ORNL/TM-2023/3027

# Dataset Repository for Investigating Suicide Risk Using Social and Environmental Determinants of Health

Hilda B. Klasky
Heidi A. Hanson
Kevin Sparks
Matthew Whitehead
Blair Christian
Jodie Trafton
Anuj J. Kapadia

**August 2023**

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/TM-2023/3027**

Computational Sciences & Engineering Division

# DATASET REPOSITORY FOR INVESTIGATING SUICIDE RISK USING SOCIAL AND ENVIRONMENTAL DETERMINANTS OF HEALTH

Hilda B. Klasky
Heidi A. Hanson
Kevin Sparks
Matthew Whitehead
Blair Christian
Jodie Trafton
Anuj J. Kapadia

August 2023

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

# TABLE OF FIGURES

# TABLE OF TABLES

# ABSTRACT

Suicide is frequently modeled as a function of genetics and environment, where the latter refers to factors other than direct biological consequences, such as air quality, financial level, social connectivity, transportation and food access, and homelessness status. According to the World Health Organization, clean air, a stable climate, adequate water, sanitation and hygiene, safe chemical use, radiation protection, healthy and safe workplaces, sound agricultural practices, health-supportive cities and built environments, and a preserved natural environment are all prerequisites for good health. Understanding the relationships between these determinants and mental health outcomes requires standardized data that can be included in healthcare programs and health outcome models. There is a wealth of publicly available data on social and environmental factors provided by various US organizations that can benefit the design of health care systems and public health interventions, as well as improve our comprehension of factors that impact health. Such information would not only help improve the understanding of individual and community risk but also identify new risk factors that have not previously been therapeutically targeted, especially in terms of their impact on mental health. However, curating and standardizing such datasets is challenging because they are often recorded at numerous geographical and temporal resolutions and with varying spatial and temporal granularities. To address this challenge, we launched an endeavor in conjunction with the Veterans Health Administration to collect publicly available socioeconomic and environmental determinants of health statistics in the US. In this manuscript, we describe a social and environmental determinants of health (SEDH) datasets repository, data curation documentation, and a pipeline framework for data generation; This effort started in 2020, when we began constructing a scalable pipeline to automate the download, extraction, preparation, analysis, and production of datasets. These datasets have been made available to the VHA and may be shared upon agreement with collaborating organizations.

# 1. INTRODUCTION

The US Department of Veterans Affairs (VA) is the world's most comprehensive system of support for veterans [1], providing support in three primary areas: health, benefits, and burial. While highly effective already, rapidly evolving health challenges and clinical innovation force the VA to constantly review its practices and their impacts on America's Veterans; for example: rapidly expanding technology in numerous disciplines; social and environmental uncertainty; shifting social and demographic trends; and tight economic limits [2]. As a result, the Veterans Health Administration (VHA) has adopted the following public health strategy: combining universal, targeted, and suggested actions to minimize global risk; attending to high-risk groups; and treating those with established clinical needs [3].

Suicide is commonly modeled as a function of genetics and the environment, where the environment refers to factors other than direct biological effects, such as air quality, socioeconomic status, social connectedness, transportation and food access, and homelessness status. According to the World Health Organization (WHO), clean air, stable climate, adequate water, sanitation and hygiene, safe use of chemicals, radiation protection, healthy and safe workplaces, sound agricultural practices, health-supportive cities and built environments, and a preserved natural environment, are all prerequisites for good health [4]. Understanding the links between these factors and mental health outcomes necessitates collected, standardized data that can be included into the VA's predictive model-based targeted prevention programs, such as Recovery Engagement and Coordination for Health-Veterans Enhanced Treatment (REACH VET) and Stratification Tool for Opioid Risk Mitigation, and/or be used in strategic planning models and efforts to optimize health care access and delivery [5].

There is a variety of publicly available data on social and environmental health variables that may provide valuable context for the patient's predicted health and community-related challenges and concerns [6-23]. This information may aid in our knowledge of risk and the identification of novel risk variables that have not previously been therapeutically targeted. When used in risk surveillance and clinical operations, such data may aid in bringing new risks to clinical attention, targeting therapeutic concerns to specific times of elevated risk, improving identification of patients most likely to commit suicide or overdose, informing public health and community outreach efforts to address suicide risk factors, and matching availability of specific treatments to health care locations based on local patient needs (Figure 1).

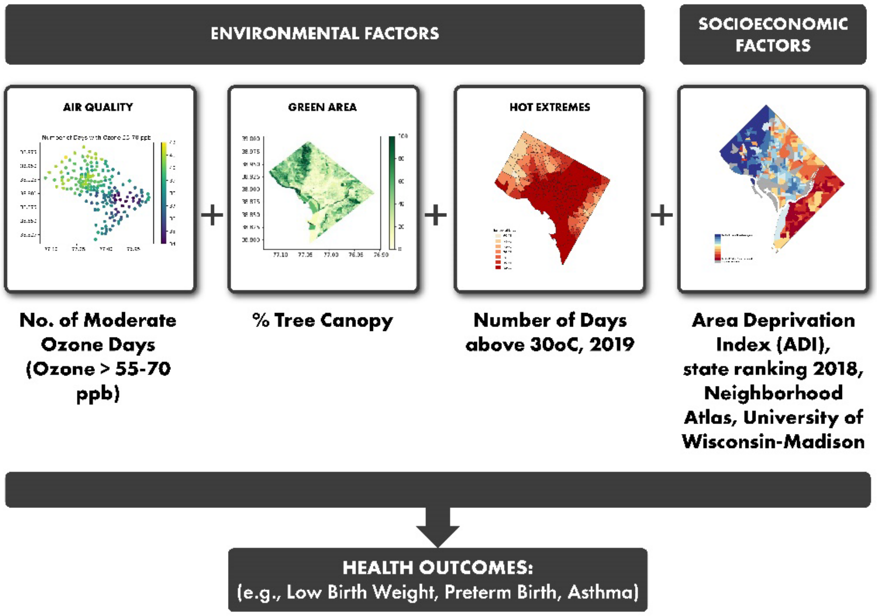
ENVIRONMENTAL FACTORS
SOCIOECONOMIC FACTORS
AIR QUALITY
GREEN AREA
HOT EXTREMES
+
+
+
No. of Moderate Ozone Days (Ozone > 55-70 ppb)
% Tree Canopy
Number of Days above 30oC, 2019
Area Deprivation Index (ADI), state ranking 2018, Neighborhood Atlas, University of Wisconsin-Madison
HEALTH OUTCOMES:
(e.g., Low Birth Weight, Preterm Birth, Asthma)

