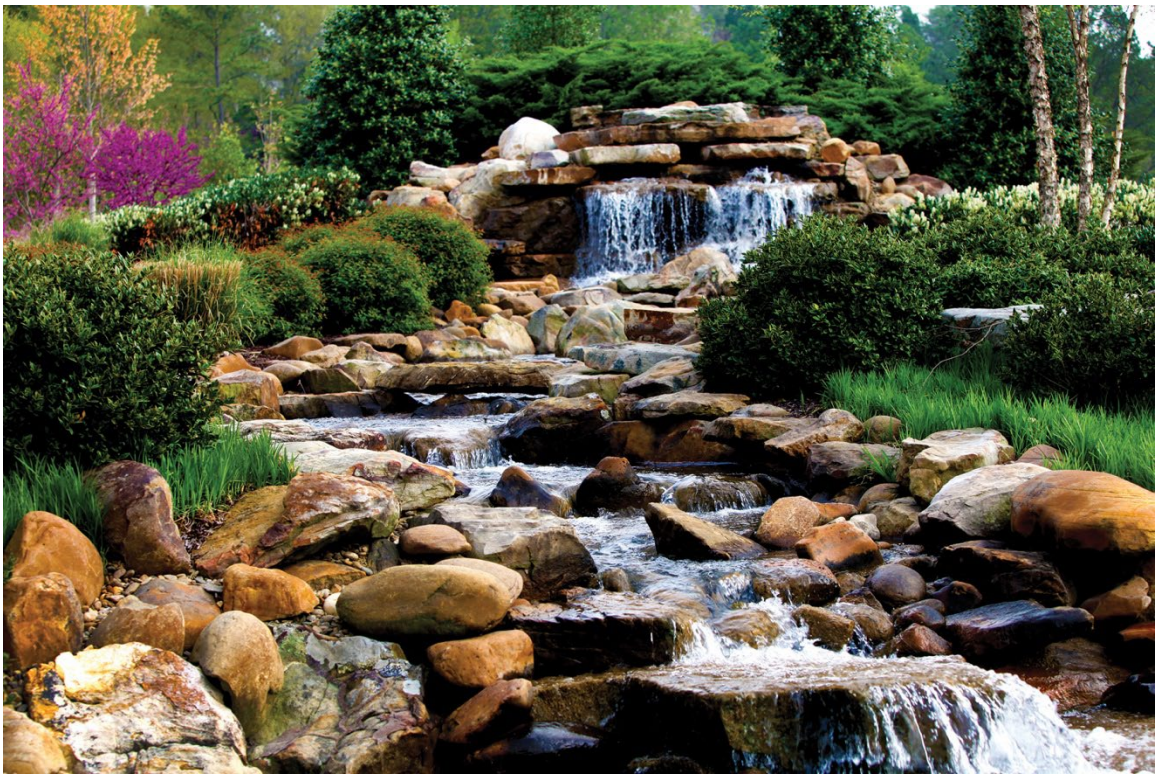
ORNL/SPR-2024/3299
PUB ID 210685

VA EDH Data Curation Documentation FY24-Q2



Hilda Klasky
Kevin Sparks
Alina Peluso
Jeremy Logan
Michael McGee
Kelly Callaway
Hope Cook
Dallas Sacca
Paul Reszczynski
Heidi Hanson
Rochelle Watson
Susana Martins
Jodie Trafton
Anuj Kapadia

March 2024

OAK RIDGE National Laboratory

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

ORNL/SPR-2024/3299
PUB ID 210685

Division Computational Sciences & Engineering

VA EDH DATA CURATION DOCUMENTATION FY24-Q2

Hilda Klasky
Kevin Sparks
Alina Peluso
Jeremy Logan
Michael McGee
Kelly Callaway
Hope Cook
Dallas Sacca
Paul Reszczynski
Heidi Hanson
Rochelle Watson
Susana Martins
Jodie Trafton
Anuj Kapadia

March 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

CONTENTS iii
1. Introduction 1
1.1 Background 1
2. Documentation Overview 1
2.1 Recommended Citation for FY24-Q2 Data Curation Documentation's Sponsor Report: 4
2.2 Previous Document Releases 4
3. Contents and Structure 6
3.1 Dataset Curation Documentation Standard Format 6
3.2 Dataset Conventions 7
3.3 Dataset Versioning 8
3.4 Metadata Table 8
3.5 Reports Table 9
3.6 FIPS as Geographic Identifiers and Primary Keys 9
3.7 Mapping Zip Codes to FIPS Codes for Counties: Our Methods 10
3.7.1 How to link cohorts to our datasets 11
3.8 Error Checking 12
4. High Intensity Drug Trafficking Areas (HIDTA) 15
4.1 Data Source 15
4.2 Description 15
4.3 Inclusion 15
4.4 Resources 15
4.5 Update Frequency 15
5. HUD USPS Zip Code Crosswalk Files 18
5.1 Data Source 18
5.2 Description 18
5.3 Inclusion 18
5.4 Resources 19
5.5 Update Frequency 19
6. Social Capital Index 2019 22
6.1 Data Source 22
6.2 Description 22
6.3 Inclusion 22
6.4 Resources 22
6.5 Update Frequency 22
7. References 25
Appendix A. ERROR CHECKING A-2

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, Second Quarter (FY24-Q2) dataset curation documentation for the Environmental Determinants of Health (EDH) project.

