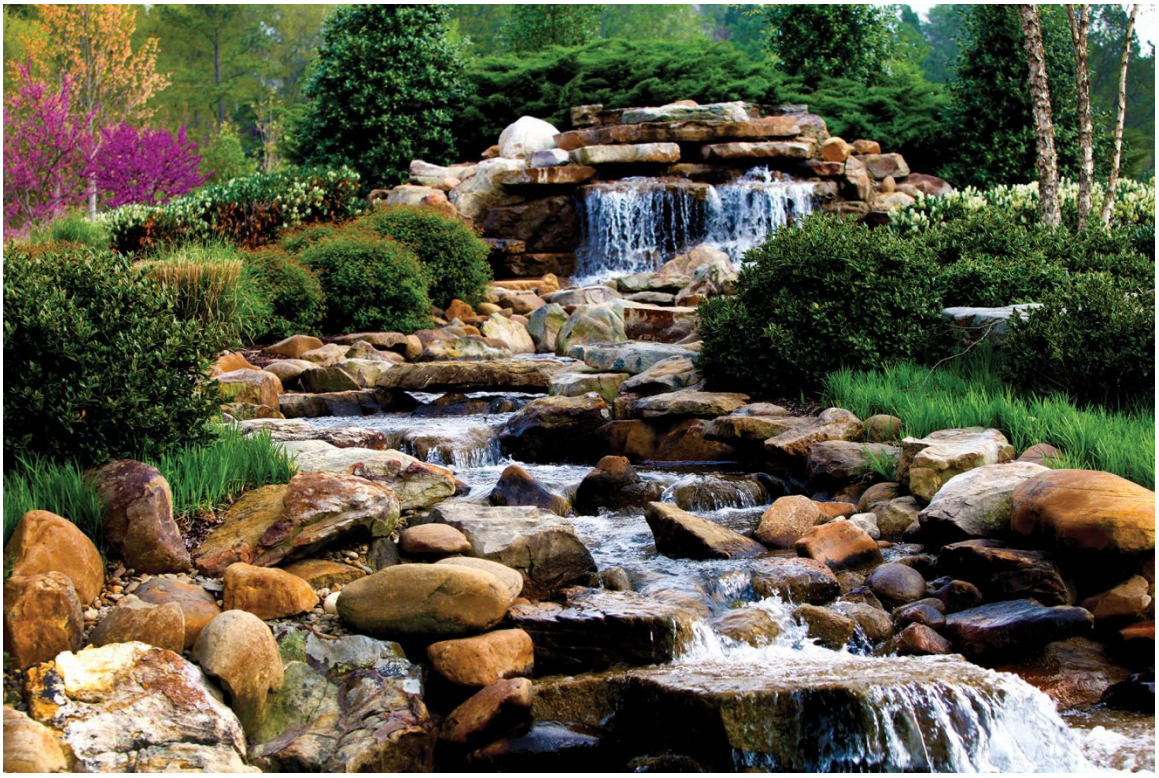
ORNL/SPR-2024/3689
RES PUB ID 226014

VA EDH Data Curation Documentation FY25-Q1



Hilda Klasky
Kevin Sparks
Alina Peluso
Josh Grant
Joe Tuccillo
Adam Spannaus
Jeremy Logan
Michael McGee
Kelly Callaway
Midgie MacFarland
Hope Cook
Heidi Hanson
Susana Martins
Jodie Trafton
Anuj Kapadia

December 2024

OAK RIDGE National Laboratory

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

ORNL/SPR-2024/3689
RES PUB ID 226014

Computational Sciences and Engineering Division

VA EDH DATA CURATION DOCUMENTATION FY25-Q1

Hilda Klasky
Kevin Sparks
Alina Peluso
Josh Grant
Joe Tuccillo
Adam Spannaus
Jeremy Logan
Michael McGee
Kelly Callaway
Midgie MacFarland
Hope Cook
Heidi Hanson
Susana Martins
Jodie Trafton
Anuj Kapadia

December 2024

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

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.

Health outcomes, including suicide, are typically influenced by both genetics and environmental factors, such as air quality, transportation access, food availability, homelessness, and more. Mental health outcomes are associated with various stressors across socioeconomic, economic, and physical environments. Analyzing the connections between these stressors, covariates, and health outcomes relies on standardized data, which can be integrated into models like the VA's Recovery Engagement and Coordination for Health, Veterans Enhanced Treatment (REACH VET).

The World Health Organization (WHO) defines Environmental Determinants of Health (EDH) as factors like clean air, stable climate, water and sanitation, chemical safety, radiation protection, safe workplaces, sustainable agriculture, healthy urban environments, and nature preservation, all of which are crucial for good health.

1.1 BACKGROUND

With funding from the VA Office of Mental Health and Suicide Prevention (OMHSP), the 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) Social Determinants of Health (SDoH) covariates, addressing key gaps by introducing finer spatial resolution (Census Tract) and additional environmental covariates.

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 air pollutants from the US Environmental Protection Agency (EPA) and 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 EDH 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 data source documentation report provides researchers with valuable insights into the structure, contents, and data sources used to compile the datasets. It specifically covers the Fiscal Year 2024, Fourth Quarter (FY25-Q1) dataset curation documentation for the Environmental Determinants of Health (EDH) project.