**Figure 1. Environmental and socioeconomic data integration.**

Because these datasets are typically measured at multiple spatial and temporal resolutions and have varied spatial and temporal granularities, curating, and standardizing them is a difficult task. For example, most US Census [6] data products use census blocks, block groups, or counties,  whereas the US Environmental Protection Agency's air pollutants and weather data are accessible on 1 km grids, and some economic data are only available at the zip code level [19]. In the context of our study, standardized means that all the datasets are available at the same geographic scale (e.g., US Census Tract, County, or km grids), and curated means that the process is repeatable, has data provenance, and employs standardized procedures for converting variables.

Most modeling frameworks [24-26] that utilize such data fall into two categories: a) "Traditional" approaches in which features are handcrafted; and b) Machine learning approaches in which measures of exposure are estimated. The key significant issues with existing datasets can be summarized as follows: a) Large amounts of "missing data"; b) Non-random missingness, for example, certain data are available at the state or county level only, not at the block group level; c) Mobility data (work/home) are not available; d) Unharmonized data across different datasets, for example, data are reported at different spatiotemporal scales; e) Presence of correlation and confounding in the datasets, which may lead to collinearity during modeling.

Our overall vision is to automate the construction of data sets that can be used by the broad research community to estimate the association between environmental exposures and health outcomes. We aim to do so by a) Generating community data products (datasets), b) Designing opensource toolkits in R and Python to analyze the datasets, c) Putting data into 'model ready' format usable by a broad community of government operations staff and researchers, and d) Developing and implementing a methodology for multi-modal data that includes statistical and deep learning, expertise with text, image, genomic, and spatial data, and methods for more complex outcomes such as trajectories.

With this vision in mind, we developed an approach that includes: 1) a reproducible and reusable data pipeline for standardizing data collection; 2) change support (e.g., traversing different spatial/temporal granularities); 3) small area estimation, Gaussian processes, and raster functions; 4) dasymetric modeling; 5) packaging these in software containers for easy re-use, scalability, and reproducibility; and 6) standardized meta-data collection. With each release of the datasets, we have incorporated robust documentation that describes the datasets and the methods used for data curation. This customized strategy has the potential to overcome the data and technique limitations that exist in current data. In contrast to traditional feature development, we have designed a data pipeline that automates the creation of covariates (e.g., change of support) that may deliver a multi-modal experience using text, photos, genomics, geographical, and other types of data. In addition to datasets, we provide modeling assistance using classical statistical modeling skills (Bayesian/Frequentist), machine learning analysis, multimodal data integration (i.e., text, picture, genomic variables), and trajectory data approaches (statistical (Functional Data Analysis, Hidden Markov Model, etc.)). In collaboration with the VA, we initiated an effort to gather publicly available US datasets on social and environmental determinants of health (SEDH).

The SEDH repository includes the following main components: a) novel datasets associated with select health outcomes; b) documentation of the data curation process; c) methodology for converting spatiotemporal data from one spatial reference to another (e.g., from 1 km grid to US Census Tracts); and d) health outcomes modeling capabilities (ongoing) with funding from the VA Office of Mental Health and Suicide Prevention (OMHSP) [3]. The datasets are an improvement on the Social Determinants of Health (SDoH) variables developed by the Agency for Healthcare Research and Quality (AHRQ) [27], address important gaps, provide a better geographical resolution (Census Tract), and include environmental covariates.

This manuscript is organized as follows: Section 1 presents the introduction; Section 2 describes the curation and standardization of community datasets, including the pipeline software framework and the change of support; Section 3 describes the key community datasets; Section 4 presents documentation; Section 5 lists the limitations and plans for future work; Section 6 summarizes the conclusions. And finally, the acknowledgements and references are presented at the end.

## 2. CURATION AND STANDARDIZATION OF COMMUNITY DATASETS

We have developed a flexible framework for the reproducible processing of SEDH that prioritizes data provenance and transparency; it contains a "toolbox" that allows for the selection of variables across time at standardized spatial and temporal resolutions. This framework consists of a scalable pipeline to automate the creation of key SEDH datasets and allows for the harmonization of SEDH data from a vast number of sources. This approach greatly improves the ability to incorporate relevant measures into downstream research and prediction tasks.

There are two key parts to this effort: 1) building the software architecture of the computational pipeline for community datasets; and 2) construction of the change of support algorithms. Both parts are described below.

### 2.1 COMPUTATIONAL PIPELINE FOR COMMUNITY DATASETS

We have developed computational tools to harmonize multimodal social and environmental datasets so that disparate data can be used in the modeling process with minimal effort. As there is a large amount of spatial, temporal, and input variability in the community social and environmental data (e.g., 1 km by 1km grids of daily data, yearly zip code data, quarterly county level data, raster files, csv files), we must standardize the measurements to the same temporal and spatial references (yearly, census tract level).