The main focus on FY24Q2 delivery has been addressing action items requested by our VA sponsors as described on the following Table 1.

Table 1. OMHSP VA EDH's project FY 24Q2 Error Log Table.

Issue No.	Issue Description	Correction Description
1.	Include pdf content in reports table	We have included the pdf report content in the reports table. And linked the reports table to the meta data table via three columns 'schema', 'ornl_res_pub_id' and 'OSTI_identifier'. These three columns can be joined to the meta data table independently or together. A view that joins both table is also deployed with both of the SEDH_meta and reports tables.
2.	FY22Q3_hidta_2018_21_county- Has Michigan as a county, yet FIPS has 2 digits, not 5. Outlier from county level granularity for this dataset	The HIDTA data was originally received during FY22Q3. Following discussions and analysis of the data, we believe that these are instances of drug seizures at locations. However, the most likely explanation for why these data appear to be at the state level rather than at the county level is that county-level information was not recorded in Michigan. The data team will not be providing aggregated data at the state level for the time being because the data provider no longer works at HIDTA. We are in the process of contacting the new representative to learn more about the data and possibly obtain a new data set. Following discussions with the VA, it was requested that this table (dataset) be delivered in FY24Q2 as two tables: one at the state level and another at the county level. The state data will only include Michigan data, not cumulative data from all other states.
3.	OMHSP_FY22Q3_housing_characteristics_2019_blockgroup - discrepancy in length of FIPS (min 11, max 12). Missing leading zeros?	Added the leading zero to the columns that are missing the 0 and have a FIPS length of 11 characters.
4.	Missing metadata: Spatial_resolution IS NULL OMHSP_FY24Q1_daymet_county_2017_2021 OMHSP_FY24Q1_va_vso_county_2010_2022 OMHSP_FY24Q1_va_vso_state_2010_2022 Can you ask for the metadata table with spatial_resolution be sent again?	As we are not able to reproduce this issue, after meeting with Susana and looking at our different work environments at ORNL and VA, we were able to identify that this issue only happens in the VA CDW that Susana is able to see; thus, we scheduled a meeting to talk to VA database administrators and find out why the datasets in the transmit database. We have a conference call with the VA database administrators on Wednesday, March 6th, at 3:30 EST. The FY24Q2 delivery will be performed in the last week of March.

5.	Inconsistent metadata- description does not align with spatial_resolution OMHSP_FY22Q1_adi_DC_2019_blockgroup OMHSP_FY22Q4_national_survey_on_drug_use_and_health OMHSP_FY22Q4_b3a_num_patients_24hour_inpatient_mental_health OMHSP_FY22Q4_b3c_num_inpatient_beds_for_mental_health OMHSP_FY22Q4_b4a_num_clients_receiving_24hr_res_mental_health OMHSP_FY22Q4_b4c_num_beds_designated_mental_health_treat OMHSP_FY22Q4_b5a_num_clients_less24hr_mental_health_treat OMHSP_FY22Q4_b7_num_mental_health_treat_admin_12month OMHSP_FY22Q4_b8_est_mental_health_admin_vet	OMHSP_FY22Q1_adi_DC_2019_blockgroup had spatial_resolution of 'state', corrected to blockgroup. In addition, the column descriptions have been improved.
6.	there are typos in the SEDH_meta_table - data_categories and data_source_description columns.	Resolved: The issue has been addressed by restructuring the data_categories column to ensure consistency in granularity. Subsequently, the data from this column has been redistributed into three distinct columns. Additionally, the original data_categories column has been removed from the dataset. The restructuring resulted in the creation of the following columns: • Dimension: Social, Environmental • Determinants: Economic Stability, Education, Health Care, Neighborhood, Food, and Social and Community Context. These determinants were sourced from Healthy People 2030 and Dang Y, Li F, Hu X, Keloth VK, Zhang M, Fu S, et al. Systematic Design and Evaluation of Social Determinants of Health Ontology (SDoHO) 2022. (doi: 10.48550/arXiv.2212.01941) • Source_Attribute: Authoritative, Derivative This adjustment enhances the clarity and organization of the dataset, ensuring accuracy and reliability for future analyses.
7.	Standardization of all datasets table names.	We will follow the convention of <dataset_name>_<spatial_resolution>_<year(s)>
8.	Text standardization: inconsistent naming conventions [spatial_resolution] values	Standardized [EDH_VA].[OMHSP].[SEDH_meta_table].[spatial_resolution] in [table_name] for the tables 'social_vulnerability_index_DC_2018_tract' and 'internet_access_2019_tract'. The old value was 'tract', and the new value is 'census tract' to maintain consistency with other datasets in

		[table_name] where [spatial_resolution] is set to 'census tract'.
9.	Standardization of fips column to all small letters on all tables	Note: FY24Q1 has name fips_code

In addition, the datasets included in this delivery and documentation are as follows:

1. **HUD USPS Zip Code Crosswalk Files, ZIP-to-tract for 2023 (*new*)**
2. **HUD USPS Zip Code Crosswalk Files, ZIP-to-county for 2023 (*new*)**
3. **Social Capital Index 2019, by county for 2019 (*updated*)**
4. **High Intensity Drug Trafficking Areas (HIDTA) state-level for 2018 - 2021 (*re-delivered*)**
5. **High Intensity Drug Trafficking Areas (HIDTA) county-level for 2018 - 2021 (*re-delivered*)**

2.1 RECOMMENDED CITATION FOR FY24-Q2 DATA CURATION DOCUMENTATION'S SPONSOR REPORT:

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). United States: N. p., 2024. ORNL/SPR-2024/3299 RES ID 210685.