Change Log

1. **Table Renaming**
    - OMHSP_FY23Q1.opioid_prescribing_rate_county_2019 renamed to OMHSP_FY23Q1.opioid_dispensing_rate_county_2019.
    - OMHSP_FY23Q2.ruca_county_2013 renamed to OMHSP_FY23Q2.rucc_county_2013.
2. **Metadata Updates**
    - Updated all renamed table references in the following metadata tables:
        - [OMHSP].[SEDH_table_metadata] (table_id, table_name).
        - [OMHSP].[SEDH_column_metadata] (table_id).
    - Updated [OMHSP].[SEDH_reports].

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

1. **Medicare Part D Opioid Prescribing Rates at the county level for 2013-2022** (*update*)
2. **Rural-Urban Continuum Codes at the county level for 2023** (*update*)
3. **Drive Time Data at county level for 2024** (*new*)
4. **Food Pantry List at the census tract level for 2024** (*new*)
5. **Food Bank List at the census tract level for 2024** (*new*)
6. **Homeless Shelter List Data, version 2, at the census tract level for 2024** (*update*)

These datasets are only available for the VA's EDH project. Please contact your program manager for access of these datasets.

2.1 RECOMMENDED CITATION FOR FY25-Q1 DATA CURATION DOCUMENTATION'S SPONSOR REPORT

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.

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

2.2 PREVIOUS DOCUMENT RELEASES

Since the inception of the EDH 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 EDH 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 Inequality (*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)
  - Social Capital Atlas - Civil Engagement (*resolution:* county, 2022, *source:* Social Capital Atlas)
  - Social Capital Atlas - Cohesiveness (*resolution:* county, 2022, *source:* Social Capital Atlas)
  - Social Capital Atlas - Economic Connectedness (*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 Attainment (*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 and Climatological 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)

Please note that the URL for the FY25-Q1 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 EDH 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 EDH datasets*: Lists the social or environmental determinants of health domains to which the data source has contributed variables. Includes additional information relevant to the EDH 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 EDH dataset 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 SEDH 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 SEDH datasets serve as a consistent and user-friendly resource for researchers, we adhered to the following conventions:

- **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 ID variables, all variable names begin with a data source acronym, followed by an underscore and a descriptive title.
- **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

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. SEDH stands for the Social and Environmental Determinants of Health repository.
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.SEDH_table_metadata Table

This table contains the following columns:

1. **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.
2. **schema_name**: Quarterly release schema names in the database (e.g., OMHSP_FY22Q4, OMHSP_FY23Q1, OMHSP_FY23Q2, OMHSP_FY23Q3, and so on).
3. **table_name**: The table name as it appears in the MS SQL Server database.
4. **table_name_description**: A description of the table name.
5. **availability_across_years**: The years for which data is available.
6. **data_source**: The name of the source organization that provided the raw data (starting in FY23Q4).
7. **data_source_description**: Description of the source organization (starting in FY23Q4).
8. **data_source_url**: URL of the source organization (starting in FY23Q4).
9. **spatial_resolution**: Spatial resolution or geography (e.g., state, county, block group, census tract, and zip code) (starting in FY23Q4).
10. **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.
11. **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).
12. **dimension**: Two option values: social and environmental (started in FY24Q2).
13. **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).
14. **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).
15. **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.SEDH_column_metadata Table

This table contains the following columns:

1. **column_id**: A sequential number.

2. **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.
3. **column_name**: Column names within each dataset as they appear in the MS SQL Server table.
4. **column_name_description**: Descriptions of each column name.
5. **column_type**: The column type in the MS SQL Server table.
6. **column_length**: The column length in the MS SQL Server table.
7. **total_rows**: The number of rows in each column in each dataset (starting in FY23Q2).
8. **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.SEDH_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.

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:

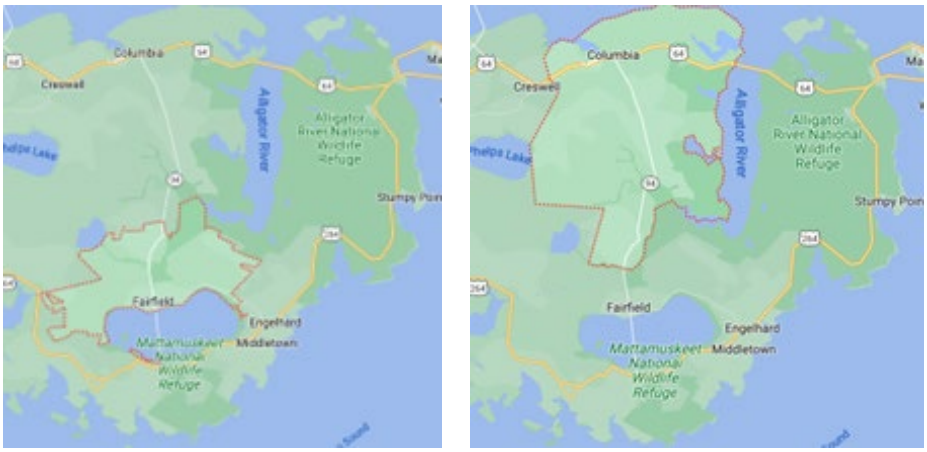


Figure 1. *Example*: - *Left*: Zip Code 27826 - *Right*: County FIPS 37177

3.7 HOW TO LINK COHORTS TO OUR DATASETS

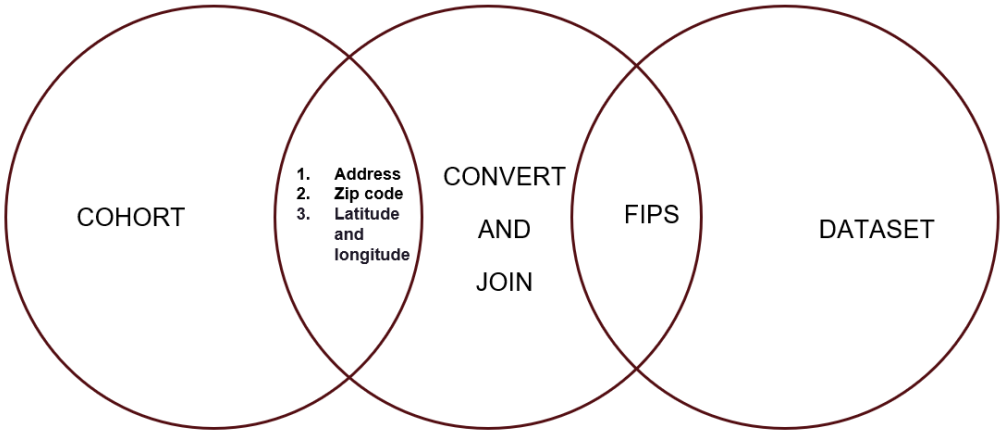


Figure 2. Joining geoids to our datasets

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 social and environmental determinant of health 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 social and environmental 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 FY25Q1 follows:

			Development		QA-Intra		QA-KDI (VIEWS)		Production (Transmit)		
Dataset Name	Rows	Column	Passes	Fails	Passes2	Fails3	Passes4	Fails5	Passes6	Fails7	Error ratio
[OMHSP].[SEDH_table_metadata]	65	15	4	1	4	1	5	0	5	0	0.11
[OMHSP].[SEDH_column_metadata]	938	8	4	1	4	1	5	0	5	0	0.11
[OMHSP].[SEDH_reports]	13	7	4	1	4	1	5	0	5	0	0.11
[OMHSP_FY25Q1].[foodbanklist_tract_2024]	11202	5	5	0	5	0	5	0	5	0	0
[OMHSP_FY25Q1].[foodpantrylist_tract_2024]	16005	5	5	0	5	0	5	0	5	0	0
[OMHSP_FY25Q1].[homeless_shelterlist_v2_tract_2024]	3277	7	5	0	5	0	5	0	5	0	0
[OMHSP_FY25Q1].[opioid_prescribing_rates_county_2013_2022]	31714	13	4	1	4	1	5	0	5	0	0.11
[OMHSP_FY25Q1].[rucc_county_2023]	3233	4	4	1	4	1	5	0	5	0	0.11
[OMHSP_FY25Q1].[VADriveTimeData_county_2024]	12643488	9	5	0	5	0	5	0	5	0	0

Appendix A presents descriptive statistics of error-checking results.

4. MEDICARE PART D OPIOID PRESCRIBING RATES

4.1 DATA SOURCE

Centers for Medicare and Medicaid Services

4.2 DESCRIPTION

In FY25Q1, we extended our opioid prescription dataset to include data for 2021 and 2022. The original dataset, delivered in FY23Q2, covered 2013 to 2020. To ensure consistency between the updates, we performed a two-part assessment:

1. County FIPS Code Consistency

- We identified missing county FIPS codes between the datasets for years 2013–2020.
- Out of more than 25,000 county-year combinations:
- 76 pairs were present in FY23Q2 but missing in FY25Q1.
- 83 pairs were present in FY25Q1 but missing in FY23Q2.
- Most discrepancies were due to updates in Census FIPS codes, particularly in Connecticut.

2. Data Value Consistency

- We checked for retrospective changes to data values between updates.
- Approximately 1% of counties (36 in total) showed modified values in the updated dataset.

Due to these differences, we will deliver a fully updated dataset in FY25Q1, with data from 2013-2022, to ensure the most accurate and consistent data for all users.

Additional Notes

- This dataset is a historical update of the 2020 data from the Centers for Medicare and Medicaid Services (CMS) and provides data from 2013 onward.
- The previous FY22Q3 release included only CMS 2019 opioid prescribing rates and had 25,341 rows. This updated dataset contains 31,714 additional rows.

Dataset Details

- The dataset provides the geographic distribution of retail opioid prescriptions dispensed per 100 people per year at the county level in the United States.
- Both first-time prescriptions and refills are included as medicines dispensed.
- The age denominator is not explicitly defined in the source documentation, but age groups include values from 0 to 65 and older.
- It is worth noting the definitions of the following columns, which are not intuitively deduced from the column names:
  - opioid_prscrbng_rate_5y_chg: The percentage point difference in the opioid prescribing rate from five years previous to the data year, which is calculated by subtracting the rate five years previous from the rate in the data year.
  - opioid_prscrbng_rate_1y_chg: The percentage point difference in the opioid prescribing rate from one year previous to the data year, which is calculated by subtracting the rate one year previous from the rate in the data year.

- la_tot_opioid_clms: The number of Medicare Part D opioid drug claims that are considered long-acting, including original prescriptions and refills.
- la_opioid_prscrbng_rate: The number of Long-Acting Opioid Claims divided by the Opioid Claims and multiplied by 100.
- la_opioid_prscrbng_rate_5y_chg: The percentage point difference in the long-acting opioid prescribing rate from five years previous to the data year, which is calculated by subtracting the rate five years previous from the rate in the data year.
- la_opioid_prscrbng_rate_1y_chg: The percentage point difference in the long-acting opioid prescribing rate from one year previous to the data year, which is calculated by subtracting the rate one year previous from the rate in the data year.