The harmonization of community factor datasets is required to construct a spatial data pipeline and generate foundational data for use in statistical models. This effort included the development of automated workflows to provide end-to-end support for integrating spatial data into clinical research and predictive models. Expanding upon our previous work [28], existing "Extract, Transform, and Load" (ETL) functions have been developed to allow additional data sources to be collected using the pipeline. The high-level architecture of the pipeline is shown in Figure 2.

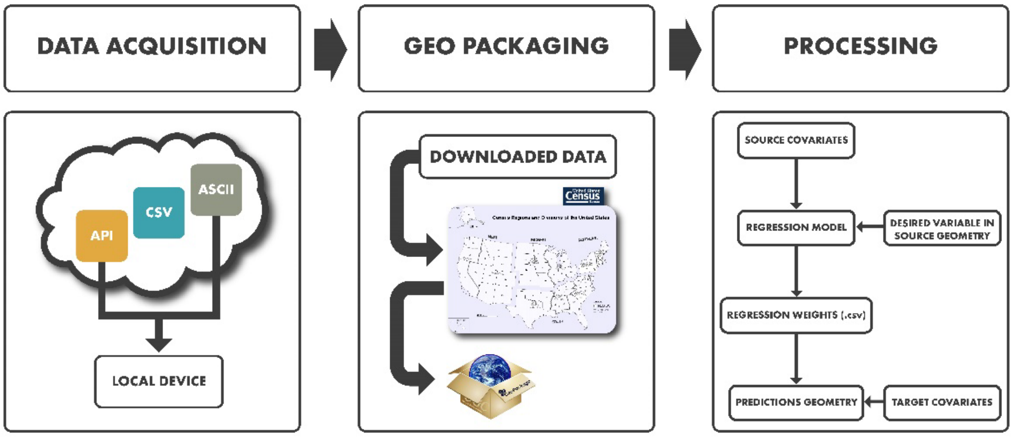

**Figure 2. The computational pipeline acquires data from multiple input formats, creates intermediate cleaned versions of the data, and then applies change of support if needed.**

The key aims in this spatial data pipeline include: a) Creating a data- source- agnostic input format to enable the pipeline to ingest the data from multiple original sources (such as web application programming interfaces (API), comma separated value (CSV) files on a website, or zipped files from a secure file share); b) Performing necessary geospatial transformations, estimation, and/or data cleanup to create an initial internal representation of the raw data; c) Maintaining and documenting provenance of any touches to the data in a standardized schema with accompanying metadata indicating the source, any parameters used, the relevant temporal and spatial scales, and any other information pertinent to the data and to the specific request; and e) Producing and storing standardized data products using CoS as needed.

## 2.2 CHANGE OF SUPPORT (COS)

A key step in the harmonization of multimodal spatial data is the CoS step, sometimes referred to as scaling, with specific methods including small area estimation (SAE) or Gaussian Process approaches. This standardization creates a flexible package that allows data to be formatted for use in a Geographic Information System (GIS), such as a geo-package, exported into modeling software as data frames, or queried in a Structured Query Language (SQL) system. Note that the CoS methodology creates estimates and confidence intervals for the variables of interest at the new spatial reference and that many common datasets, such as the US Census Bureau's Americas Community Survey (ACS), typically provide data of this type (i.e., estimates and confidence intervals). An example of CoS is using a statistical method, such as SAE, to take an original variable at one spatial resolution (such as county-level opioid use rates) and predict it at another, finer, spatial resolution where it is desired (such as US Census Tract 2020 boundaries). Because this CoS methodology is a key algorithm needed in our pipeline, we have broken it out from the pipeline task in subsection 2.1.

In service of the previously mentioned spatial data analysis objectives, we expect a substantial need for the development of algorithms and models for reconciling data occurring at multiple spatial and population scales. This type of problem may arise when harmonizing data captured at different resolutions, such as at the county or zip code level, by downscaling from a coarse resolution to a more finely grained resolution. To accurately capture all available multimodal spatial data, we apply statistical models that consider group-level restrictions, i.e., summary statistics from smaller units must be consistent with those recorded from larger geographical regions.

We anticipate high demand for data products that cover the contiguous 48 states in the continental US, which often requires spatial and temporal interpolation with a robust measure of uncertainty associated with these estimates. Our experience so far in developing probabilistic interpolation workflows for environmental variables suggests that including uncertainty quantification for estimated or interpolated data products can significantly improve the rigor of downstream data analyses, which would otherwise report unwarranted confidence in findings if uncertainty in imputation is not considered.

For small units of study, e.g., census tracts nested within counties, the desired variables may not be available at a fine resolution. To produce synthetic estimates, we fit regression models using covariate variables shared between fine and coarse resolutions and generate posterior predictive distributions of the target variables at a finer resolution (see Figure 3). Another problem occurring in spatial analysis results from the desire to capture cross-variable correlations to infer the spatial distribution of data from one partially observed process, given ubiquitous observational data for another process that is likely to be highly correlated with the first. These two tasks are described as *small area estimation* and *coregionalization*, respectively, within the spatial statistics literature [29].