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
  - 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

Please note that the URL for the FY24-Q2 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 an 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 TABLE

Starting from FY23Q1, the ORNL team provides an updated metadata table, known as SEDH_meta_table, located in the OMHSP schema. SEDH stands for the Social and Environmental Determinants of Health repository. This table contains the following columns:

- **schema**: Quarterly release schema names in the database (e.g., OMHSP_FY22Q4, OMHSP_FY23Q1, OMHSP_FY23Q2, OMHSP_FY23Q3, and so on).
- **table_name**: The table name as it appears in the MS SQL Server database.
- **table_name_description**: A description of the table name.
- **column_name**: Column names within each dataset as they appear in the MS SQL Server table.
- **column_name_description**: Descriptions of each column name.
- **availability_across_years**: The years for which data is available.
- **reference_report**: A reference to the ORNL report containing data curation documentation.
- **report_url**: URL link to the ORNL report.
- **column_type**: The column type in the MS SQL Server table.
- **column_length**: The column length in the MS SQL Server table.
- **total_rows**: The number of rows in each column in each dataset (starting in FY23Q2).
- **null_rows**: The number of null rows for each column in each dataset (starting in FY23Q2).
- **data_source**: The name of the source organization that provided the raw data (starting in FY23Q4).
- **data_source_description**: Description of the source organization (starting in FY23Q4).
- **data_source_url**: URL of the source organization (starting in FY23Q4).
- **spatial_resolution**: Spatial resolution or geography (e.g., state, county, block group, census tract, and zip code.) (starting in FY23Q4).
- **data_categories**: General data categories, such as social, economic, educational, etc. (started in FY23Q4, but removed in FY24Q2, as it was replaced with the 'determinant', 'source_attribute', and 'dimension' columns, to provide a more precise category granularity, and to follow SDoH ontologies.

- **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.
- **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. See reference [11].
- **dimension**: two option values: social and environmental
- **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 osti 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.

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 reports table which will include 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:

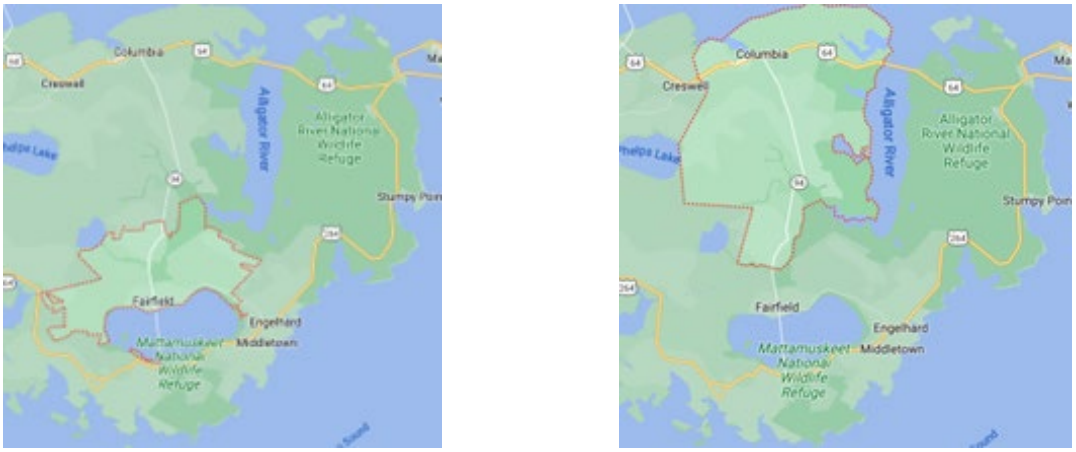
1. The National Mental Health Services Survey (table: national_mental_health_services_survey), delivered in FY22Q4.
2. The Veterans Service Organizations (VSO) 2010-2022, by zip code, delivered in FY24Q1.
3. HUD USPS Zip Code Crosswalk Files, ZIP-to-tract for 2023, delivered in FY24Q2.

