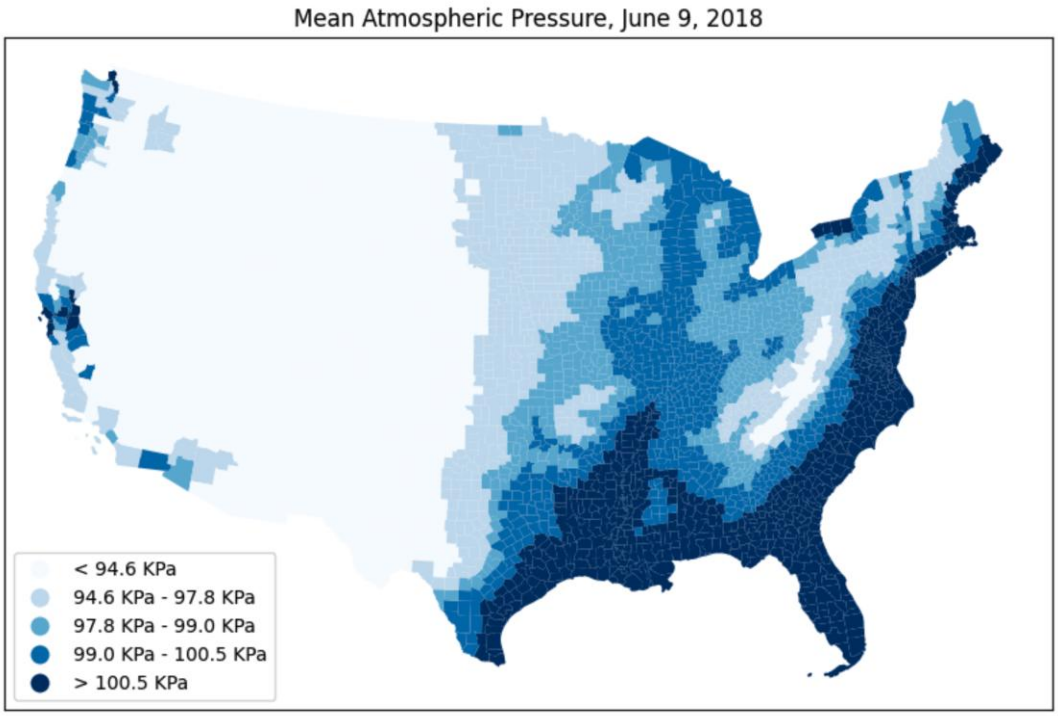
6. For each year, the dataset was loaded using the 'rioxarray' Python library. Pixel geometries were extracted to represent the area covered by each raster cell, first in pixel space and then reprojected to EPSG:4326.

7. Using 'numpy' Python library, daily minimum, mean, and maximum statistics were computed for each pixel.

8. The 2018 U.S. county boundary shapefiles, published by the U.S. Census Bureau, were loaded and used in combination with the 'tobler' module from the PySAL library to perform areal interpolation. This process yielded interpolated atmospheric pressure estimates for each county on each day.

These steps ensure the dataset offers spatially consistent, county-level estimates that are suitable for integration into VA predictive modeling efforts such as ReachVet 2.0 Phase II and STORM 2.0.

Data version: General availability release as of August 6, 2024.



Mean Atmospheric Pressure on June 9, 2018 in Kilo Pascals (ERA5 Data)

5.3 INCLUSION

Year: 2017 - 2020

Geographical unit: county fips continental US

5.4 RESOURCES

For more information on the ERA5 Atmospheric Pressure Data, see:

- NOAA ERA5
- ERA5 Data Documentation
- Cartographic Boundary Files - Shapefile
- Download Instructions

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 . Atmospheric Pressure Data (ERA5)

variable name	variable label	total_rows	null_rows
fips	Federal Information Processing Standards (FIPS), county-level fips codes.	4537680	0
year	The year corresponding to the data.	4537680	0
day	The day of the month corresponding to the atmospheric pressure measurements. Format is one or two digits, such as D (e.g., 5) or DD (e.g., 05).	4537680	0
pres_mean	The daily mean surface atmospheric pressure in Pascals (Pa), averaged over all hourly values within a given day for a specific geographic unit (e.g., county).	4537680	0
pres_min	The daily minimum surface atmospheric pressure in Pascals (Pa), representing the lowest pressure value recorded during a 24-hour period for a given geographic unit (e.g., county).	4537680	0
pres_max	The daily maximum surface atmospheric pressure in Pascals (Pa), representing the highest pressure value recorded during a 24-hour period for a given geographic unit (e.g., county).	4537680	0

Table 4 . Variable availability across years, (ERA5)

variable name	2017	2018	2019	2020
fips	X	X	X	X
year	X	X	X	X
day	X	X	X	X
pres_mean	X	X	X	X
pres_min	X	X	X	X
pres_max	X	X	X	X

6. FEDERAL FIREARMS LISTINGS

6.1 DATA SOURCE

ATF - Bureau of Alcohol, Tobacco, Firearms and Explosives

6.2 DESCRIPTION

This dataset provides a count of active Federal Firearms Licensees (FFLs) within each U.S. county for the year 2025. Each entry includes the county's Federal Information Processing Standards (FIPS) code, the type of license (e.g., dealer, manufacturer, importer), and a count number of the give license type for the given FIPS code for 2025.