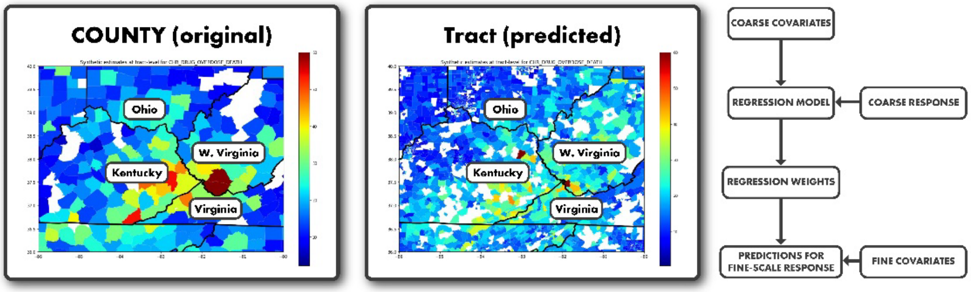

**Figure 3. Example of small area estimation showing original data (left) at county level for predicted drug overdose deaths, along with small area estimates at the tract level (center), and the description of SAE methodology (right).**

## 3. KEY COMMUNITY DATASETS

In collaboration with the VHA, we selected and generated key community datasets, and we delivered the curated data and its documentation to the VHA. The key community dataset categories include the following: a) economic distress (ED), b) social capital and connectedness (SCC), c) lethal means access (LMA), d) healthcare quality and access (HQ&A), and e) environmental determinants of health (EDH). For the key community datasets (i.e., a-e), we provided at least one dataset as well as its detailed documentation. In addition, these datasets are divided into two categories: a) *derivative*, i.e., datasets produced from other datasets by applying a model and creating an index value; and b) *authoritative*, i.e., datasets that have not been modified other than ensuring the inclusion of required geographic administrative boundary identifiers such as FIPS codes. Each key community dataset is explained in detail in the following sections.

### 3.1 ECONOMIC DISTRESS DATASETS

Table 1 lists the Economic Distress datasets included in the SEDH repository. All datasets in this table are authoritative except the Social Vulnerability Index data on rows 19 and 22, which are derivative.

In addition to include a large variety of US Census Bureau data, we included data from the High Intensity Drug Trafficking Areas Program's activities from 2018 to 2021, using FIPS codes at the county level. Also included were small area estimates of selected housing characteristics at the census block level for 2019, including year of construction, residential type and density, rent burden, homeownership costs, and property values for this dataset. In addition, we provided data on the status of Internet access services at the census tract level as of June 30, 2019; for this dataset, we supplied data from residential fixed connections per 1000 homes by census tract. We provided the Veteran population status for the civilian population 18 years of age and over at the county level. And finally, we provided the occupational employment and wage statistics estimates for April 1, 2020, to July 1, 2021, at the state level from the US Census Bureau and the US Bureau of Labor Statistics [30].

**Table 1. Economic Distress Datasets**

| No. | Dataset | Years | Source |
|---|---|---|---|
| 1. | American Community Survey Income Inequality Measures based on Income to Poverty Ratio by Census Block Group [28] | 2015-2019 | US Census Bureau |
| 2. | American Community Survey Income Inequality Measures based on Income to Poverty Ratio, Census Tracts [28] | 2015-2019 | US Census Bureau |
| 3. | American Community Survey Income Inequality Measures based on Household Income Quintiles [28] | 2019 | US Census Bureau |
| 4. | Local Area Unemployment Statistics [28] | 01/2010 - 06/2021 | Bureau of Labor Statistics |
| 5. | Quarterly Census of Employment and Wages [28] | Q1 2016 - Q4 2020. | Bureau of Labor Statistics |
| 6. | Individual-Oriented Social Vulnerability Index, Census Block Groups [28] | 2015-2019 | US Census Bureau |
| 7. | Individual-Oriented Social Vulnerability Index, Census Tracts [28] | 2015-2019 | US Census Bureau |
| 8. | Veteran Segments by Vulnerability Level by Census Block Group [28] | 2015-2019 | US Census Bureau |
| 9. | Veteran Segments by Census Block Group [28] | 2015-2019 | US Census Bureau |
| 10. | Profiles of Veteran Segments by Vulnerability Level [28] | 2015-2019 | US Census Bureau |
| 11. | Veteran Segment Vulnerability Profiles by Vulnerability Level [28] | 2015-2019 | US Census Bureau |
| 12. | Veteran Segment Vulnerability Profiles [28] | 2015-2019 | US Census Bureau |
| 13. | Veteran Segment Service-Connected Disability Profiles by Vulnerability Level [28] | 2015-2019 | US Census Bureau |
| 14. | Veteran Segment Service-Connected Disability Profiles [28] | 2015-2019 | US Census Bureau |
| 15. | Veteran Segments by Vulnerability Level by Census Tract [28] | 2015-2019 | US Census Bureau |
| 16. | Veteran Segments by Census Tract [28] | 2015-2019 | US Census Bureau |
| 17. | CDC/ATSDR Social Vulnerability Index Dataset [31] | 2018 | US Centers for Decease Control/ Agency for Toxic Substances and Disease Registry |

|  |  |  |  |
|---|---|---|---|
| 18. | Block Group Area Deprivation Index Dataset for Washington, DC [31] | 2019 | Neighborhood Atlas, University of Wisconsin, Department of Medicine |
| 19. | Low Food Access Area Dataset for Washington, DC [31] | 2017 | OPEN DATA DC |
| 20. | Eviction Rates (by county) [32] | 2000-2016 | Eviction Lab |
| 21. | Income Inequality (American Community Survey Income Inequality Measures Based on Income to Poverty Ratio by Census Block Group) [32] | 2019 | US Census Bureau |
| 22. | Individual-Oriented Social Vulnerability Index (IOSVI), Census Block Groups [32] | 2019 | US Census Bureau |
| 23. | High Intensity Drug Trafficking Areas (HIDTA) [33] | 2018-2021 | Washington/Baltimore High Intensity Drug Trafficking Areas Program |
| 24. | Small-Area Estimates of Housing Characteristics [33] | 2019 | US Census Bureau |
| 25. | Internet Access Services [33] | 2019 | Federal Communications Commission |
| 26. | Veteran Population Status for the Civilian Population [33] | 2020 | Census Reporter and American Community Survey (ACS) |
| 27. | Occupational Employment and Wage Statistics [30] | 2020-2021 | US Census Bureau and US Bureau of Labor Statistics |