4.3 INCLUSION

Year: 2013-2022 data.
Geographical unit: FIPS County Level, Continental US.

4.4 RESOURCES

For more information on the Medicare Part D Opioid Prescribing Rate:

- CMS Data Portal

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. Medicare Part D Opioid Prescribing Rates (OPR)

variable name	description	total_rows	null_rows
fips	Federal Information Processing Standards (FIPS) code at county level.	31714	0
year	The year of the data.	31714	0
tot_prscrbrs	The number of providers prescribing Medicare Part D drugs.	31714	42
tot_opioid_prscrbrs	The number of Medicare Part D providers prescribing opioid drugs.	31714	42
tot_opioid_clms	The number of opioid drug claims under Medicare Part D, including initial prescriptions and refills.	31714	243
tot_clms	The number of Medicare Part D drug claims, including initial prescriptions and refills.	31714	42
opioid_prscrbng_rate	The number of Opioid Claims divided by the Overall Claims and multiplied by 100.	31714	243
opioid_prscrbng_rate_5y_chg	Opioid prescribing rate 5-year change.	31714	16135
opioid_prscrbng_rate_1y_chg	Opioid prescribing rate 1-year change.	31714	3498
la_tot_opioid_clms	Long acting total opioids claims.	31714	1489
la_opioid_prscrbng_rate	Long acting opioid prescribing rate.	31714	1633
la_opioid_prscrbng_rate_5y_chg	Long acting opioid prescribing rate 5-year change.	31714	17380
la_opioid_prscrbng_rate_1y_chg	Long acting opioid prescribing rate 1-year change.	31714	5246

Table 2. Variable availability across years, (OPR)

variable name	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
fips	X	X	X	X	X	X	X	X	X	X
year	X	X	X	X	X	X	X	X	X	X
tot_prscrbrs	X	X	X	X	X	X	X	X	X	X
tot_opioid_prscrbrs	X	X	X	X	X	X	X	X	X	X
tot_opioid_clms	X	X	X	X	X	X	X	X	X	X
tot_clms	X	X	X	X	X	X	X	X	X	X
opioid_prscrbng_rate	X	X	X	X	X	X	X	X	X	X
opioid_prscrbng_rate_5y_chg	X	X	X	X	X	X	X	X	X	X
opioid_prscrbng_rate_1y_chg	X	X	X	X	X	X	X	X	X	X
la_tot_opioid_clms	X	X	X	X	X	X	X	X	X	X
la_opioid_prscrbng_rate	X	X	X	X	X	X	X	X	X	X
la_opioid_prscrbng_rate_5y_chg	X	X	X	X	X	X	X	X	X	X
la_opioid_prscrbng_rate_1y_chg	X	X	X	X	X	X	X	X	X	X

5. RURAL-URBAN CONTINUUM CODES

5.1 DATA SOURCE

USDA Economic Research Service and the US Department of Agriculture

5.2 DESCRIPTION

This dataset includes the 2023 Rural-Urban Continuum Codes, developed by the USDA Economic Research Service and the U.S. Department of Agriculture. It supersedes the 2013 version, last updated on 12/10/2020, and was provided to the VA in the FY23Q2 delivery.

Overview of the 2023 Rural-Urban Continuum Codes - These codes classify U.S. counties based on: - Metro population size. - Urbanization level for nonmetropolitan counties. - Proximity of nonmetropolitan counties to metro areas. - The classification uses three metro and six nonmetro categories, offering detailed county-level insights beyond a simple metro/nonmetro division.

Changes in the 2023 Update - Aligns with the 2023 Office of Management and Budget (OMB) delineation. - Adopts updated U.S. Census Bureau standards, raising the urban area population threshold from 2,500 to 5,000 people. - Reflects changes to urban and metro criteria, which may affect comparability with prior versions.
These updates provide a more nuanced analysis of rural and urban trends, supporting research and policy efforts.

5.3 INCLUSION

Year: 2023 data.
Geographical unit: FIPS County Level, Continental US.

5.4 RESOURCES

For more information on the Rural-Urban Continuum Codes, see:
https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx

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. Rural-Urban Continuum Codes (RUCC)

variable name	description	total_rows	null_rows
fips	Federal Information Processing Standards (FIPS) code at county level.	3233	0
population_2020	The population of the county in 2020 according to the US Census Bureau.	3233	0
rucc_2023	The rural-urban continuum code as of 2023.	3233	0
rucc_description	The rural-urban continuum code classification description.	3233	0

Table 4. Variable availability across years, (RUCC)

variable name	2023
fips	X
population_2020	X
rucc_2023	X
rucc_description	X

6. DRIVE-TIME

6.1 DATA SOURCE

VA and ORNL.

6.2 DESCRIPTION

The Drive-Time dataset provides driving times and driving distances from an origin address to the nearest US Department of Veterans Affairs (VA) locations or healthcare facilities.
This dataset contains 12,643,488 rows. Of these, 922 rows have null ZIP code values. Some rows also include "unknown" values for non-essential variables, whereas all essential variables are fully populated. In this FY25Q1 release, county-level Federal Information Processing Standard (FIPS) codes have been added.

Notes on Naming Conventions

Variables joined from other tables retain their original naming conventions (e.g., camel case) to ensure consistency with VA's Corporate Data Warehouse (CDW). Newly created variables, however, follow the standard format used in this dataset (i.e., lower case with underscores).