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:

Table 2. *Example*: - *Left*: Zip Code 27826 - *Right*: County FIPS 37177



3.7.1 How to link cohorts to our datasets

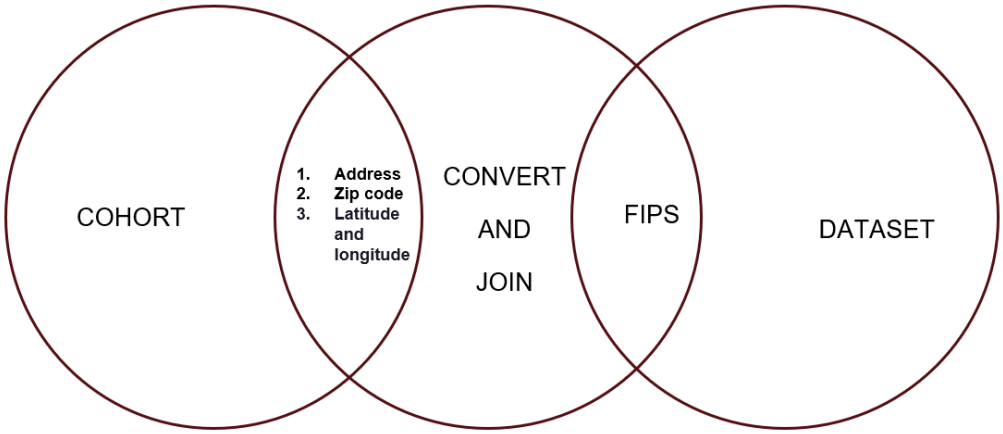


Figure 1. Joining geoids to the VA EDH OMHSP datasets.

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 [11]:

"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." [11]

The error-checking results for FY24Q2 follows:

			Development		QA-Intra		QA-KDI (VIEWS)		Production (Transmit)		
Dataset Name	Rows	Columns	Passes	Fails	Passes2	Fails3	Passes4	Fails5	Passes6	Fails7	Error ratio
OMHSP_FY24Q2.hidta_county_2018_2021	348280	14	5	0	5	0	5	0	5	0	0
OMHSP_FY24Q2.hidta_state_2018_2021	304	7	5	0	5	0	5	0	5	0	0
OMHSP_FY24Q2.hud_zip_county_2023	54503	8	5	0	5	0	5	0	5	0	0
OMHSP_FY24Q2.hud_zip_tract_2023	377692	8	5	0	5	0	5	0	5	0	0
OMHSP_FY24Q2.sci_county_2019	3108	56	3	2	4	1	5	0	5	0	0.18
OMHSP.SEDH_meta_table	852	19	4	1	4	1	5	0	4	1	0.18
OMHSP.SEDH_reports	10	7	5	0	4	1	5	0	4	1	0.11
VIEWS.SEDH_meta_data_view	852	23	5	0	5	0	5	0	5	0	0

In addition, we performed visual error checks on the housing characteristics dataset (table: OMHSP_FY22Q3_housing_characteristics_2019_blockgroup), which had a difference in FIPS length (min 11, max 12) and missing leading zeros. Refer to issue number 3 in the error log table. The image below illustrates the rectified leading zeros in Alabama as an example of an error check efforts.

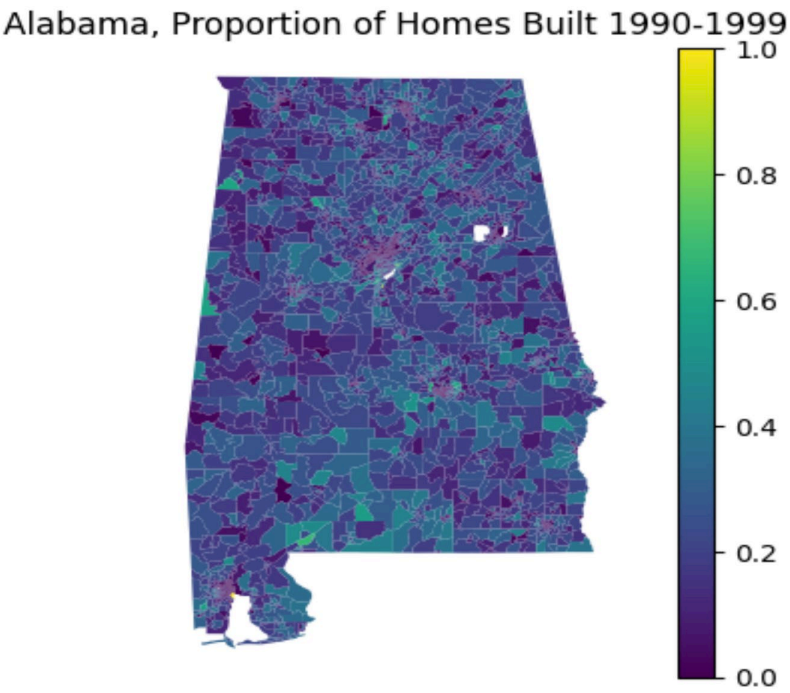


Figure 2. Error checking on corrected leading zero for housing characteristics dataset.

Appendix A presents descriptive statistics of error-checking results.

4. HIGH INTENSITY DRUG TRAFFICKING AREAS (HIDTA)

4.1 DATA SOURCE

Washington/Baltimore High Intensity Drug Trafficking Areas Program.

4.2 DESCRIPTION

The High Intensity Drug Trafficking Areas (HIDTA) program, established by Congress with the Anti-Drug Abuse Act of 1988, provides support to federal, state, local, and tribal law enforcement agencies operating in locations identified as significant drug-trafficking areas in the United States.