## 3.2 SOCIAL CAPITAL AND CONNECTEDNESS

We recreated a social capital index, originally developed and published by Rupasingha et al. [34], at the county level for each of the 50 states for the years 1990, 1997, 2005, 2009, and 2014. We also created an updated 2019 version of this index based on the method used for the previous years (see social capital in [28] and [31]). We backfilled missing data back to 1990 and collected and processed the relevant variables (relevant establishments per county, voter turnout, census participation, and number of non-profit organizations) for a 2019 update. Four factors were used for the computation of the 2019 index: 1) Establishments per 10,000 population; 2) Voter turnout; 3) Census response rate; and 4) Non-profit organizations per 10,000 population. The social capital index was created using principal component analysis using the above four factors. The four factors were standardized to have a mean of zero and a standard deviation of one, and the first principal component was considered the index of social capital.

We developed a Social Connectedness Index based on Facebook's Social Connectedness Index and data [35] to compute the likelihood of one person being socially connected with someone else within 50, 100, and 500 miles. We investigated the geographical distribution of social networks in the US using anonymized and aggregated data from Facebook. To protect user anonymity, our search only included

FIPS with a total population of at least 500 individuals. These social network statistics were combined with information from the 2015 Census Bureau 5-year and the 2014 Internal Revenue Service (IRS) Individual Income Tax Statistics. In addition, we provided two more measures of connectedness, along with an associated rank for each. The first assessed within-county connectedness, focusing only on connections that fell within the same county. The second measured the total connectedness of each county, regardless of the location of those connections. Both measures were recorded on an integer scale ranging from 1 to $10^9$, following the scale established by the original Facebook data. For each of these measures, a rank was also provided to indicate where this county fell when counties were ordered using that measure.

Table 2 lists the Social Capital and Connectedness Datasets included in the SEDH repository. All datasets in this table are derivatives.

**Table 2. Social Capital and Connectedness Datasets**

| No. | Dataset | Years | Source |
|---|---|---|---|
| 1. | Social Capital Index [28] | 1997, 2005, 2009, 2014, 2019 | US Department of Veterans Affairs |
| 2. | Social Capital Index Dataset [31] | 2019 - Update | US Census Bureau, MIT Election Lab, National Center for Charitable Statistics |
| 3. | Facebook Social Connectedness Index [33] | 2021 | Facebook |

## 3.3 LETHAL MEANS ACCESS

Datasets from the National Instant Criminal Background Check System (NICS) and the Gun Violence Archive (GVA) from the Federal Bureau of Investigation, which record the number of permits and firearm transactions from 1998 to present by state and month, were aggregated by year (See National Instant Criminal Background Check System (NICS) in [28]. Through this study, we investigated the impact of firearm availability on suicide risk. In analyzing NICS background checks on people who wanted to purchase a gun, we identified 300 million checks and 1.5 million denials. The GVA has provided news articles and other information about suicide deaths due to guns since 2013 in aggregated yearly reports. An update on this dataset was issued in [32]. Table 3 lists the lethal means access datasets included in the SEDH repository. All datasets in this table are authoritative.

**Table 3. Lethal Means Access Datasets**

| No. | Dataset | Years | Source |
|---|---|---|---|
| 1. | National Instant Criminal Background Check System (NICS) [28] | 1998-2021 | Federal Bureau of Investigation |
| 2. | National Instant Criminal Background Check System (NICS), Lethal Means Access [32] | 1998-2021 (updated: included the month the data was collected) | Federal Bureau of Investigation |

## 3.4 HEALTHCARE QUALITY AND ACCESS

We collected datasets from the Centers for Medicare and Medicaid Services (CMS). CMS provides health coverage to more than 100 million people through Medicare, Medicaid, the Children's Health Insurance

Program, and the Health Insurance Marketplace. People who are geographically isolated, economically disadvantaged, or medically vulnerable, are served by the Health Resources and Services Administration (HRSA) programs. People living with HIV/AIDS, pregnant women, mothers, and their families, as well as those who are otherwise unable to access high-quality health care are all included in HRSA. HRSA also promotes rural health care access, health professional training, the distribution of providers to areas where they are most needed, and health care delivery improvements. The data include (See Healthcare Quality and Access in [28]): 1) Medicare disparities: quality of care, cost of care, hospital metrics and performance scores; 2) Market saturation and utilization: number of fee-for-service beneficiaries, number of providers, average number of users per provider, etc.; 3) Health professional shortage areas: primary care, dental health, and mental health; 4) Medically underserved areas or populations.

We provided the Medicare Part D Opioid Prescribing Rates by Geography dataset, which includes FIPS geographic comparisons of the quantity and proportion of Medicare Part D opioid prescriptions. The CMS publishes this data in yearly updates; this is the 2019 version of this data [33]. In addition, we provided specific features that were agreed upon during our conversations with our VA's sponsors for the state-level National Mental Health Services Survey from the Substance Abuse and Mental Health Data Archive. We also provided a subset from the 2018-2019 state-level National Survey on Drug Use and Health from the Substance Abuse and Mental Health Services Administration [30].