The data is sourced from the Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) and reflects the distribution of federally licensed firearms businesses across the country. Such information is valuable for analyzing regional patterns in firearms commerce, assessing regulatory compliance, and informing policy decisions related to public safety and firearms regulation.

Federal Firearms Listings refer to databases or records maintained by the Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) that catalog information about firearms, manufacturers, dealers, and classifications under U.S. federal law. The term can apply to several specific resources:

ATF Federal Firearms Licensee (FFL) Listings are public records of individuals and businesses that have a Federal Firearms License to legally engage in the manufacture, import, or sale of firearms. The listings typically include: - Licensee name - Business address - License type (e.g., dealer, manufacturer) - Expiration date

The types of licenses follow:

- code: 01

  license_type: Dealer in Firearms Other Than Destructive Devices

  description: Typical firearms dealer (handguns, rifles, shotguns).

- code: 02

  license_type: Pawnbroker in Firearms Other Than Destructive Devices

  description: Allows pawnbroking of standard firearms.

- code: 03

  license_type: Collector of Curios and Relics (C&R)

  description: Non-commercial license for collecting historic firearms.

- code: 06

  license_type: Manufacturer of Ammunition for Firearms

description: Allows manufacturing of ammunition (not including destructive devices).

- code: 07

  license_type: Manufacturer of Firearms Other Than Destructive Devices

  description: Can manufacture and deal in firearms; often used by gunsmiths or small makers.

- code: 08

  license_type: Importer of Firearms Other Than Destructive Devices

  description: Allows importation of standard firearms.

- code: 09

  license_type: Dealer in Destructive Devices

  description: Can deal in weapons such as grenades, missiles, or explosive weapons.

- code: 10

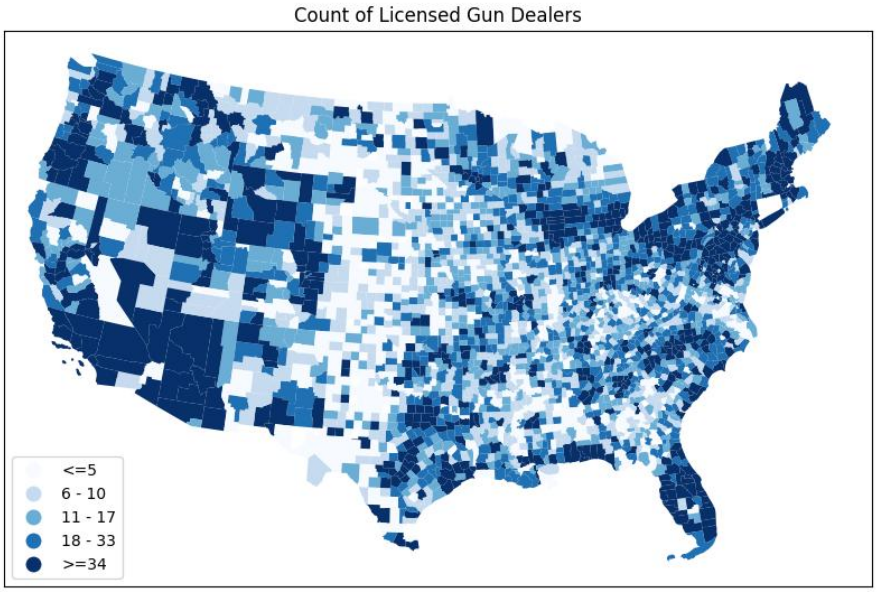
  license_type: Manufacturer of Destructive Devices

  description: Can manufacture explosive or high-power military-grade weapons.

- code: 11

  license_type: Importer of Destructive Devices

  description: Can import high-power military-grade or explosive weapons.



U.S. Count of Licensed Gun Dealers, 2025

6.3 INCLUSION

Year: 2025 data
Geographical unit: FIPS County Level, Continental US.

6.4 RESOURCES

Federal Firearms Listings:

- Link to ATF Website
- Link to ATF Data

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 . Federal Firearms Listings (FIREARMS)

variable name	description
fips	Federal Information Processing Standards (FIPS) - at county level.
lic_type	The license type issued by the Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF), indicating the classification of the Federal Firearms License (FFL), such as dealer, manufacturer, or importer.
count	The count number of the give license type for the given FIPS.

Table 6 . Variable availability across years, (FIREARMS)

variable name	2025	total rows	null rows
fips	X	9628	0
lic_type	X	9628	0
count	X	9628	0

7. HOUSING INSECURITY

7.1 DATA SOURCE

CDC PLACES

7.2 DESCRIPTION

PLACES is a CDC-led initiative launched in partnership with the Robert Wood Johnson Foundation and CDC Foundation to provide small-area health data across the U.S. Using model-based estimates, it offers detailed local data on chronic diseases, health behaviors, and social determinants of health at the county, place, census tract, and ZIP Code Tabulation Area (ZCTA) levels.

