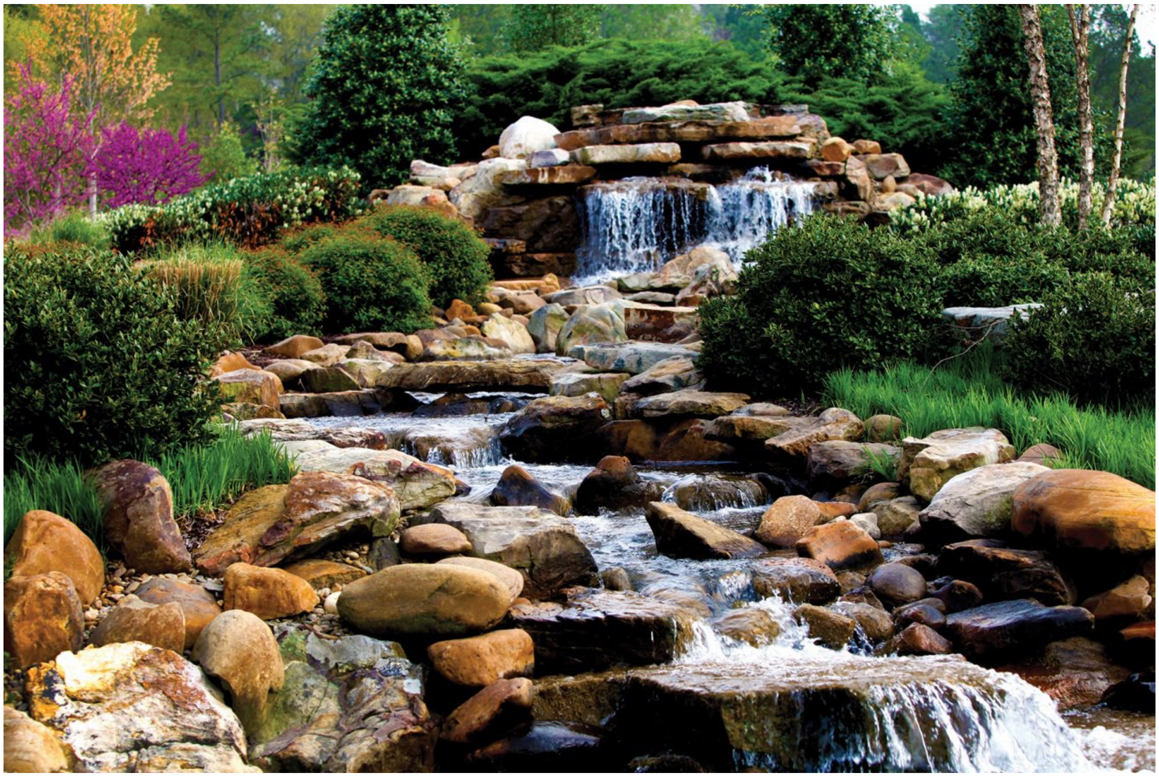
ORNL/SPR-2024/3419
RES PUB ID 215289

VA EDH Data Curation Documentation FY24-Q3



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

June 2024

OAK RIDGE
National Laboratory

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

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

ORNL/SPR-2024/3419
RES PUB ID 215289

Computational Sciences & Engineering Division

VA EDH DATA CURATION DOCUMENTATION FY24-Q3

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

June 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 is designed to provide researchers with valuable insights into the structure, contents, and the data sources utilized to compile the datasets. It specifically covers the Fiscal Year 2024, Third Quarter (FY24-Q3) dataset curation documentation for the Environmental Determinants of Health (EDH) project.

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

1. **High Intensity Drug Trafficking Areas (HIDTA) at state-level for 2022 - 2023 (new)**
2. **High Intensity Drug Trafficking Areas (HIDTA) at county-level for 2022 - 2023 (new)**
3. **Drive-time sample at selected lat/long points in the state of Tennessee for 2024 (new)**

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 FY24-Q3 DATA CURATION DOCUMENTATION'S SPONSOR REPORT

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., Stratford, J., and Kapadia, A. VA EDH Data Curation Documentation (FY24-Q3). United States: N. p., 2024. ORNL/SPR-2024/3419 PUB ID 215289.