How the Nearest Facility is Identified
Starting from an origin location, the three nearest facilities are first identified based on the straight-line distance between points. This process calculates the distances between pairs of latitude and longitude coordinates. Once the three closest facilities are identified, routing algorithms are applied to an OpenStreetMap road network to compute driving routes from the origin to each facility. The algorithm estimates the driving time for each route, and the facility with the shortest driving time is selected, while the other two are disregarded.

This dataset is exclusively available for the VA's EDH project. For further inquiries, please contact your program manager.

6.3 INCLUSION

Year: 2024
Geographical unit: FIPS at county level of the origin address. Continental USA.
Data included: Selected FIPS at county level of the origin addresses.

6.4 RESOURCES

Link to VA Locations
Link to OpenStreetMap

6.5 UPDATE FREQUENCY

This dataset will be updated as requested by the sponsor.

Table 5. Drive-Time (DRIVETIME)

variable name	variable label
fips	County fips code of the origin address.
State	State of the origin address.
County	County name of the origin address.
StreetAddress1	Origin address.
City	City of the origin address.
Zip	Zip code of the origin address.
drive_time(min)	Drive time in minutes.
drive_distance(km)	Drive distance in kilometers.
FacilityID	The VA's Facility identification number.

Table 6. Variable availability across years, (DRIVETIME)

variable name	2024	total rows	null rows
fips	X	12643488	0
State	X	12643488	0
County	X	12643488	0
StreetAddress1	X	12643488	0
City	X	12643488	0
Zip	X	12643488	922
drive_time(min)	X	12643488	0
drive_distance(km)	X	12643488	0
FacilityID	X	12643488	0

7. FOOD PANTRY LIST DATA

7.1 DATA SOURCE

https://www.foodpantries.org/

7.2 DESCRIPTION

FoodPantries.org is a comprehensive directory of food pantries, food banks, soup kitchens, and non-profit organizations in the United States dedicated to combating hunger. The website provides detailed information about local resources, including addresses, hours of operation, contact details, and services offered. Each entry is manually curated to ensure the information is as accurate and current as possible. By connecting individuals and families with essential food resources in their communities, FoodPantries.org supports food security and assists organizations in their mission to reduce hunger. However, as a crowdsourced platform, some data may be unstructured or contain inaccuracies and entry errors.

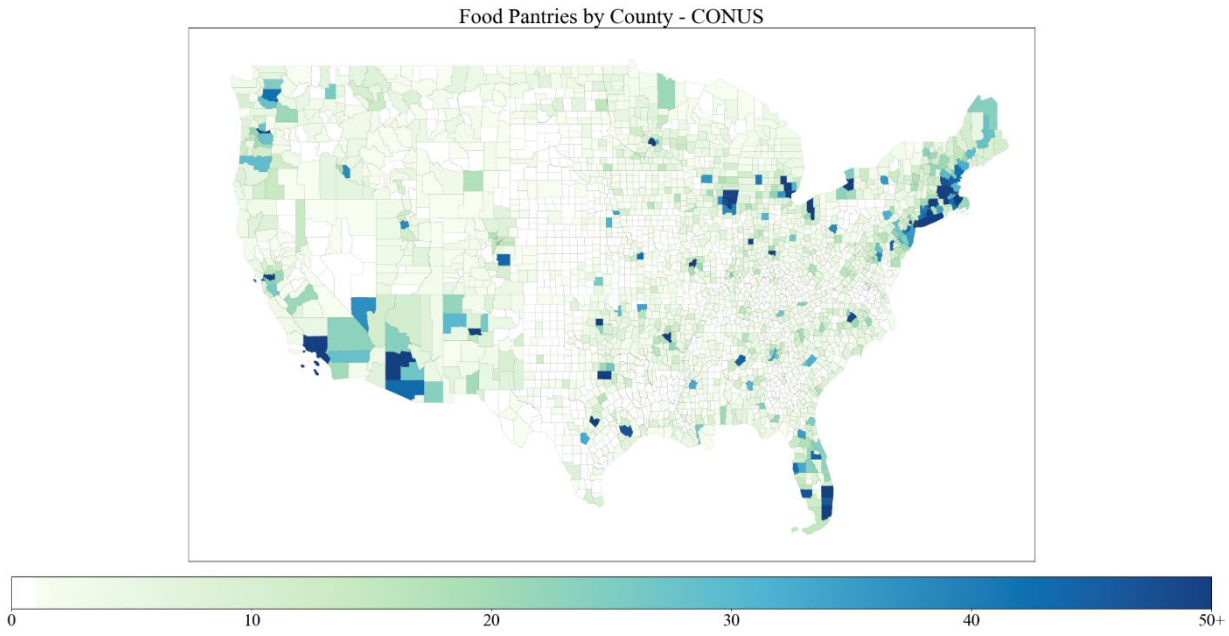
Definitions

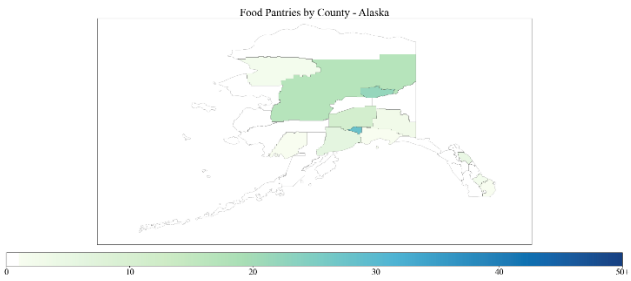
- A food bank is a warehouse and distribution center where food is collected, stored, and distributed to local pantries and soup kitchens.