Key Features:

- Free, interactive web tool with data for areas with at least 50 adult residents.
- Covers 40+ health and social measures, including 12 health outcomes, 9 social determinants, and 4 health risk behaviors.
- Data sourced from BRFSS, U.S. Census, and the American Community Survey.
- Supports health officials, policymakers, and community planners in identifying local health disparities, planning targeted interventions, and improving health equity.

In 2024, PLACES adopted Census 2020 geographies.

History: Originally launched in 2015 as the 500 Cities Project, it was expanded in 2020 to become PLACES, now offering nationwide small-area estimates to support local health decision-making across both urban and rural communities.

In this dataset, we provide housing insecurity dataset. This measure estimates the prevalence of adults unable to pay mortgage, rent, or utility bills in the past year using model-based data derived from the Behavioral Risk Factor Surveillance System (BRFSS) and American Community Survey (ACS) through a multi-level regression and post-stratification method. It reports both crude and age-adjusted prevalence, with a 95% confidence interval based on Monte Carlo simulations.

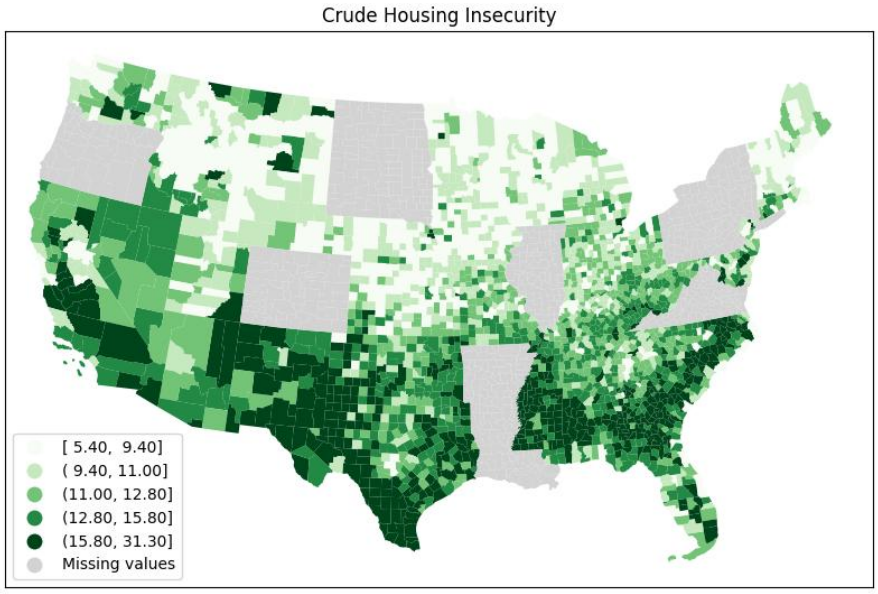
Key Insights:

- Housing insecurity is linked to poor health outcomes and reduced healthcare access.
- In 2022, over 42 million U.S. households were housing cost-burdened (spending >30% of income on housing).
- Renters, urban residents, and some minority groups face greater risk.
- Unaffordable housing correlates with hypertension, cardiovascular disease, and mental health issues.
- Supportive housing programs and rent subsidies can reduce homelessness and improve health outcomes.

Limitations:

- This indicator does not cover all aspects of housing instability (e.g., quality, crowding, frequent moves).
- Data only available for states using the Social Determinants and Health Equity BRFSS module.

Policy Link: Supports Healthy People 2030 objective SDOH-04 — reducing families spending over 30% of income on housing.



U.S. Unadjusted Housing Insecurity Rate, 2024

7.3 INCLUSION

Year: 2024

Geographical unit: county level fips

7.4 RESOURCES

For more information on the CDC Places, see:

- Link to CDC PLACES
- Link to Housing insecurity dataset

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. Housing Insecurity (HOUSING)

variable name	variable label	total_rows	null_rows
fips	Federal Information Processing Standards (FIPS), county-level fips codes.	3144	0
CrudePrev	Crude Prevalence. Crude refers to that the estimate is unadjusted for factors like age, sex, or other demographics.	3144	0
Crude95CI_low	This value represents the lower bound of the 95% confidence interval for the crude prevalence estimate.	3144	0
Crude95CI_high	Upper limit of the 95% confidence interval for the crude prevalence.	3144	0
AdjPrev	Age-adjusted prevalence estimate of the outcome being measured.	3144	0
Adj95CI_low	Lower limit of the 95% confidence interval for the age-adjusted prevalence estimate.	3144	0
Adj95CI_high	Upper limit of the 95% confidence interval for the age-adjusted prevalence estimate.	3144	0

Table 8. Variable availability across years, (HOUSING)

variable name	2024
fips	X
CrudePrev	X
Crude95CI_low	X
Crude95CI_high	X
AdjPrev	X
Adj95CI_low	X
Adj95CI_high	X

8. SOCIAL AND ENVIRONMENTAL COUNTY-LEVEL INDICATORS FOR INCLUSION IN SUICIDE AND OVERDOSE NEXT-GENERATION PREDICTIVE MODELING (REACH VET AND STORM)

