



ScHARe

ScHARe Repository Introduction

November 20, 2024

Deborah Duran, PhD • NIMHD

Elif Dede Yildirim, PhD • NIMHD

Mark Aronson, PhD • NIMHD



ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

Outline

- 15'** ScHARe Overview
- 5'** Repository Background
- 5'** Getting Started
- 15'** Uploading your first Data Set
- 15'** **HANDS ON: Uploading Data**
- 15'** CDE Mapping and Dataviews
- 15'** **HANDS ON: Dataviews for CDE Mapping**
- 5'** Sharing Data
- 10'** Data Aggregation and Analysis - Overview
- 15'** Conclusion and Q&A

Experience poll

Please check your level of experience with the following:

	None	Some	Proficient	Expert
Python	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
R	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Cloud computing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Terra	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Health disparities research	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Health outcomes research	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Algorithmic bias mitigation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Interest poll

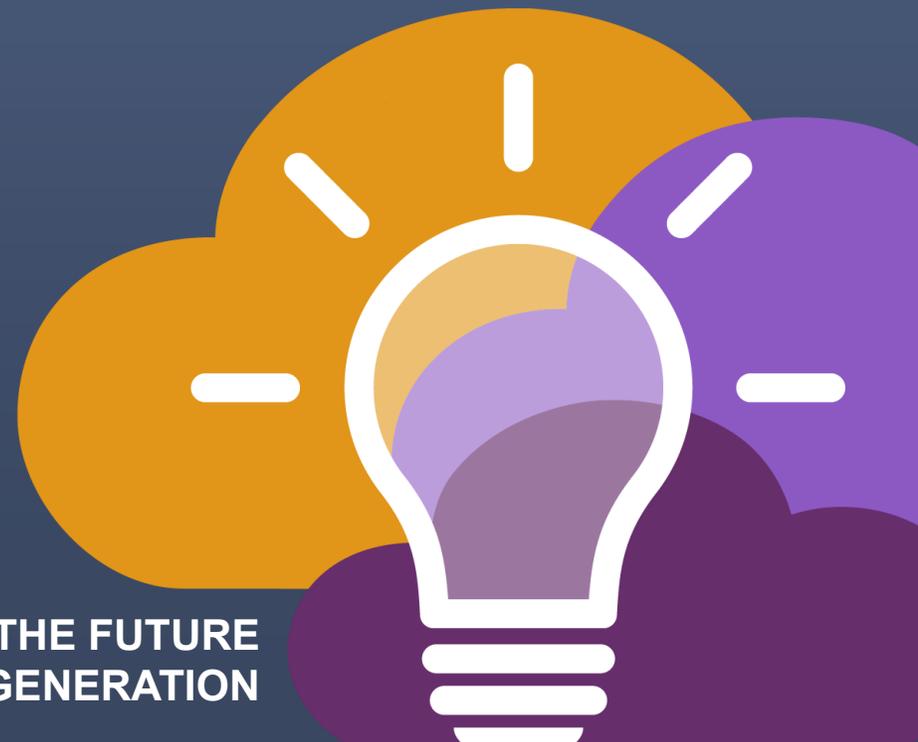
I am interested in (check all that apply):

- Learning about Health Disparities and Health Outcomes research to apply my data science skills
- Conducting my own research using AI/cloud computing and publishing papers
- Connecting with new collaborators to conduct research using AI/cloud computing and publish papers
- Learning to use AI tools and cloud computing to gain new skills for research using Big Data
- Learning cloud computing resources to implement my own cloud
- Developing bias mitigation and ethical AI strategies
- Other

ScHARe

What is ScHARe?