The HIDTA data was originally received as provided to our VA sponsors in FY22Q3. After discussions and studying the data, we understand that it consists of instances of drug seizures at various locations. After analyzing the data we realized that the county-level information was not recorded for Michigan. We are currently in the process of reaching out to the new HIDTA representative to gather more information about the data and obtain a new dataset if possible. Following discussions with our VA sponsors, it was requested to re-deliver this dataset in FY24Q2, dividing it into two tables: one at the state level and another at the county level.

The state-level data will only include Michigan data and not be cumulative for all other states.
The county-level data will only include the rest of data and exclude the state-level data of Michigan.

4.3 INCLUSION

The data was obtained from the High Intensity Drug Trafficking Areas Program's activities from 2018 to 2021.
Both the state-level and the county-level data sets have the columns listed below.

4.4 RESOURCES

For more information about HIDTA:
Main: https://www.dea.gov/operations/hidtas

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

variable name	variable label
fips	Federal Information Processing Standards (FIPS), county level fips codes.
state	US state name.
county	The county 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	2018	2019	2020	2021
fips	X	X	X	X
state	X	X	X	X
county	X	X	X	X
seizure_date	X	X	X	X
drug	X	X	X	X
quantity	X	X	X	X
unit	X	X	X	X

5. HUD USPS ZIP CODE CROSSWALK FILES

5.1 DATA SOURCE

US Housing and Urban (HUD) Development's Office of Policy Development and Research.

5.2 DESCRIPTION

When attempting to relate USPS ZIP codes to Census Bureau geographies, social science researchers and practitioners frequently face challenges. Despite the availability of useful data at the ZIP code level, combining it with demographic data tabulated at various Census geography levels remains challenging, limiting exploration opportunities. While there are acceptable methods for merging ZIP codes and Census geography, they are limited. To address this issue, PD&R made available the HUD-USPS Crosswalk Files. These distinct files are derived from data in the quarterly USPS Vacancy Data, which is obtained directly from the USPS. They are updated quarterly, respond quickly to changes in ZIP code configurations, and include both business and residential address locations.

To facilitate the mapping process between cohort addresses and the social and environmental determinants of health datasets provided by this project, we use crosswalk tables from the US Housing and Urban Development's Office of Policy Development and Research. Instead of spatial data types such as Geometry and Geography, which allow for more precise conversion of latitude and longitude coordinates to FIPS codes, we recommend using the crosswalk tables from the US Housing and Urban Development's Office of Policy Development and Research. Crosswalk Tables Link.

In fact, we used these files to generate the Veterans Service Organizations datasets included in the FY24Q1 release. These crosswalk tables serve as a reference for associating ZIP codes with their corresponding FIPS codes, making it easier to integrate our datasets.

However, it is important to recognize that this process has limitations. In approximately 45% of cases, ZIP codes cannot be perfectly mapped to county-level FIPS codes. This flaw highlights the difficulties inherent in geographic data integration and emphasizes the importance of careful thought and validation when performing these conversions.

According to the HUD website, the relationship between the two types of crosswalk files is not completely inverse. That is, the ZIP to Tract crosswalk file cannot be used to map data from census tracts to ZIP codes. In that case, you must use the Tract to ZIP crosswalk file. Despite its flaws, using crosswalk tables is still a useful method for connecting cohort data to our project's datasets, providing a foundational step in the analysis and interpretation of social and environmental determinants of health.

When utilizing this dataset, please include the following citation: Din, Alexander and Wilson, Ron, 2020. "Crosswalking ZIP Codes to Census Geographies: Geoprocessing the U.S. Department of Housing & Urban Development's ZIP Code Crosswalk Files," Cityscape: A Journal of Policy Development and Research, Volume 22, Number 1, Link

5.3 INCLUSION

We provide the ZIP-to-tract and ZIP-to-county (FIPS) crosswalk tables for the 4th quarter of 2023. Note that ZIP code boundaries frequently do not align with administrative/political boundaries. For more information, the HUD recommends seeing USPS City Versus Census Geography located in their website.

5.4 RESOURCES

- Link to source tables: HUD USPS Crosswalk Tables

- Wilson, Ron and Din, Alexander, 2018. "Understanding and Enhancing the U.S. Department of Housing and Urban Development's ZIP Code Crosswalk Files," Cityscape: A Journal of Policy Development and Research, Volume 20 Number 2, 277 – 294. Link

- Din, Alexander and Wilson, Ron, 2020. "Crosswalking ZIP Codes to Census Geographies: Geoprocessing the U.S. Department of Housing & Urban Development's ZIP Code Crosswalk Files," Cityscape: A Journal of Policy Development and Research, Volume 22, Number 1, Link

5.5 UPDATE FREQUENCY

This dataset will be updated and disseminated annually, or as per the sponsor's request. Minimal quarterly updates may be necessary to rectify minor data inaccuracies.