Table 4 lists the healthcare quality and access datasets included in the SEDH repository. All datasets in this list are authoritative.

**Table 4. Healthcare Quality and Access Dataset**

| No. | Dataset | Years | Source |
|---|---|---|---|
| 1. | Healthcare Quality and Access [28] | 2014-2019 | Centers for Medicare & Medicaid Services (CMS), Health Resource & Services Administration (HRSA) |
| 2. | Medicare Part D Opioid Prescribing Rates [33] | 2019 | Centers for Medicare & Medicaid Services (CMS) |
| 3. | National Mental Health Services Survey [30] | 2018 | Substance Abuse & Mental Health Data Archive |
| 4. | National Survey on Drug Use and Health [30] | 2018-2019 | Substance Abuse and Mental Health Services Administration |

## 3.5 ENVIRONMENTAL DETERMINANTS OF HEALTH

From the US National Cancer Institute, we included county-level UV exposure data for the Continental US [18], which is a 30-year average global solar radiation measure aggregated at the county level. Finally, from the US National Transportation Noise Database, we provided data combining road, aviation, and passenger rail for 2018 [17], as received from the source. Table 5 lists the environmental determinants of health datasets included in the SEDH repository.

**Table 5. Environmental Determinants of Health Datasets**

| No. | Dataset | Years | Source |
|---|---|---|---|
| 1. | USA National Transportation Noise Database 2018 Noise combined data for | 2018 | National Transportation Noise Database |

|  |  |  |  |
|---|---|---|---|
|  | road, aviation, and passenger rail [17] [28] |  |  |
| 2. | County Level UV Exposure Data for the Continental US [18] [28] | 1961-1990, i.e., 30-year average. | USA National Cancer Institute (NIH) |

## 4. ERROR-CHECKING

We give great importance to maintaining the integrity of our datasets. Before a dataset can be deemed production-ready, it must undergo a thorough error-checking assessment by numerous subject matter experts to ensure the accuracy and rigor of the data. We prepared and refined our datasets using standard data and software development methodologies in four work environments: 1) a team-shared work environment where data selection, extraction, and preparation were performed, which we call "development"; 2) a team-shared work environment in the ORNL intranet focused on quality assurance testing (QA), which we call "QA-Intra"; 3) a team-shared work environment in the ORNL Knowledge Discovery Infrastructure (KDI) secure work environment that stores highly sensitive data and ensures its security, which we call "QA-KDI"; and finally, 4) a production environment housed within the KDI environment and accessible to our VA sponsors, which we refer to as "production". As the datasets progressed through the four work environments, we performed test iterations in each work environment to ensure data integrity and compatibility with multiple computing systems. Several test groups were run at each iteration, and a distinct error-checking approach was used for authoritative vs. derivative datasets.

Authoritative datasets, which make up the majority of both our environmental and social data, were error-checked using a data profiling plan that included the following test groups: 1) evaluating missingness by randomly checking for missing data; 2) gathering descriptive statistics like row count, column count, and variable data types; 3) adding checksums to selected columns on both the source and target copies to verify consistency; 4) consistently using FIPS codes as geographic administrative boundaries to represent the social and physical environment and verifying that the FIPS codes matched the geographic administrative boundaries of the original data; and 5) manually comparing the first, last, and five additional randomly chosen rows for consistency between the source and target datasets.

Derivative datasets, which account for only about 5% of our datasets, were error-checked in each of the four work environments using a combination of statistical methodologies based on each dataset's properties in addition to the data profiling methodology used for authoritative datasets as described above. Statistical error-checking included the following:

a) Social capital dataset: Principal Component Analysis (PCA) was used to generate the social capital datasets using an existing composite index approach based on the work of Rupasingha et al. [34], followed by Factor Analysis to error-check the PCA findings.

b) (b) Individual-focused social vulnerability index dataset: The strength and direction of the linear link between variables were assessed using Pearson's correlation ($p < 0.001$ for social factors and $p < 0.05$ for environmental factors) for the individual-focused social vulnerability index dataset [36].

c) (c) Social connectedness dataset: Visualization and statistical correlation methods were used to error-check the values of the social connectedness dataset, based on the work of Bailey et al. [35]. The index values were transformed to provide distance-fading social relationships for counties rather than national connections for each county to every other county, providing a fresh viewpoint on the raw data.

Table 6 summarizes the cumulative error-checking test findings for the 25 datasets delivered in fiscal year 2022 and displays the outcomes per test iteration. The table's rows indicate the test group, and the columns represent each test iteration completed in a particular work environment.

**Table 6. Cumulative Error-Checking Test Results Performed during the Four Quarters of Fiscal Year 2022.**

| Test Group | Development | | QA-Intra | | QA-KDI | | Production | | Error Ratio |
|---|---|---|---|---|---|---|---|---|---|
| | Passes | Fails | Passes | Fails | Passes | Fails | Passes | Fails | |
| 1 | 21 | 4 | 25 | 0 | 23 | 2 | 25 | 0 | 0.064 |
| 2 | 25 | 0 | 25 | 0 | 25 | 0 | 25 | 0 | 0 |
| 3 | 25 | 0 | 25 | 0 | 25 | 0 | 25 | 0 | 0 |
| 4 | 25 | 0 | 25 | 0 | 25 | 0 | 25 | 0 | 0 |
| 5 | 25 | 0 | 25 | 0 | 25 | 0 | 25 | 0 | 0 |

We have partially automated the profiling plan described above using the pipeline architecture to reduce data inaccuracies caused by inadvertent human error while moving datasets from various work environments. Future error-checking methods will be modified based on the types of customized datasets encountered in the project. We plan to completely automate our error-checking processes for each dataset as we add new ones during each quarterly release.