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, 2017-2021, by county (new)**
  - Veterans Service Organizations (VSO) 2010-2022, by state
  - Veterans Service Organizations (VSO) 2010-2022, by county
  - Veterans Service Organizations (VSO) 2010-2022, by zip code

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

- Link to Documentation
- Included Datasets:
  - HUD USPS Zip Code Crosswalk Files, ZIP-to-tract for 2023
  - HUD USPS Zip Code Crosswalk Files, ZIP-to-county for 2023
  - Social Capital Index 2019, by county for 2019 (updated)
  - High Intensity Drug Trafficking Areas (HIDTA) state-level for 2018 - 2021 (re-delivered)
  - High Intensity Drug Trafficking Areas (HIDTA) county-level for 2018 - 2021 (re-delivered)

Please note that the URL for the FY24-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 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.

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:

- **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).

- **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.

3.5 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.

3.6 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.

3.7 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:

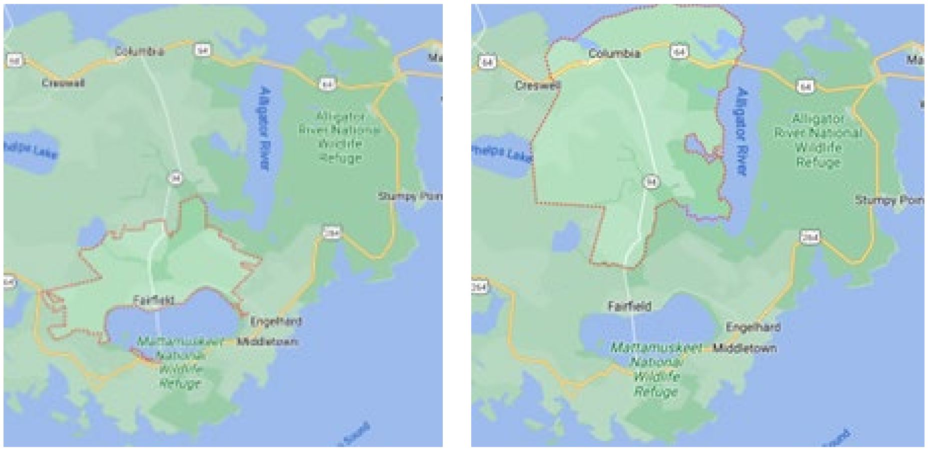


Figure 1, *Example*: - *Left*: Zip Code 27826 - *Right*: County FIPS 37177

3.7.1 How to link cohorts to our datasets

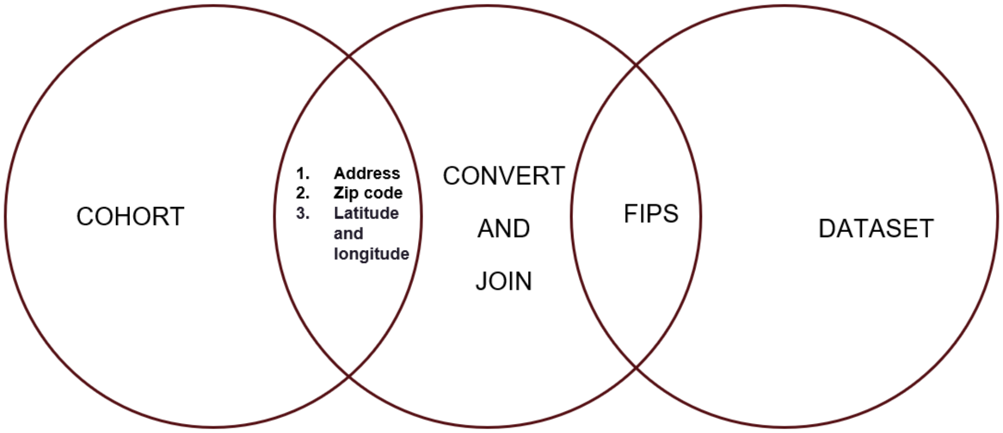


Figure 2 ***Joining geoids to our datasets.***

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.