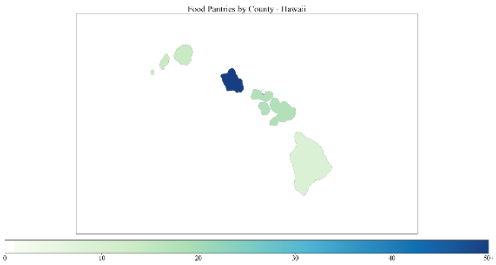
- A food pantry is a charity that directly distributes food and groceries to those in need.

Data Preparation

For this dataset, addresses and phone numbers have been standardized, and duplicates removed to enhance data accuracy. The census tract FIPS code is not included with this data by default. It was computed using the street address as input, applying the same geocoding process used in other deliverables to derive the census tract FIPS.







7.3 INCLUSION

Year: 2024
Geographical unit: census tract level data.
Continental US.

7.4 RESOURCES

Home: https://www.foodpantries.org/ar/about

7.5 UPDATE FREQUENCY

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

Table 7. Food Pantry List Data (PANTRYLIST)

variable name	variable label
fips	Census tract-level Federal Information Processing Standards (FIPS) code for the food pantry's address.
name	The official name of the food pantry.
address	The physical location of the food pantry, including street, city, and state.
phone_number	Contact number for reaching the food pantry.
description	Brief overview or details about the services or offerings of the food pantry.

Table 8. Variable availability across years, (PANTRYLIST)

variable name	2024	total rows	null rows
fips	X	16005	0
name	X	16005	0
address	X	16005	0
phone_number	X	16005	0
description	X	16005	238

8. FOOD BANK LIST DATA

8.1 DATA SOURCE

https://www.freefood.org/

8.2 DESCRIPTION

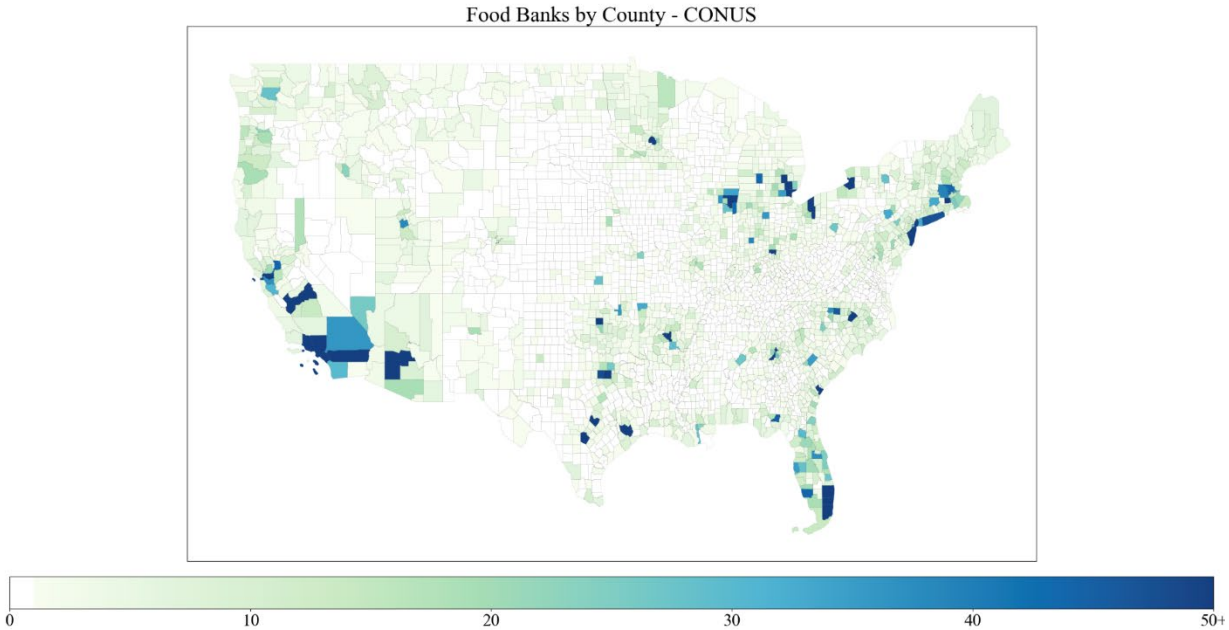
FreeFood.org is an anonymous organization dedicated to helping people find and add free food resources both locally and nationwide. The organization is not affiliated with any government agency or nonprofit organization.

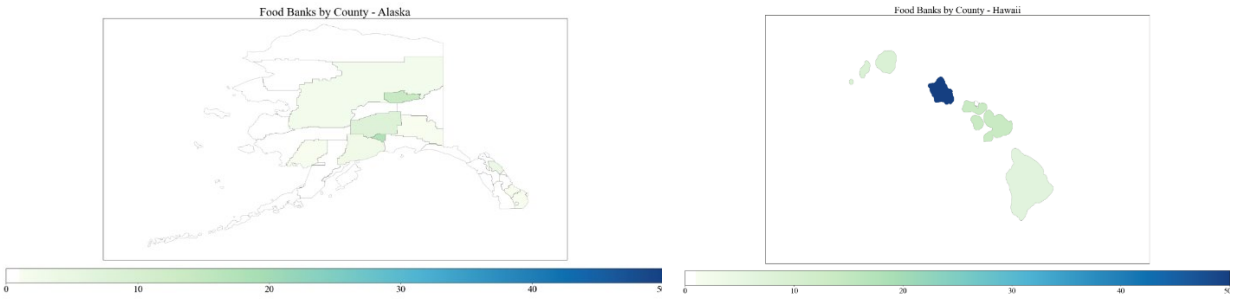
What FreeFood.org Does
FreeFood.org provides a list of free food centers across the U.S. and allows users to add missing food resources to its database.