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction



Register: nimhd.nih.gov/schare

ScHARe is a **cloud-based population science data platform** designed to accelerate research in health disparities, health and healthcare delivery outcomes, and artificial intelligence (AI) bias mitigation strategies

ScHARe aims to fill **five critical gaps**:

- Increase participation of **women & underrepresented populations with health disparities** in data science through data science skills training, cross-discipline mentoring, and multi-career level collaborating on research
- Leverage population science, SDoH, and behavioral Big Data and cloud computing tools to foster a **paradigm shift** in health disparity and healthcare delivery outcomes research
- **Advance AI bias mitigation and ethical inquiry** by developing innovative strategies and securing diverse perspectives
- Provide a **data science cloud computing resource** for community colleges and low resource minority serving institutions and organizations
- Offer a **project data repository** centered on core common data elements for enhanced data interoperability and compliance with NIH Data Management and Sharing Policy



ScHARe



Google Platform Terra Interface

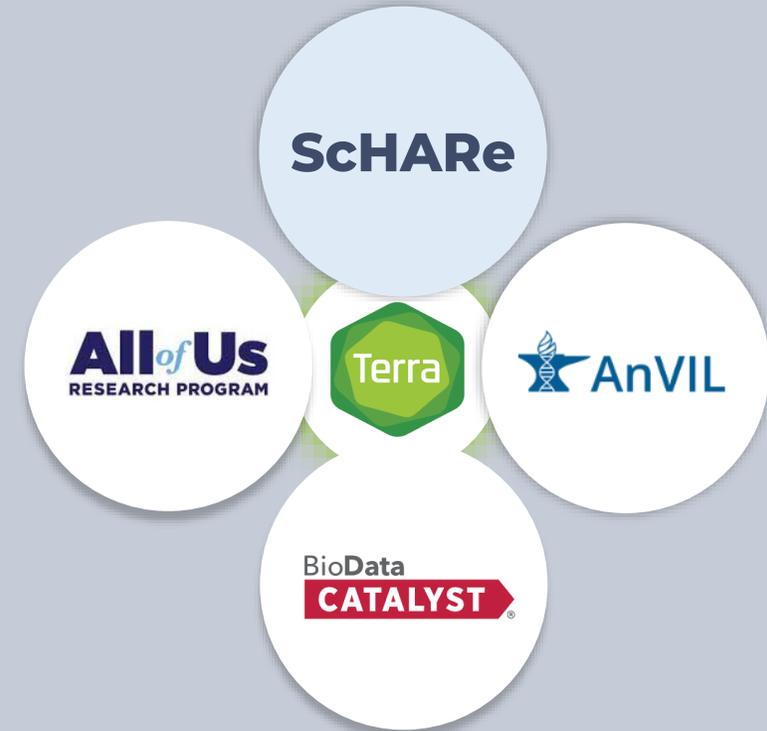
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Copy-and-paste code in Python and R
- Learning Terra on ScHARe prepares you to use other NIH platforms