Table 5. HUD USPS Zip Code Crosswalk Files (HUD)

variable name	variable label
ZIP	5 digit USPS ZIP code
fips	originally found as 'COUNTY' is the 5 digit unique 2000 or 2010 Census county GEOID consisting of state FIPS + county FIPS.
TRACT	11 digit unique 2000 or 2010 Census tract GEOID consisting of state FIPS + county FIPS + tract code. The decimal is implied and leading and trailing zeros have been preserved.
USPS_ZIP_PREF_CITY	USPS preferred city name. Note, ZIP code preferred names frequently do not align with administrative/political names.
USPS_ZIP_PREF_STATE	USPS preferred state address state.
RES_RATIO	The ratio of residential addresses in the ZIP – Tract, County, or CBSA part to the total number of residential addresses in the entire ZIP.
BUS_RATIO	The ratio of business addresses in the ZIP – Tract, County, or CBSA part to the total number of business addresses in the entire ZIP.
OTH_RATIO	The ratio of other addresses in the ZIP – Tract, County, or CBSA part to the total number of other addresses in the entire ZIP.
TOT_RATIO	The total ratio of all addresses in the ZIP – Tract, County, or CBSA part to the total number of all types of addresses in the entire ZIP.

Table 6. Variable availability across years, (HUD)

variable name	2023
ZIP	X
fips	X
TRACT	X
USPS_ZIP_PREF_CITY	X
USPS_ZIP_PREF_STATE	X
RES_RATIO	X
BUS_RATIO	X
OTH_RATIO	X
TOT_RATIO	X

6. SOCIAL CAPITAL INDEX 2019

6.1 DATA SOURCE

United States Department of Veterans Affairs

6.2 DESCRIPTION

This dataset presents the "Social Capital Index 2019," as delineated in the study conducted by Peluso, Alina, et al., entitled "Spatial Analysis of Social Capital and Community Heterogeneity at the United States County Level," published in Applied Geography (Volume 162, 2024, Article 103168). Study Link

Initially delivered to our sponsors at the VA during the FY22Q1, this Social Capital Index surpasses its FY22Q2 counterpart in comprehensiveness. The aforementioned reference underscores its integration of Putnam-like or collective definitions of social capital, as proposed in the literature. The authors leveraged this updated measure to scrutinize the correlation between social capital and community heterogeneity. A thorough examination of this correlation enhances our comprehension of the role of community diversity and elucidates disparities in our measurement of homogeneous and heterogeneous environments. Furthermore, both the study and dataset underwent peer review.

When utilizing this dataset, please include the following citation:

Peluso, A., Tuccillo, J., Sparks, K., Kapadia, A., & Hanson, H. A. (2024). Spatial Analysis of Social Capital and Community Heterogeneity at the United States County Level. Applied Geography, 162, 103168. Study Link

6.3 INCLUSION

Geographic Unit: County level (county FIPS codes) for the contiguous US states. This dataset comprises a total of 55 columns and 3808 rows.

Please note that the tables below provide descriptions of the basic data columns due to space limitations. For each basic column, the dataset includes numerical values as well as per capita values (indicated by columns suffixed with '*_percap'), 'f_'* prefixed columns with Boolean values, meaning whether the value has been imputed (true) or not (false), and columns suffixed with '_imp' representing the same variables with imputed values for any values less than 3. For additional details, kindly refer to the aforementioned research manuscript.

6.4 RESOURCES

By using free, publicly available, and reliable data, we generated a 2019 US county-level social capital index to be employed in contemporary studies. Our 2019 US county-level social capital index is based on the definition of social capital generated by Rupasingha et al. in 2006. The original study was conducted by Rupasingha et al. in 2006 is found here. Original Study Link

6.5 UPDATE FREQUENCY

This dataset will be updated as new data will allow, or as per the sponsor's request.

Table 7. Social Capital Index 2019 (SCI)

variable name	variable label
fips	Federal Information Processing Standards (FIPS), county FIPS code,
year	The year of the data, 2019.
SCI	2019 Social Capital Index (normalized first principal component from associations, vote, response, and nonprofit organizations).
bowling	No. of bowling centers.
civic	No. of civic and social associations establishments.
fitness	No. of fitness and recreational sports centers.
golf	No. of golf courses and country clubs.
religion	No. of religious organizations.
sport	No. of sports teams and clubs.
political	No. of political organizations.
professional	No. of professional organizations.
business	No. of business associations.
labor	No. of labor organizations.
associations	Average of all 10 above variables divided by population × 10,000.
response	Census response rate.
population	Population estimate.
vote	Voter turnout rate.
nonprofits	No. of nonprofit organizations. Note that this column was originally named ‘nccs’ for Nonprofits or National Center for Charitable Statistics (NCCS) data.
county_name	County name.
state_abbr	State abbreviation.
state_name	State name.
long_name	County and state.
state	State code.
county	County code.
region_name	Region name.
division_name	Division name.

Table 8. Variable availability across years, (SCI)

variable name	2019
fips	X
year	X
bowling	X
civic	X
fitness	X
golf	X
religion	X
sport	X
political	X
professional	X
business	X
labor	X
associations	X
response	X
population	X
vote	X
nonprofits	X
county_name	X
state_abbr	X
state_name	X
long_name	X
state	X
county	X
region_name	X
division_name	X

7. 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 191790. 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.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 outlines the statistical characteristics of the datasets available for the FY24Q2 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_FY24Q2. hidta_county_2018_2021:

fips	state	county	seizure_date	drug	quantity	unit
Length:304	Length:304	Length:304	Length:304	Length:304	Min. : 0.000	Length:304
Class :character	Class :character	Class :character	Class :character	Class :character	1st Qu.: 0.062	Class :character
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Median : 3.000	Mode :character
NA	NA	NA	NA	NA	Mean : 47.075	NA
NA	NA	NA	NA	NA	3rd Qu.: 15.340	NA
NA	NA	NA	NA	NA	Max. :4001.000	NA