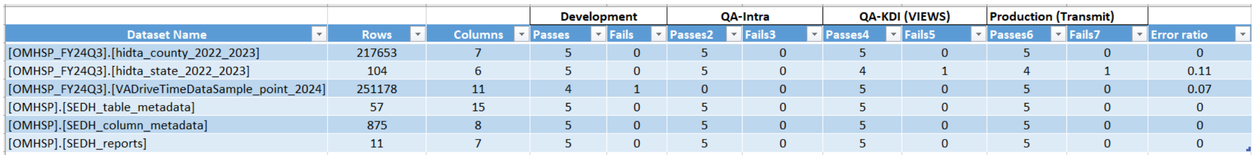
All datasets were error-checked using a data profiling strategy that includes at least two reviewers and the following test groups:

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 FY24Q3 follows:



			Development		QA-Intra		QA-KDI (VIEWS)		Production (Transmit)		
Dataset Name	Rows	Columns	Passes	Fails	Passes2	Fails3	Passes4	Fails5	Passes6	Fails7	Error ratio
[OMHSP_FY24Q3].[hidta_county_2022_2023]	217653	7	5	0	5	0	5	0	5	0	0
[OMHSP_FY24Q3].[hidta_state_2022_2023]	104	6	5	0	5	0	4	1	4	1	0.11
[OMHSP_FY24Q3].[VADriveTimeDataSample_point_2024]	251178	11	4	1	0	0	5	0	5	0	0.07
[OMHSP].[SEDH_table_metadata]	57	15	5	0	5	0	5	0	5	0	0
[OMHSP].[SEDH_column_metadata]	875	8	5	0	5	0	5	0	5	0	0
[OMHSP].[SEDH_reports]	11	7	5	0	5	0	5	0	5	0	0

Appendix A presents descriptive statistics of error-checking results.

4. DRIVE-TIME SAMPLE

4.1 DATA SOURCE

VA and ORNL

4.2 DESCRIPTION

The Drive-Time dataset provides driving times from an origin address to the nearest US Department of Veterans Affairs (VA) locations or healthcare facilities. This is an alpha version sample table for the state of TN. Because this is a proof of concept, this data should not be used for any analysis.

4.2.1 How the nearest facility is identified.

Starting from an origin location, we first identify the three nearest facilities based on straight-line distance between two points. This involves calculating the distance between pairs of latitude and longitude coordinates. Once the three closest facilities are determined, we employ routing algorithms on an OpenStreetMap road network to establish the driving route from the origin to each of these facilities. The routing algorithm provides the estimated driving time for each route. We then select the route with the shortest driving time and disregard the other two.

For verification purposes, while this dataset is being refined, only a subset of entries for the state of Tennessee (TN) is included in this release.

This dataset is only available for the VA's EDH project. Please contact your program manager if you have any questions. This dataset is depicted in Figure 1 below.