## 5. DATA CURATION DOCUMENTATION

We developed technical documentation and metadata for the key community variables described in the previous section, including the provenance of the source data, a summary of modeling strategies employed in the curation processes, assumptions made in the modeling strategies, limitations in the interpretation of curated variables, and/or other caveats that should be considered in the use of the curated data. The format of our documentation follows that of the Agency for Healthcare Research and Quality's (AHRQ) Social Determinants of Health (SDoH) datasets.

We provided technical document reports with overviews for each dataset using the AHRQ's SDoH documentation as a template but adding in additional information for CoS when used. These documents include the following fields for each dataset:

- Sponsor (name of the organization that provided the raw data, e.g., Health Resources and Services Administration [HRSA] for the Area Health Resources Files [AHRF]).
- Description (a brief, general description of the data, including: a summary of modeling strategies employed in the curation, years of coverage, assumptions made in the modeling strategy employed, limitations on interpretation of curated variables, and/or other caveats that should be considered in the use of the curated data).
- Inclusion in the datasets: a) Lists the domains to which the data source has contributed variables; b) Includes additional information about the data source relevant to the dataset.
- Resources (links to original data source documentation, data download sites, and other relevant information).
- Variable definitions and specifications (in tabular format) for each column value: Variable name, Variable label, Source table (if multiple data tables were available from the original data source); Numerator (for derived variables) and Denominator (for derived variables) or original variable (when renamed for the SEDH repository). The numerators and denominators for the variables and their sources are shown following each data source description.
- Variable availability across years (in tabular format), which include a) Variable name, b) Variable label, and Data year availability (e.g., 2009–2018).

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

- Variable assignment to annual datasets. Variables appear in the annual datasets that correspond with (1) the single year represented by the original data source (e.g., Nursing Home Compare data for facilities in 2016 appears in the 2016 county dataset) or (2) the last year in a period represented by the data (e.g., ACS data aggregated over 2012 to 2016 is in the 2016 dataset).
- Variable availability. The availability of each variable changes across data years. Following each data source description, we provide a table showing the availability of each variable in the annual datasets.
- Variable naming. Except for the geographic ID variables, all variable names begin with a data source acronym followed by an underscore and a descriptive title.
- Missing values. The datasets use a blank, or 'NA', to denote a missing value almost exclusively. The one exception is the provider ratio variables from the County Health Rankings (CHR) data, which have negative values for counties where the number of providers is zero. This is described further in the description of the CHR data.

Currently, documentation publicly available include the following sponsor reports:

- 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*. United States: N. p., 2022. Web. doi:10.2172/1892396, [30].
- 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*. United States: N. p., 2022. Web. doi:10.2172/1876283, [33].
- Christian, 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*. United States: N. p., 2022. Web. doi:10.2172/1862127, [32].
- Christian, B., Klasky, H. B., Sparks, K., Peluso, A., Tuccillo, J., Devineni, P., and Watson, R. *VA EDH Data Curation Documentation FY22-Q1, Rev. 2*. United States: N. p., 2021. Web. doi:10.2172/1854460, [31].
- Christian, B., Branstetter, M., Klasky, H.- B., Rastogi, D., Sparks, K., Tuccillo, J., Watson, R., Yoon, HJ, and Kim, Y. *VA EDH Data Curation Documentation – FY21, Rev. 2*. United States: N. p., 2022. Web. doi:10.2172/1854468, [28].

## 6. LIMITATIONS AND FUTURE WORK

There are several limitations in this work. The following datasets were missing at the time this manuscript was written: outdoor ambient air pollutants, climate such as CDC's Wide-ranging Online Data for Epidemiologic Research (WONDER) for the environment, industrial facility maps, green space areas, drinking water quality, addictive substances in the community, state and local policies related to substance abuse, and the Geographic microdata from the US Environmental Protection Agency's Risk-Screening Environmental Indicators.

In addition, as the source data are updated, the existing datasets will be provided with finer resolutions and fresh updates. Future plans include investigating the relationship between economic conditions, social vulnerability, and individual vulnerability (including for veteran population segments in the United States) based on individual data from the PUMS and developing small-area poverty rates adjusted for region-specific cost of living using population estimates derived from the PUMS and criteria on household composition, homeownership status, and income from the US Census Bureau's Supplemental Poverty Measure.

## 7. CONCLUSION

This manuscript provides an overview of the Social and Environmental Determinants of Health (SEDH) dataset repository that contains novel datasets associated with health outcomes, data curation documentation, a pipeline framework for data generation, a methodology for converting spatiotemporal data from one spatial reference (such as a 1 km grid) to another (such as US Census Tracts), and health outcomes modeling capabilities with funding from the VA Office of Mental Health and Suicide Prevention (OMHSP).

## 8. ACKNOWLEDGEMENTS

This work is sponsored by the US Department of Veterans Affairs. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

We thank our colleague Dr. Chris Krapu for helpful discussions regarding CoS.

# 9. REFERENCES

[1] "Department of Veterans Affairs. 2021 National Veteran suicide prevention annual report." https://www.mentalhealth.va.gov/mentalhealth/suicide_prevention/data.asp (accessed April 2022.

[2] H. Hedegaard and M. Warner, "Suicide mortality in the United States, 1999-2019," 2021.