Terra recommends using **Chrome**
Must have a **Gmail** friendly account

PREPARING FOR AI RESEARCH AND HEALTHCARE USING BIG DATA

Mapping across cloud platforms with
Terra interface for collaborative research



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

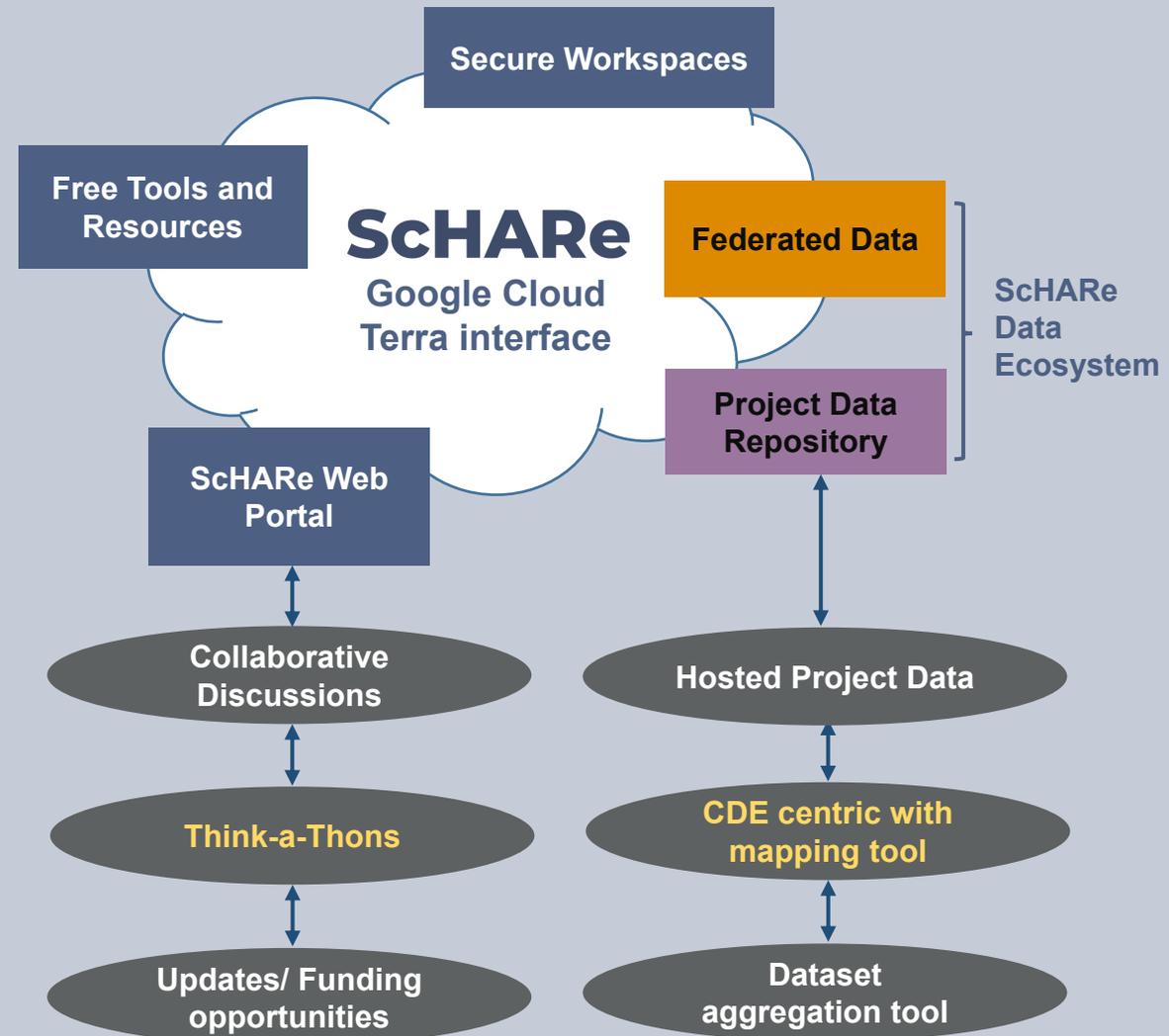


ScHARe Components

ScHARe co-localizes within the cloud:

1. **Datasets** (including social determinants of health and social science data) relevant to minority health, health disparities, and healthcare outcomes research
2. **CDE-focused data repository** to comply with the required hosting and sharing of data from NIMHD-/NINR-funded programs
3. **User-friendly computational capabilities and secure, collaborative workspaces** for students and all career level researchers
4. **Tools for collaboratively evaluating and mitigating biases** associated with datasets and algorithms utilized to inform healthcare and policy decisions (*upcoming*)

Intramural and Extramural Resource



ScHARe Terra interface: secure workspace

The screenshot displays the ScHARe Terra interface with a 'Share Workspace' modal dialog open. The background shows a workspace list with 'ScHARe' and 'ScHARe Think-a-Thons'. The dialog box is titled 'Share Workspace' and contains the following elements:

- User email:** A text input field with the placeholder 'Add people or groups' and an 'ADD' button.
- Current Collaborators:** A list of collaborators with their roles and permissions:
 - calzonil2@nih.gov:** Role: Owner (dropdown), Permissions: Can share, Can compute.
 - ScHARe-Contractors@firecloud.org:** Role: Writer (dropdown), Permissions: Can share, Can compute. Includes a close button (X).
 - ScHARe-Read-Only-Access@firecloud.org:** Role: Reader (dropdown), Permissions: Can share, Can compute. Includes a close button (X).
- Share with Support:** A toggle switch currently set to 'No'.
- Buttons:** 'CANCEL' and 'SAVE' buttons at the bottom right.