- A **food bank** is a warehouse and distribution center where food is collected, stored, and distributed to local pantries and soup kitchens.

- A **food pantry** is a charity that directly distributes food and groceries to those in need.
These organizations differ in purpose and structure, and FreeFood.org's listings reflect this distinction. Some data sources prioritize food pantries, while others emphasize food banks.

Data Preparation
For this dataset, addresses and phone numbers have been standardized to ensure consistency, and duplicates have been removed to improve accuracy. Census tract FIPS codes have also been included for each location. These codes were not part of the original data but were computed using the street addresses and a geocoding process consistent with other deliverables.





8.3 INCLUSION

Year: 2024
Geographical unit: census tract level data.
Continental US.

8.4 RESOURCES

Home: https://www.freefood.org/

8.5 UPDATE FREQUENCY

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

Table 9. Food Bank List Data (FOODBANKLIST)

variable name	variable label
fips	Census tract-level Federal Information Processing Standards (FIPS) code for the food bank's address.
name	The official name of the organization that provides free food.
address	The physical location of the organization that provides free food, including street, city, and state.
phone_number	Contact number for reaching the organization that provides free food.
description	Brief overview or details about the organization that provides free food.

Table 10. Variable availability across years, (FOODBANKLIST)

variable name	2024	total rows	null rows
fips	X	11202	0
name	X	11202	1
address	X	11202	0
phone_number	X	11202	0
description	X	11202	534

9. HOMELESS SHELTER LIST DATA

9.1 DATA SOURCE

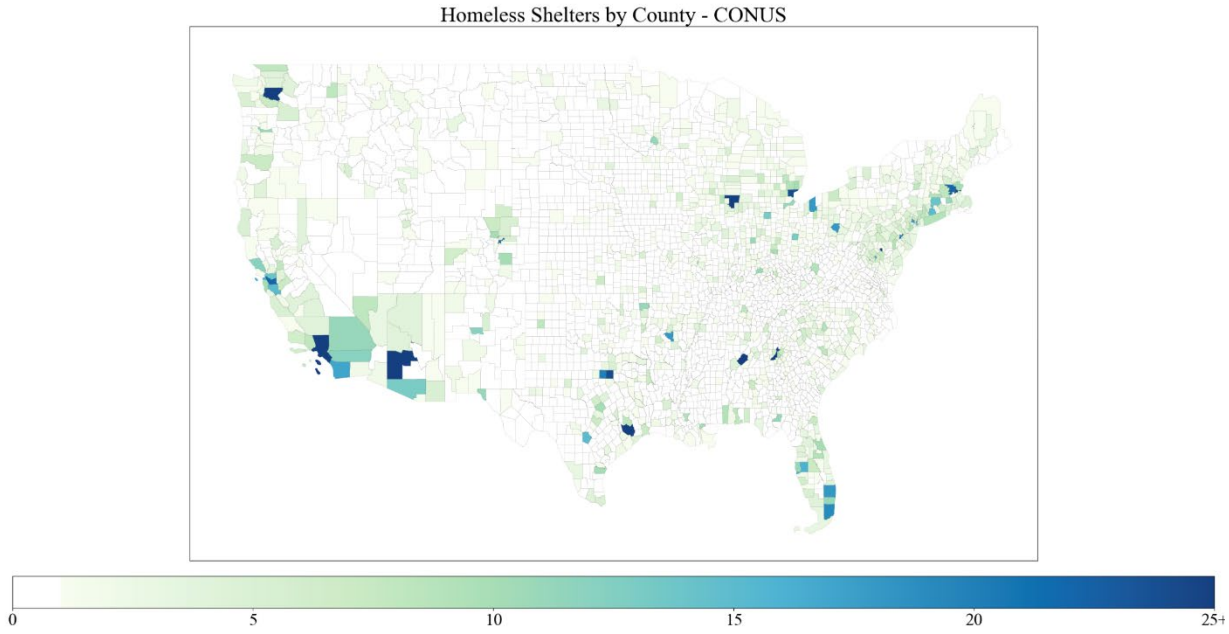
ShelterList.com

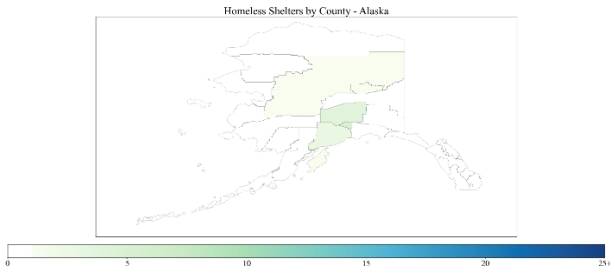
9.2 DESCRIPTION

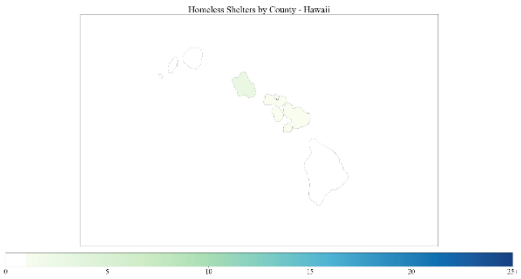
The homeless shelter dataset was initially delivered in FY24Q4. We provide an updated version of the data, processed with an enhanced algorithm that leverages Nominatim.org and OpenStreetMap data, along with custom code for address formatting and standardization, and census tract mapping. As a result, the number of homeless shelter entries has increased from 2,308 in FY24Q4 to 3,277 in this release. Therefore, the previous dataset (FY24Q4) should now be considered deprecated. And this dataset is considered version 2 or v2.