[3] (2021). *Performance Work Statement - Part B Task Identifying and Modeling Clinically Important Patterns in VA Medical Record Data Relevant to Suicide and Overdose Prevention, Pain Mangement and Opioid Safety*

[4] T. Kiserud *et al.*, "The World Health Organization Fetal Growth Charts: A Multinational Longitudinal Study of Ultrasound Biometric Measurements and Estimated Fetal Weight," (in en), *PLOS Medicine,* vol. 14, no. 1, p. e1002220, Jan 24, 2017 2017, doi: 10.1371/journal.pmed.1002220.

[5] J. F. McCarthy *et al.*, "Evaluation of the Recovery Engagement and Coordination for Health–Veterans Enhanced Treatment Suicide Risk Modeling Clinical Program in the Veterans Health Administration," *JAMA Network Open,* vol. 4, no. 10, pp. e2129900-e2129900, 2021, doi: 10.1001/jamanetworkopen.2021.29900.

[6] "US Census Bureau." https://www.census.gov/ (accessed 2020).

[7] "US Centers for Disease Control and Prevention." https://www.cdc.gov/ (accessed 2020).

[8] "Neighborhood Atlas - Center for Heaalth Disparities Research, University of Wisconsin, School of Medicine and Public Health." https://www.neighborhoodatlas.medicine.wisc.edu/ (accessed 2020).

[9] "Open Data DC." https://opendata.dc.gov/ (accessed 2020).

[10] "Eviction Lab." https://evictionlab.org/ (accessed 2020).

[11] "High Intensity Drug Trafficking Areas (HIDTA) - US Drug Enforcement Administration." https://www.dea.gov/operations/hidta (accessed 2022).

[12] "MIT Elecction Lab." https://electionlab.mit.edu/ (accessed 2020).

[13] "Urban Institute - National Center for Charitable Statistics." https://nccs.urban.org/ (accessed 2020).

[14] "Meta Data for Independent Research." https://research.facebook.com/data/ (accessed 2022).

[15] "National Instant Criminal Background Check - Federal Bureau of Investigation." https://www.fbi.gov/services/cjis/nics (accessed 2020).

[16] "Centers for Medicare & medicaid Services - Research, Statistics, Data & Systems." https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems (accessed.

[17] "National Trasnportation Noise Data." https://hub.arcgis.com/documents/usdot::2018-noise-data/about (accessed 2020).

[18] "County Level UV Exposure Data for the Continental United States." https://gis.cancer.gov/tools/uv-exposure/ (accessed 2020).

[19] "US Environmental Protection Agency." https://www.epa.gov/ (accessed 2020).

[20] "American Community Survey (ACS)." https://www.census.gov/programs-surveys/acs (accessed 2020).

[21] "Food Access Research Atlas." https://www.ers.usda.gov/data-products/food-access-research-atlas.aspx (accessed 2020).

[22] "US Department of Houseing and Urban Development." https://www.hud.gov/ (accessed 2020).

[23] "US Bureaud of Labor Statistics." https://www.bls.gov/ (accessed 2020).

[24] E. C. Fradelos, I. V. Papathanasiou, D. Mitsi, K. Tsaras, C. F. Kleisiaris, and L. Kourkouta, "Health based geographic information systems (GIS) and their applications," *Acta Informatica Medica,* vol. 22, no. 6, p. 402, 2014.

[25] A.-K. Lyseen *et al.*, "A review and framework for categorizing current research and development in health related geographical information systems (GIS) studies," *Yearbook of medical informatics,* vol. 23, no. 01, pp. 110-124, 2014.

[26] C. I. Nykiforuk and L. M. Flaman, "Geographic information systems (GIS) for health promotion and public health: a review," *Health promotion practice,* vol. 12, no. 1, pp. 63-73, 2011.

[27] "Agency for Health Research and Quality." https://www.ahrq.gov/ (accessed 2022).

[28] B. Christian *et al.*, "VA EDH Data Curation Documentation – FY21, Rev. 2," Oak Ridge National Laboratory, United States, 2022. [Online]. Available: https://www.osti.gov/biblio/1854468-va-edh-data-curation-documentation-fy21-rev

[29] J. N. Rao and I. Molina, *Small area estimation*. John Wiley & Sons, 2015.

[30] H. B. Klasky *et al.*, "VA EDH Data Curation Documentation FY22-Q4," Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2022.

[31] B. Christian *et al.*, "VA EDH Data Curation Documentation FY22-Q1, Rev. 2," Oak Ridge National Laboratory, United States, 12 2021. [Online]. Available: https://www.osti.gov/biblio/1854460-va-edh-data-curation-documentation-fy22-q1-rev

[32] B. Christian *et al.*, "VA EDH Data Curation Documentation FY22-Q2, Rev. 2," Oak Ridge National Laboratory, United States, 3 2022. [Online]. Available: https://www.osti.gov/biblio/1862127-va-edh-data-curation-documentation-fy22-q2-rev

[33] H. Klasky, K. Sparks, and J. Logan, Tuccillo, Joe, Whitehead, Matthew, Hamaker, Alec, Hanson, Heidi, Watson, Rochelle, and Kapadia, Anuj., "VA EDH Data Curation Documentation - FY22-Q3. ," Oak Ridge National Laboratory, 2022. [Online]. Available: https://www.osti.gov/biblio/1876283-va-edh-data-curation-documentation-fy22-q3

[34] A. Rupasingha, S. J. Goetz, and D. Freshwater, "The production of social capital in US counties," *The journal of socio-economics,* vol. 35, no. 1, pp. 83-101, 2006.

[35] M. Bailey, P. Farrell, T. Kuchler, and J. Stroebel, "Social connectedness in urban areas," *Journal of Urban Economics,* vol. 118, p. 103264, 2020.

[36] J. V. Tuccillo, "A Multiscalar Index of Social Vulnerability for the United States," 2022.