- Secure workspace for self or collaborative research
- Assign roles: review or admin
- Host own data and code

ScHARe Terra interface: analyses

Notebooks for analytics and tutorials

WORKSPACES
Workspaces > ScHARe/ScHARe > Analyses

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

Your Analyses + START

Application	Name ↓
Jupyter	00_List of Datasets Available on ScHARe.ipynb
Jupyter	01_Introduction to Terra Cloud Environment.ipynb
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb
Jupyter	03_R Environment setup.ipynb
Jupyter	04_Python 3 Environment setup.ipynb
Jupyter	05_How to access plot and save data from public BigQuery datasets using R.ipynb
Jupyter	06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

Modular codes

- Easy-to-use copy and paste analytics

WORKSPACES
Workspaces > ScHARe/ScHARe > ANALYSES

DASHBOARD DATA ANALYSES

WORKFLOWS

Find a Workflow

Suggested Workflows

- haplotypecaller-gvcf-gatk4
Runs HaplotypeCaller from GATK4 in GVCF mode on a single sample.
- mutect2-gatk4
Implements GATK4 Mutect 2 on a single tumor-normal pair.
- processing-for-variant-discovery-gatk4

Find Additional Workflows

Dockstore
Browse WDL workflows in Dockstore, an open platform used by the GA4GH for sharing Docker-based workflows.

- Modular codes developed for reuse
- Adding SAS

ScHARe Terra interface: access to datasets

What data?

The screenshot shows the ScHARe Terra interface in the 'Analyses' tab. The notebook '00_List of Datasets Available on ScHARe.ipynb' is open, displaying the 'The ScHARe Data Ecosystem' section. This section provides an overview of the data ecosystem and lists various dataset categories:

- Google Cloud Public Datasets** - Publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program. Examples: US Census Data: American Community Survey (ACS)
- ScHARe Hosted Public Datasets** - Publicly accessible, de-identified datasets hosted by ScHARe. Examples: Social Vulnerability Index (SVI), Behavioral Risk Factor Surveillance System (BRFSS)
- Funded Datasets on ScHARe** - Publicly accessible and controlled-access, funded program/project datasets shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy. Example: Jackson Heart Study (JHS).

A detailed list of the datasets available in the ScHARe Data Ecosystem, including links to documentation and other helpful resources for each dataset, is available in the sections below. The datasets are categorized as follows, based on their content:

A - SOCIAL DETERMINANTS OF HEALTH

- A1 Multiple Categories:** Datasets that include data on multiple Social Determinants of Health (SDoH) factors/indicators
- A2 Economic Stability:** Datasets that include data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.
- A3 Education Access and Quality** Datasets that include data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.
- A4 Health Care Access and Quality** Datasets that include data on health literacy, use of health IT, emergency room waiting times, evidence-based preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.
- A5 Neighborhood and Built Environment** Datasets that include data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.
- A6 Social and Community Context** Datasets that include data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc. For incidence and prevalence of anxiety, depression, and other mental health conditions, see section 'B1 - Diseases and conditions' below
- A7 Health Behaviors** Datasets that include data on health behaviors

B - HEALTH OUTCOMES

In the **Analyses** tab, the notebook **00_List of Datasets Available on ScHARe** lists all datasets

Where?

The screenshot shows the ScHARe Terra interface in the 'Data' tab. A table of datasets is displayed, with the 'EconomicStability' category selected. The table includes columns for dataset names and their sizes in gigabytes (Gb).

Dataset Name	SizeGb
EconomicStability_Id	
FoodAccessResearchAtlasData2010	0.0297
CurrentPopulationSurvey_FoodSecuritySupplement_2011	0.184
CurrentPopulationSurvey_FoodSecuritySupplement_2012	0.185
CurrentPopulationSurvey_FoodSecuritySupplement_2013	0.184
CurrentPopulationSurvey_FoodSecuritySupplement_2014	0.188
AHS_National_Household_2015	0.491
AHS_National_Mortgage_2015	0.002
AHS_National_Person_2015	0.057
AHS_National_Project_2015	0.004
CurrentPopulationSurvey_FoodSecuritySupplement_2015	0.185