8.1 DATA SOURCE

Several

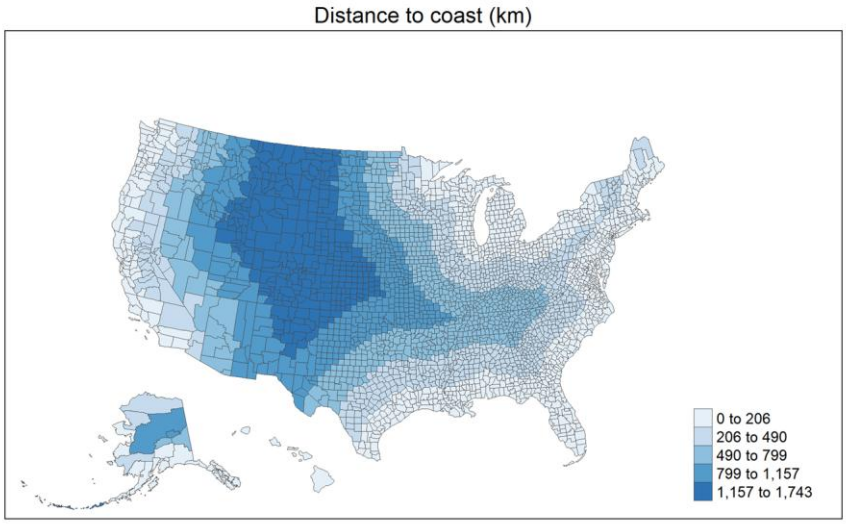
8.2 DESCRIPTION

The Social and Environmental county-level indicators for inclusion in Suicide and Overdose Next-Generation Predictive Modeling (REACH VET and STORM) follow.

8.2.1 Distance to Coast (km)

Source: https://catalog.data.gov/dataset/tiger-line-shapefile-2019-nation-u-s-coastline-national-shapefile
Notes: Final values calculated using distance from the centroid of the region to the coast according to census data.
Steps:

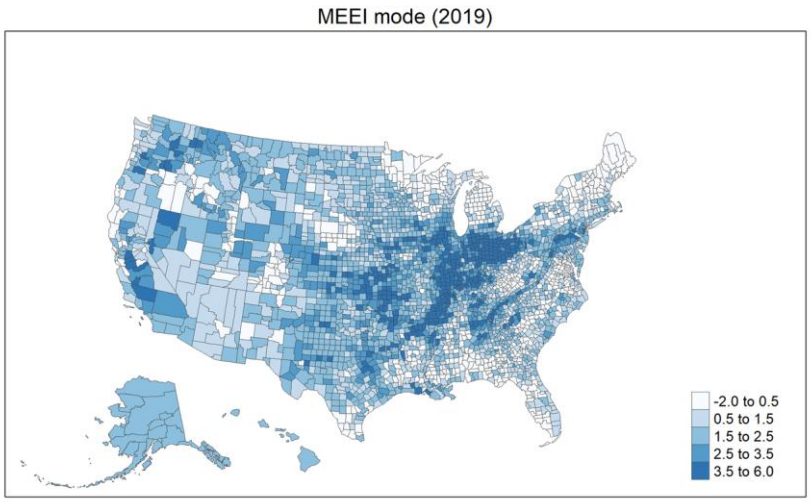
1. Downloaded county shapefile from Census
2. Calculated centroid for each county
3. Downloaded Coastline National Shapefile for the US from Census
4. Calculated minimum distance for each centroid to the coast



Distance to Coast (km)

8.2.2 Multi-Environmental Exposures Index (MEEI)

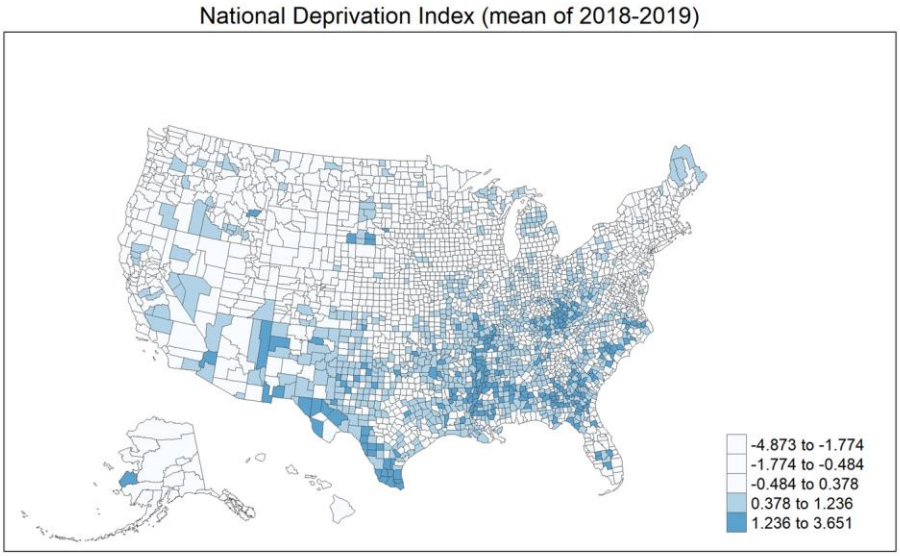
Source: Peluso, A., Rastogi, D., Klasky, H. B., Logan, J., Maguire, D., Grant, J., … & Hanson, H. A. (2024). Environmental determinants of health: Measuring multiple physical environmental exposures at the united states census tract level. Health & Place, 89, 103303. https://doi.org/10.1016/j.healthplace.2024.103303
Data: https://github.com/AlinaPeluso/MEII/blob/main/MEII_data.csv
Notes: The MEEI captures multiple environmental exposures in the same area that can result in additive and synergistic effects on health outcomes.