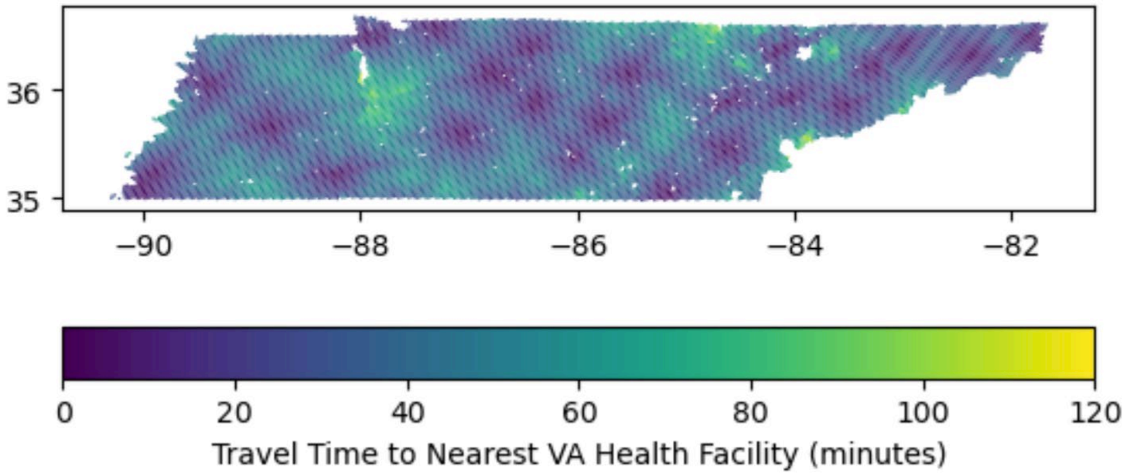


Figure 3. Travel time to nearest VA Healthcare Facility in TN (minutes).

4.3 INCLUSION

Year: 2024

Geographical unit: Point, lat/long of the origin address.

Data included: Selected latitude and longitude points in the state of TN. Total rows: 251178. Null rows: 0.

4.4 RESOURCES

Link to VA Locations

Link to OpenStreetMap

4.5 UPDATE FREQUENCY

This dataset will be updated as requested by the sponsor.

Table 1 . Drive-Time Sample (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.
Longitude	Longitude of the origin address.
Latitude	Latitude of the origin address.
DriveTime	Drive time in minutes.
FacilityID	The VA's Facility identification number.

Table 2 . Variable availability across years, (DRIVETIME)

variable name	**2024**
fips	X
State	X
County	X
StreetAddress1	X
City	X
Zip	X
Longitude	X
Latitude	X
DriveTime	X
FacilityID	X

5. HIGH INTENSITY DRUG TRAFFICKING AREAS (HIDTA)

5.1 DATA SOURCE

Washington/Baltimore High Intensity Drug Trafficking Areas Program

5.2 DESCRIPTION

The High Intensity Drug Trafficking Areas (HIDTA) program, established by Congress with the Anti-Drug Abuse Act of 1988, offers support to federal, state, local, and tribal law enforcement agencies working in locations identified as important drug-trafficking areas in the United States. The data was obtained from the High Intensity Drug Trafficking Areas Program's activities.

Previously, ORNL delivered the 2018-2021 HIDTA datasets to the VA in the second quarter of the Fiscal Year 2024 (the FY24Q2 delivery). This release includes data for both 2022 and 2023. The data is transferred to the VA as received from the source. However, at our sponsor's request, we provide two datasets: one with county-level data, and another with state-level data. The county-level data includes all US states, but Michigan. The state-level data, includes data from the state of Michigan data, which did not provide county-level data.

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

For information about the use of HIDTA data, read the section "Publishing Using PMP Data" in the following document: https://www.hidta.org/wp-content/uploads/2021/04/PERFORMANCE-MANAGEMENT-PROCESS-Research-Guidance-and-Procedures-1-1.pdf

5.3 INCLUSION

Year: from 2022 to 2023.

Geographical unit: county level, and the state of Michigan at state level.

State-level dataset: total rows: 104, null rows: 0.

County-level dataset: total rows: 217653, null rows: 0.

5.4 RESOURCES

For more information about HIDTA:

HIDTA Source Link

HIDTA Performance Management Process – Using HIDTA Data

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 . High Intensity Drug Trafficking Areas (HIDTA) (HIDTA)

variable name	**variable label**
fips	Federal Information Processing Standards (FIPS), county or state level fips codes.
county	County name.
state	US state name.
seizure_date	The date of seizure.
drug	The type of drug seized.
quantity	The quantity of drugs seized.
unit	The weight unit of measurement (kilogram - Kg, or deci atomic mass unit = D.U.)