This dataset was sourced from ShelterList.com, a comprehensive national resource for locating emergency and transitional housing for individuals and families across the USA. ShelterList.com includes detailed information on various types of shelters, such as homeless shelters, women's shelters, family shelters, and youth shelters.

The website categorizes shelters by city and state and provides specific details including contact information, available services, eligibility requirements, and operating hours. ShelterList.com's primary mission is to assist those in need by quickly connecting them with safe and reliable shelter options. For this dataset, addresses and phone numbers have been standardized, and duplicates have been removed to ensure data accuracy.







9.3 INCLUSION

Year: 2024
Geographical unit: census tract level data.
Continental US.

9.4 RESOURCES

Home: https://www.shelterlist.com/page/about

9.5 UPDATE FREQUENCY

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

Table 11. Homeless Shelter List Data (HOMELESS_SHELTERLIST)

variable name	variable label
fips	Federal Information Processing Standards (FIPS) code at census tract level.
name	Name of shelter facility.
added_date	Date that shelter was added to website.
updated_date	Most recent date that shelter info was updated.
address	Street address of shelter.
phone_number	Telephone number for shelter.
description	Information listed on website to describe shelter (non-standard, info included may range from no data to populations served, list of services, available resources, etc.).

Table 12. Variable availability across years, (HOMELESS_SHELTERLIST)

variable name	2024	total rows	null rows
fips	X	3277	0
name	X	3277	0
added_date	X	3277	2
updated_date	X	3277	1
address	X	3277	0
phone_number	X	3277	0
description	X	3277	61

10. 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

APPENDIX A. ERROR CHECKING

APPENDIX A. ERROR CHECKING

The following descriptive statistics were obtained using the pandas df.describe function. The definitions of the main included statistics are as follows::

- **count**: Number of non-null entries in each column.
- **unique**: Number of unique entries in each column
- **top**: Most frequent entry
- **freq**: Frequency of the most common entry

OMHSP_FY25Q1.opioid_prescribing_rates_county_2013_2022

	fips	year	tot_prscrbrs	tot_opioid_prscrbrs	tot_opioid_clms	tot_clms	opioid_prscrbng_rate	opioid_prscrbng_rate_5y_chg	opioid_prscrbng_rate_1y_chg	la_tot_opioid_clms	la_opioid_prscrbng_rate	la_opioid_prscrbng_rate_5y_chg	la_opioid_prscrbng_rate_1y_chg
count	31714	31714	31672	31672	31471	31672	31471	15579	28216	30225	30081	14334	26468
unique	3189												
top	12086												
freq	19												
mean		2017.5030585861134	373.71413235665574	254.42302349078 05	22723.99710844905	466249.2937926244	4.850322836897462	-1.387506	-0.234578	2789.3285028949545	10.285122835012134	-2.625645	-0.442273
std		2.8732250470959158	1342.0088194060165	869.06349278 91017	60294.606387 43187	138783.0811 54134	2.303897041952321	2.495660715031962	1.048575733037 3466	7352.7414727 17101	6.180301875 7691595	6.292362094 9549965	2.284886936 5462265
min		2013	1	0	0	11	0	-57.11	-35.46	0	0	-73	-28.44
25%		2015	16	13	1796	41960.5	3.5	-2.41	-0.47	141	5.61	-6.07	-1.23
50%		2018	47	36	5280	116227	4.66	-1.28	-0.21	487	9.83	-2.66	-0.37
75%		2020	181	130	17130	337294	5.98	-0.29	0.02	1902	14.27	0.61	0.35
max		2022	36568	22396	166872 5	419438 67	63.14	48.17	52.01	163821	81.82	38.1	36.67

OMHSP_FY25Q1.rucc_county_2023

	fips	population_2020	rucc_2023	rucc_description
count	3233	3233	3233	3233
unique	3233	3172	9	9
top	01001	34668	9	Nonmetro - Urban population of fewer than 5,000, not adjacent to a metro area
freq	1	3	595	595

OMHSP_FY25Q1.foodpantrylist_tract_2024

	name	address	phone_number	description	fips
count	16005	16005	16005	15767	16005
unique	15181	15856	15001	14168	12312
top	Salvation Army	6735 Black Horse Pike, Egg Harbor Township, NJ 08234	No phone number found	Provides a food pantry.	022900002000
freq	47	3	76	514	17

OMHSP_FY25Q1.foodbanklist_tract_2024

	name	address	phone_number	description	fips
count	11201	11202	11202	10668	11202
unique	10609	11147	10560	10259	8847
top	First Baptist Church	2203 San Antonio Street Austin, TX 78705	No phone number found	For more information, please call them.	05119004400
freq	30	3	86	104	18

OMHSP_FY25Q1.homeless_shelterlist_v2_tract_2024

	name	added_date	updated_date	address	phone_number	description	fips
count	3277	3275	3276	3277	3277	3216	3277
unique	3224	884	229	3252	3092	3199	2747
top	The Salvation Army	11/21/2018 0:00	6/6/2022 0:00	901 Swartswood Road Newton, NJ 07860	No phone number listed	Emergency Shelter	22033005300
freq	6	27	99	2	4	4	7