Multi-Environmental Exposures Index (MEEI)

8.2.3 Neighborhood Deprivation Index

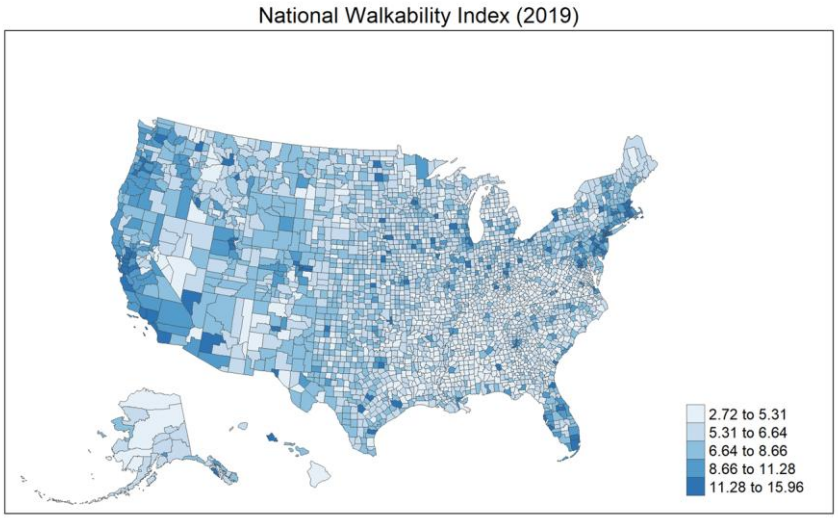
Source: https://cran.r-project.org/web/packages/ndi/vignettes/vignette.html
Notes: Neighborhood Deprivation Index (Powell-Wiley) raw score, with imputed values for missing data.



Neighborhood Deprivation Index

8.2.4 National Walkability Index

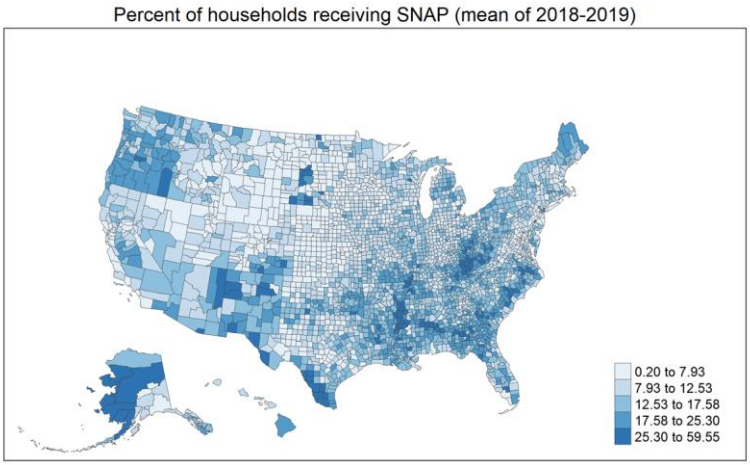
Source: https://epa.maps.arcgis.com/home/item.html?id=f16f5e2f84884b93b380cfd4be9f0bba
Notes: Calculated by the EPA at the block group level based on street intersection density, proximity to transit stops, and diversity of land uses; aggregated to county level.



National Walkability Index

8.2.5 Percent of Households Receiving SNAP

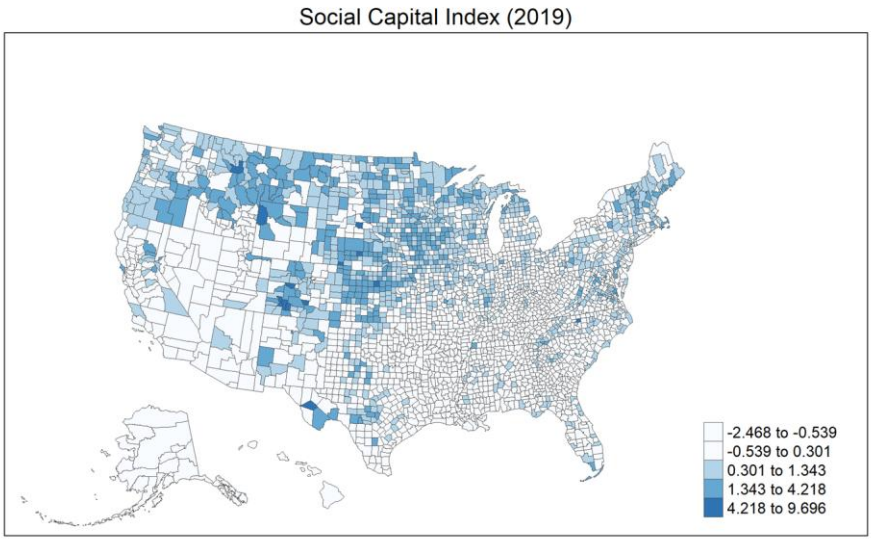
Source: Census data, ACS 5-year estimates, Table S2201
Notes: Percent of households receiving food stamps/SNAP in the past 12 months.