OMHSP_FY24Q2. hidta_state_2018_2021:

fips	state	county	seizure_date	drug	quantity	unit
Length:348280	Length:348280	Length:348280	Length:348280	Length:348280	Min. : 0	Length:348280
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 : 156	NA
NA	NA	NA	NA	NA	3rd Qu.: 1	NA
NA	NA	NA	NA	NA	Max. :10000000	NA

OMHSP_FY24Q2. hud_zip_county_2023:

ZIP	COUNTY	USPS_ZIP_PREF_CITY	USPS_ZIP_PREF_STATE	RES_RATIO	BUS_RATIO	OTH_RATIO	TOT_RATIO
Length:54503	Length:54503	Length:54503	Length:54503	Min. :0.00000	Min. :0.00000	Min. :0.0000	Min. :0.0000272
Class :character	Class :character	Class :character	Class :character	1st Qu.:0.07214	1st Qu.:0.03488	1st Qu.:0.0000	1st Qu.:0.2897553
Mode :character	Mode :character	Mode :character	Mode :character	Median :0.98945	Median :1.00000	Median :1.0000	Median :1.0000000
NA	NA	NA	NA	Mean :0.65242	Mean :0.66167	Mean :0.6583	Mean :0.7245840
NA	NA	NA	NA	3rd Qu.:1.00000	3rd Qu.:1.00000	3rd Qu.:1.0000	3rd Qu.:1.0000000
NA	NA	NA	NA	Max. :1.00000	Max. :1.00000	Max. :1.0000	Max. :1.0000000

OMHSP_FY24Q2. hud_zip_tract_2023:

ZIP	TRACT	USPS_ZIP_PREF_CITY	USPS_ZIP_PREF_STATE	RES_RATIO	BUS_RATIO	OTH_RATIO	TOT_RATIO
Length:188846	Length:188846	Length:188846	Length:188846	Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.0000172
Class :character	Class :character	Class :character	Class :character	1st Qu.:0.01307	1st Qu.:0.00418	1st Qu.:0.00000	1st Qu.:0.0182161
Mode :character	Mode :character	Mode :character	Mode :character	Median :0.07556	Median :0.04381	Median :0.03361	Median :0.0806299
NA	NA	NA	NA	Mean :0.18803	Mean :0.19074	Mean :0.18656	Mean :0.2084662
NA	NA	NA	NA	3rd Qu.:0.18910	3rd Qu.:0.19409	3rd Qu.:0.18182	3rd Qu.:0.2098630
NA	NA	NA	NA	Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.0000000

OMHSP_FY24Q2.sci_county_2019:

fips	Min. : 1001	1st Qu.:19044	Median :29212	Mean :30672	3rd Qu.:46008	Max. :56045
year	Min. :2019	1st Qu.:2019	Median :2019	Mean :2019	3rd Qu.:2019	Max. :2019
SCI	Min. :-2.4681	1st Qu.:-0.4925	Median : 0.0000	Mean : 0.1374	3rd Qu.: 0.6071	Max. : 9.6961
bowling	Min. : 0.0000	1st Qu.: 0.0000	Median : 0.0000	Mean : 0.6844	3rd Qu.: 0.0000	Max. :44.0000
civic	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 7.837	3rd Qu.: 7.000	Max. :550.000
fitness	Min. : 0.0	1st Qu.: 0.0	Median : 0.0	Mean : 12.2	3rd Qu.: 7.0	Max. :1165.0
golf	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 2.778	3rd Qu.: 3.250	Max. :135.000
religion	Min. : 0.00	1st Qu.: 11.00	Median : 24.00	Mean : 59.18	3rd Qu.: 55.00	Max. :3267.00
sport	Min. : 0.0000	1st Qu.: 0.0000	Median : 0.0000	Mean : 0.1953	3rd Qu.: 0.0000	Max. :36.0000
political	Min. : 0.0000	1st Qu.: 0.0000	Median : 0.0000	Mean : 0.6248	3rd Qu.: 0.0000	Max. :184.0000
professional	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 1.958	3rd Qu.: 0.000	Max. :308.000
business	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 4.324	3rd Qu.: 4.000	Max. :503.000
labor	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 3.941	3rd Qu.: 3.000	Max. :290.000
associations	Min. :0.000	1st Qu.:1.053	Median :1.378	Mean :1.488	3rd Qu.:1.781	Max. :6.202
response	Min. :0.1330	1st Qu.:0.5228	Median :0.6155	Mean :0.5981	3rd Qu.:0.6900	Max. :0.8490
population	Min. : 98	1st Qu.: 11177	Median : 25946	Mean : 103774	3rd Qu.: 67892	Max. :10081570
vote	Min. :0.1498	1st Qu.:0.4428	Median :0.4968	Mean :0.4975	3rd Qu.:0.5522	Max. :1.1451
nccs	Min. : 0.5131	1st Qu.: 4.7530	Median : 7.0274	Mean : 8.5528	3rd Qu.: 10.4880	Max. :116.6861
county_name	Length:3108	Class :character	Mode :character	NA	NA	NA
state_abbr	Length:3108	Class :character	Mode :character	NA	NA	NA
state_name	Length:3108	Class :character	Mode :character	NA	NA	NA
long_name	Length:3108	Class :character	Mode :character	NA	NA	NA
state	Min. : 1.00	1st Qu.:19.00	Median :29.00	Mean :30.57	3rd Qu.:46.00	Max. :56.00
county	Min. : 1.0	1st Qu.: 35.0	Median : 79.0	Mean :103.3	3rd Qu.:133.0	Max. :840.0
region_name	Length:3108	Class :character	Mode :character	NA	NA	NA
division_name	Length:3108	Class :character	Mode :character	NA	NA	NA
bowling_percap	Min. :0.000e+00	1st Qu.:0.000e+00	Median :0.000e+00	Mean :2.792e-06	3rd Qu.:0.000e+00	Max. :1.479e-04
civic_percap	Min. :0.0000000	1st Qu.:0.0000000	Median :0.0000000	Mean :0.0000744	3rd Qu.:0.0001131	Max. :0.0013810
fitness_percap	Min. :0.000e+00	1st Qu.:0.000e+00	Median :0.000e+00	Mean :5.043e-05	3rd Qu.:9.823e-05	Max. :6.370e-04
golf_percap	Min. :0.000e+00	1st Qu.:0.000e+00	Median :0.000e+00	Mean :2.502e-05	3rd Qu.:2.950e-05	Max. :5.885e-04
religion_percap	Min. :0.0000000	1st Qu.:0.0005970	Median :0.0008624	Mean :0.0009417	3rd Qu.:0.0011800	Max. :0.0062016
sport_percap	Min. :0.000e+00	1st Qu.:0.000e+00	Median :0.000e+00	Mean :2.536e-07	3rd Qu.:0.000e+00	Max. :3.960e-05
political_percap	Min. :0.000e+00	1st Qu.:0.000e+00	Median :0.000e+00	Mean :1.146e-06	3rd Qu.:0.000e+00	Max. :2.688e-04
professional_percap	Min. :0.000e+00	1st Qu.:0.000e+00	Median :0.000e+00	Mean :4.763e-06	3rd Qu.:0.000e+00	Max. :6.154e-04
business_percap	Min. :0.000e+00	1st Qu.:0.000e+00	Median :0.000e+00	Mean :3.088e-05	3rd Qu.:4.140e-05	Max. :9.655e-04
labor_percap	Min. :0.000e+00	1st Qu.:0.000e+00	Median :0.000e+00	Mean :1.917e-05	3rd Qu.:6.697e-06	Max. :5.359e-04
bowling_imp	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 1.13	3rd Qu.: 1.00	Max. :44.00
f_bowling_imp	Mode :logical	FALSE:371	TRUE :2737	NA	NA	NA
civic_imp	Min. : 0.00	1st Qu.: 1.00	Median : 2.00	Mean : 8.55	3rd Qu.: 7.00	Max. :550.00
f_civic_imp	Mode :logical	FALSE:1510	TRUE :1598	NA	NA	NA
fitness_imp	Min. : 0.00	1st Qu.: 1.00	Median : 2.00	Mean : 13.01	3rd Qu.: 7.00	Max. :1165.00
f_fitness_imp	Mode :logical	FALSE:1275	TRUE :1833	NA	NA	NA
golf_imp	Min. : 0.000	1st Qu.: 1.000	Median : 2.000	Mean : 3.467	3rd Qu.: 3.250	Max. :135.000
f_golf_imp	Mode :logical	FALSE:1032	TRUE :2076	NA	NA	NA
religion_imp	Min. : 0.00	1st Qu.: 11.00	Median : 24.00	Mean : 59.24	3rd Qu.: 55.00	Max. :3267.00
f_religion_imp	Mode :logical	FALSE:2974	TRUE :134	NA	NA	NA
sport_imp	Min. : 0.0000	1st Qu.: 0.0000	Median : 0.0000	Mean : 0.4556	3rd Qu.: 0.0000	Max. :36.0000
f_sport_imp	Mode :logical	FALSE:107	TRUE :3001	NA	NA	NA
political_imp	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 1.185	3rd Qu.: 1.000	Max. :184.000
f_political_imp	Mode :logical	FALSE:185	TRUE :2923	NA	NA	NA
professional_imp	Min. : 0.000	1st Qu.: 0.000	Median : 1.000	Mean : 2.778	3rd Qu.: 2.000	Max. :308.000
f_professional_imp	Mode :logical	FALSE:388	TRUE :2720	NA	NA	NA
business_imp	Min. : 0.000	1st Qu.: 1.000	Median : 2.000	Mean : 5.135	3rd Qu.: 4.000	Max. :503.000
f_business_imp	Mode :logical	FALSE:1138	TRUE :1970	NA	NA	NA
labor_imp	Min. : 0.000	1st Qu.: 1.000	Median : 2.000	Mean : 4.838	3rd Qu.: 3.000	Max. :290.000
f_labor_imp	Mode :logical	FALSE:782	TRUE :2326	NA	NA	NA