Table 4 . Variable availability across years, (HIDTA)

variable name	**2022**	**2023**
fips	X	X
county	X	X
state	X	X
seizure_date	X	X
drug	X	X
quantity	X	X
unit	X	X

6. 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

APPENDIX A. ERROR CHECKING

APPENDIX A. ERROR CHECKING

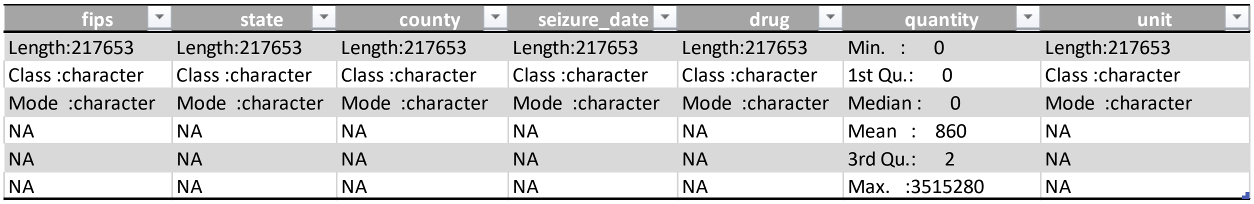
This Appendix A presents some of the statistical characteristics of two of the datasets available for the FY24Q3 delivery. The statistics presented were generated using the 'summary()' function in R.

Here's what the 'summary()' function provides:

- Length: Indicates the number of rows per column.
- Minimum (Min): The smallest value observed in the column.
- 1st Quartile (1st Qu): Represents the value at the 25th percentile, indicating the boundary below which 25% of the data falls.
- Median: The middle value of the column when arranged in ascending order.
- 3rd Quartile (3rd Qu): Represents the value at the 75th percentile, indicating the boundary below which 75% of the data falls.
- Maximum (Max): The largest observed value in the column.

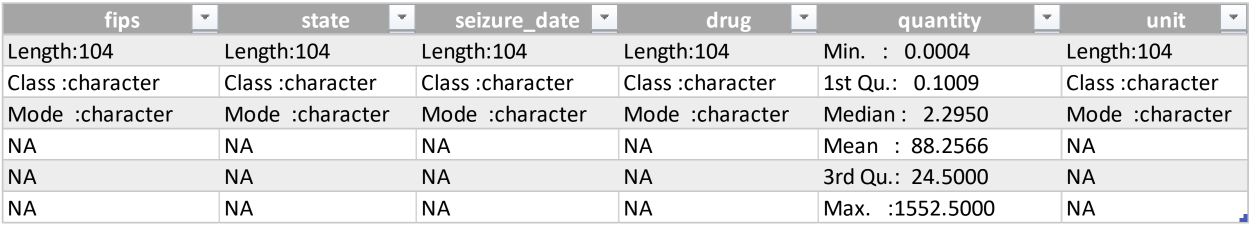
For columns of character type, mean, 1st quartile, median, 3rd quartile, and maximum values are not provided.

OMHSP_FY24Q3. hidta_county_2022_2023:



fips	state	county	seizure_date	drug	quantity	unit
Length:217653	Length:217653	Length:217653	Length:217653	Length:217653	Min. : 0	Length:217653
Class :character	Class :character	Class :character	Class :character	Class :character	1st Qu.: 0	Class :character
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Median : 0	Mode :character
NA	NA	NA	NA	NA	Mean : 860	NA
NA	NA	NA	NA	NA	3rd Qu.: 2	NA
NA	NA	NA	NA	NA	Max. :3515280	NA

OMHSP_FY24Q3. hidta_state_2022_2023:



fips	state	seizure_date	drug	quantity	unit
Length:104	Length:104	Length:104	Length:104	Min. : 0.0004	Length:104
Class :character	Class :character	Class :character	Class :character	1st Qu.: 0.1009	Class :character
Mode :character	Mode :character	Mode :character	Mode :character	Median : 2.2950	Mode :character
NA	NA	NA	NA	Mean : 88.2566	NA
NA	NA	NA	NA	3rd Qu.: 24.5000	NA
NA	NA	NA	NA	Max. :1552.5000	NA