Percent of Households Receiving SNAP

8.2.6 Social Capital Index (2019)

Source: Peluso, A., Tuccillo, J., Sparks, K., Kapadia, A., & Hanson, H. A. (2024). Spatial analysis of social capital and community heterogeneity at the United States county level. Applied Geography, 162, 103168. https://doi.org/10.1016/j.apgeog.2023.103168
Data: https://github.com/AlinaPeluso/SCIndex
Notes: Social capital is defined as the resources or benefits received through one's social connections. It supports economic efficiency, community development, and public health.

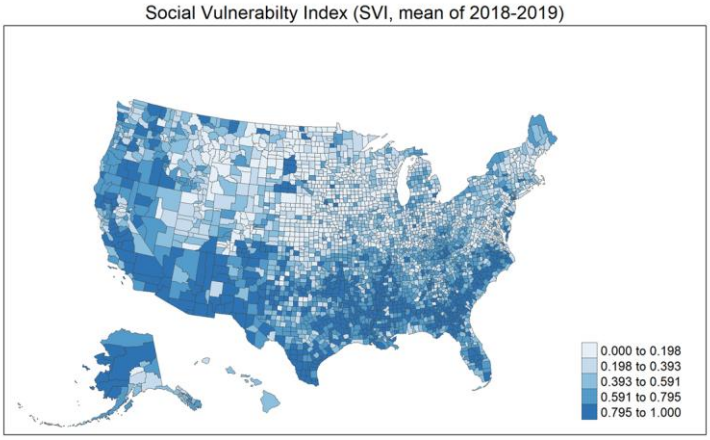


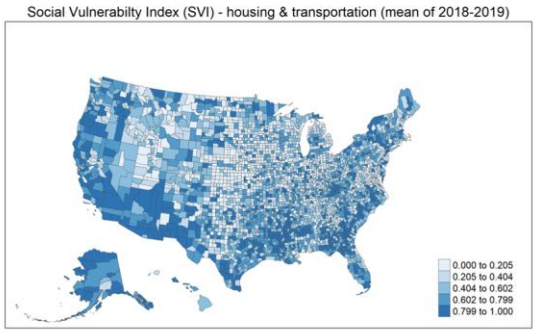
Social Capital Index (2019)

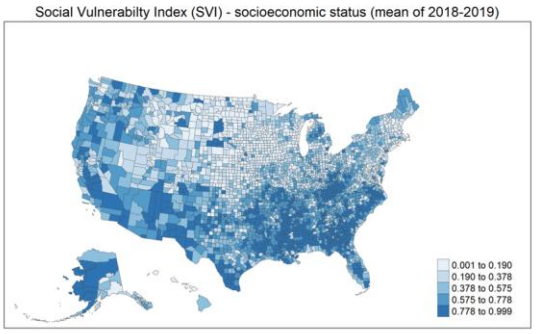
8.2.7 Social Vulnerability Index (SVI)

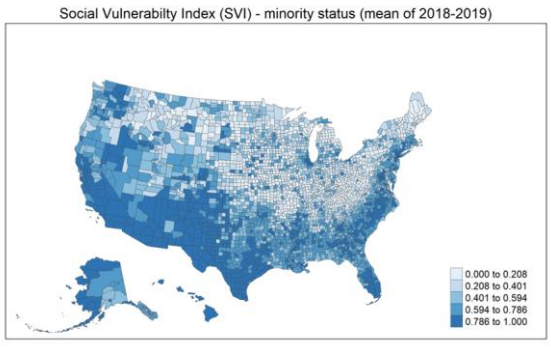
Source: The Centers for Disease Control and Prevention (CDC) and Agency for Toxic Substances and Disease Registry (ATSDR) https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html
Notes: The SVI is calculated using CDC/ATSDR's methodology (Flanagan et al., 2011), implemented via the fidnSVI R package. It includes 4 themes:

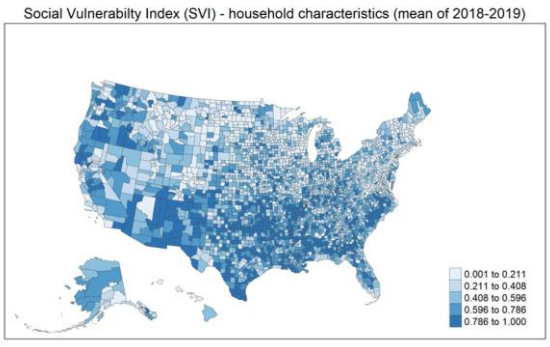
1. Socioeconomic Status
2. Household Characteristics
3. Racial & Ethnic Minority Status
4. Housing Type & Transportation
5. The final score is a composite average across all four themes.











8.3 INCLUSION

Year: 2018-2019

Geographical unit: county fips

8.4 RESOURCES

For more information on the ReachVet Program, see:

- Links to ReachVet Program:
  (https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2785078)
  (https://psychiatryonline.org/doi/full/10.1176/appi.ps.202100629)

For more information on the STORM Program, see:

- Link to STORM Program

8.5 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 9. Social and Environmental county-level indicators for inclusion in Suicide and Overdose Next-Generation Predictive Modeling (REACH VET and STORM) (SEDOH)

variable name	variable label	total_rows	null_rows
fips	Federal Information Processing Standards (FIPS), county-level fips codes.	6466	0
Year	The year corresponding to the data.	6466	0
NDI_pwi	Neighborhood deprivation Index.	6466	0
SVI_RPL_theme1	Social vulnerability index theme 1.	6466	0
SVI_RPL_theme2	Social vulnerability index theme 2.	6466	0
SVI_RPL_theme3	Social vulnerability index theme 3.	6466	0
SVI_RPL_theme4	Social vulnerability index theme 4.	6466	0
SVI_RPL_themes	The Social Vulnerability Index (SVI), summarized across all its component themes.	6466	0
ACS_PCT_HH_FD_STMP	Percentage of households on food stamps/SNAP, past 12 months.	6466	0
NatWalkInd	Walkability index.	6466	0
MEEI_mode	Multi-exposure environmental index.	6466	0
SCI	Social capital index.	6466	0
distance_to_coast_km	Distance to coast in km.	6466	0

Table 10. Variable availability across years, (SEDOH)

variable name	2018	2019
fips	X	X
year	X	X
NDI_pwi	X	X
SVI_RPL_theme1	X	X
SVI_RPL_theme2	X	X
SVI_RPL_theme3	X	X
SVI_RPL_theme4	X	X
SVI_RPL_themes	X	X
ACS_PCT_HH_FD_STMP	X	X
NatWalkInd	X	X
MEEI_mode	X	X
SCI	X	X
distance_to_coast_km	X	X

9. REFERENCES

[1] 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, Rev. 2, ORNL/SPR-2021/2366 - Pub ID 170648. 2021. https://www.osti.gov/biblio/1854468

[2] Christian, J.B., Klasky, H.B., Sparks, K., Peluso, A., Tuccillo, J., Devineni, P., and Watson, R. VA EDH Data Curation Documentation - FY22-Q1, Rev. 2, ORNL/SPR-2022/2316- Pub ID 172755. 2022. https://www.osti.gov/biblio/1854460

[3] 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, Rev. 2, ORNL/SPR-2022/2391 - Pub ID 174092. 2022. https://www.osti.gov/biblio/1862127

[4] Klasky, H.B., Sparks, K., Logan, J., Tuccillo, J., Whitehead, M., Hamaker, A., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY22-Q3, Rev. 2. ORNL/SPR-2022/2487 - Pub ID 178645. 2022. https://www.osti.gov/biblio/1876283

[5] Klasky, H.B., Sparks, K., Logan, J., Hamaker, A., Whitehead, M., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY22-Q4, ORNL/SPR-2022/2587, PUB ID 183700. 2022. https://www.osti.gov/biblio/1892396

[6] Klasky, H.B., Sparks, K., Logan, J., Hamaker, A., Whitehead, M., Peluso, A., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY23-Q1, ORNL/SPR-2022/2694, PUB ID 187842. 2022. https://www.osti.gov/biblio/1909101

[7] Klasky, H.B., Sparks, K., Peluso, A., Whitehead, M., K., Logan, J., Hamaker, A., McGee, M., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY23-Q2, ORNL/SPR-2023/2857, PUB ID 19179. 2023. https://www.osti.gov/biblio/1971721

[8] Klasky, H.B., Sparks, K., Peluso, A., K., Logan, J., Hamaker, A., McGee, M., VanDerslice, J., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY23-Q3, ORNL/SPR-2023/2930 PUB ID 195499, 2023. https://www.osti.gov/biblio/1992724

[9] Klasky, H.B., Sparks, K., Peluso, A., K., Myers, A., Hamaker, A., McGee, M., Zhang, J., Logan, J., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY23-Q4, ORNL/SPR-2023/3097 PUB ID 202517, 2023. https://www.osti.gov/biblio/2204567

[10] Klasky, H.B., Sparks, K., Peluso, A., K., Myers, A., Logan, J., McGee, M., Hamaker, A., Zhang, J., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY24-Q1, ORNL/SPR-2023/3207 PUB ID 205615, 2023. https://www.osti.gov/biblio/2229216

[11] Klasky, H., Sparks, K., Peluso, A., Logan, J., McGee, M., Callaway, K., Cook, C., Sacca, D., Reszczynski, P., Hanson, H., Watson, R., Martins, S., Trafton, J., and Kapadia, A. VA EDH Data Curation Documentation (FY24-Q2). 2024. ORNL/SPR-2024/3299 PUB ID 210685. https://www.osti.gov/biblio/2341397

[12] Klasky, H.B., Hanson, H., Sparks, K., Whitehead, M., Blair, C., and Kapadia, A., "Dataset Repository for Investigating Suicide Risk Using Social and Environmental Determinants of Health", ORNL/TM-2023/3027 Pub ID 183902. 2022. https://www.osti.gov/biblio/1997699

[13] Klasky, H.B., Sparks, K., Grant, J., Tuccillo, J., Peluso, A., Logan, J., McGee, M., Callaway, K., MacFarland, M., Cook, H., Hanson, H., Watson, R., Martins, S., Trafton, J., and Kapadia, A. VA EDH Data Curation Documentation (FY24-Q3). United States: N. p., 2024. ORNL/SPR-2024/3419 PUB ID 215289. https://www.osti.gov/biblio/2404618

[14] Klasky, H.B., Sparks, K., Grant, J., Tuccillo, J., Peluso, A., Logan, J., McGee, M., Callaway, K., MacFarland, M., Cook, H., Hanson, H., Martins, S., Trafton, J., and Kapadia, A. VA EDH Data Curation Documentation (FY24-Q4). United States: N. p., 2024. ORNL/SPR-2024/3505 PUB ID 220033. https://www.osti.gov/biblio/2472692

[15] Klasky, H.B., Sparks, K., Peluso, A., Grant, J., Tuccillo, J., Spannaus, A., Logan, J., McGee, M., Callaway, K., MacFarland, M., Cook, H., Hanson, H., Martins, S., Trafton, J., and Kapadia, A. VA EDH Data Curation Documentation (FY25-Q1). United States: N. p., 2024. ORNL/SPR-2024/3689 PUB ID 226014. https://www.osti.gov/biblio/2502173

[16] Klasky, H.B., Sparks, K., Peluso, A., Grant, J., Tuccillo, J., Spannaus, A., Logan, J., McGee, M., Callaway, K., MacFarland, M., Cook, H., Hanson, H., Martins, S., Trafton, J., and Kapadia, A. VA EDH Data Curation Documentation (FY25-Q2). United States: N. p., 2024. ORNL/SPR-2025/3765 RES PUB ID 230283. https://www.osti.gov/biblio/xxx

APPENDIX A. ERROR CHECKING

APPENDIX A. ERROR CHECKING

The following descriptive statistics were obtained using the R summary function. The definitions of the main statistics included are as follows:

In R, the summary()[1] function provides a quick statistical summary of an object, typically a **data frame**, **vector**, or **factor**. The output varies depending on the data type but usually includes the following fields:

For Numeric Data (e.g., numeric vectors, columns in a data frame)

When applied to a numeric variable, summary() returns six key statistics:

1. **Min** – The minimum value in the dataset.
2. **1st Qu. (First Quartile, Q1)** – The 25th percentile, where 25% of the data is below this value.
3. **Median (Q2)** – The 50th percentile (middle value).
4. **Mean** – The arithmetic average of the values.
5. **3rd Qu. (Third Quartile, Q3)** – The 75th percentile, where 75% of the data is below this value.
6. **Max** – The maximum value in the dataset.

Summary statistics of selected tables follow.

[OMHSP_FY25Q3].[atm_press_metrics_county_2017_2020]

	pres_mean	pres_min	pres_max	day	year
count	4537680	4537680	4537680	4537680	4537680
mean	96666.24139982635	96332.25516	96925.80739804503	183	2018.5
std	5509.5319958691325	5501.100032549731	5522.430794069153	105.3660402910978	1.1180341119443657
min	65777.98438	65258.45313	65999.20313	1	2017
25%	95853.85156	95482.46094	96141.21875	92	2017.75
50%	98342.95313	97997.28906	98611.0625	183	2018.5
75%	100018.7031	99689.65625	100276.3438	274	2019.25
max	104064.6016	103802.0859	104358.5781	365	2020

[OMHSP_FY25Q3].[housinginsecurity_county_2024]

statistic	fips	CrudePrev	AdjPrev	Crude95CI_low	Crude95CI_high	Adj95CI_low	Adj95CI_high
count	3144	2417	2417	2417	2417	2417	2417
null_count	0	727	727	727	727	727	727
mean		12.762639645066299	14.222424488399065	11.307281754240794	14.316342573438146	12.618204385601985	15.931650806785273
std		4.103091856863093	4.2394602788332465	3.757143959038917	4.454499755825319	3.8983917451014927	4.586809330205042
min	01001	5.400000095367432	6.099999904632568	4.7	6.2	5.3	6.9
25%		9.800000190734863	11.199999809265137	8.6	11.1	9.9	12.7

[1] https://www.statswithr.com/r-functions/the-summary-function-in-r

50%		11.800000190734863	13.199999809265137	10.4	13.4	11.7	14.9
75%		14.800000190734863	16.399999618530273	13.1	16.6	14.6	18.3
max	56045	31.299999237060547	33.5	28.4	34.3	30.7	36.4