In the **Data** tab, data tables help access data

ScHARe Ecosystem structure

Researchers can access, link, analyze, and export a **wealth of SDoH and population science related datasets** within and across platforms relevant to research about health disparities, health care delivery, health outcomes and bias mitigation, including:

280+
FEDERATED
PUBLIC
DATASETS

Public datasets

Publicly accessible, federated, de-identified datasets hosted by ScHARe or hosted by Google through the Google Cloud Public Dataset Program

ScHARe e.g.: *Behavioral Risk Factor Surveillance System (BRFSS)*
Google e.g.: *American Community Survey (ACS)*

CDE
FOCUSED
REPOSITORY

Funded datasets

Publicly accessible and controlled-access, funded program/project datasets using Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

e.g.: *Jackson Heart Study (JHS)*
Extramural Grant Data
Intramural Project Data

Innovative Approach:
CDE Concept Codes
Uniform Resource Identifier (**URI**)

ScHARe Ecosystem

Datasets are categorized by content based on the CDC **Social Determinants of Health** categories:

1. Economic Stability
2. Education Access and Quality
3. Health Care Access and Quality
4. Neighborhood and Built Environment
5. Social and Community Context

with the addition of:

- **Health Behaviors**
- **Diseases and Conditions**

The screenshot shows the Terra WORKSPACES Data interface. The top navigation bar includes 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'DATA' tab is active, displaying an 'IMPORT DATA' button and a search bar for tables. A list of tables is shown on the left, with 'EconomicStability (62)' highlighted. The main table on the right lists various datasets with their names and sizes in GB.

Table Name	Count	Dataset Name	SizeGb
A_MainTableDatasets	250	EconomicStability_id	0.0297
DiseaseAndConditions	27	FoodAccessResearchAtlasData2010	0.184
EconomicStability	62	CurrentPopulationSurvey_FoodSecuritySupplement_2011	0.185
EducationAccessAndQuality	54	CurrentPopulationSurvey_FoodSecuritySupplement_2012	0.184
HealthBehaviors	17	CurrentPopulationSurvey_FoodSecuritySupplement_2013	0.188
HealthCareAccessAndQuality	36	CurrentPopulationSurvey_FoodSecuritySupplement_2014	0.491
MultipleCategories	38	AHS_National_Household_2015	0.002
NeighborhoodAndBuiltEnvironment	11	AHS_National_Mortgage_2015	0.057
SocialAndCommunityContext	8	AHS_National_Person_2015	0.004
		AHS_National_Project_2015	0.004
		CurrentPopulationSurvey_FoodSecuritySupplement_2015	0.184



ScHARe Ecosystem: ScHARe hosted datasets

Organized based on the **CDC SDoH categories**, with the addition of *Health Behaviors and Diseases and Conditions*:

280+ datasets

- What are the Social Determinants of Health?

Social determinants of health (SDoH) are the **nonmedical factors that influence health outcomes**

They are the **conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life**



https://www.cdc.gov/about/priorities/social-determinants-of-health-at-cdc.html?CDC_AAref_Val=https://www.cdc.gov/about/sdoh/index.html

ScHARe Ecosystem: ScHARe hosted datasets

Education access and quality

Data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.

Health care access and quality

Data on health literacy, use of health IT, preventive healthcare, access to health insurance, etc.

Neighborhood and built environment

Data on access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, noise exposure, smoking, mass transit use, etc.

Social and community context

Data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc.

Economic stability

Data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.

* Health behaviors

Data on health-related practices that can directly affect health outcomes.

* Diseases and conditions

Data on incidence and prevalence of specific diseases and health conditions.



** Not Social Determinants of Health*

ScHARe Ecosystem: Google hosted datasets

Examples of interesting datasets include:

- **American Community Survey** (U.S. Census Bureau)
- **US Census Data** (U.S. Census Bureau)
- **Area Deprivation Index** (BroadStreet)
- **GDP and Income by County** (Bureau of Economic Analysis)
- **US Inflation and Unemployment** (U.S. Bureau of Labor Statistics)
- **Quarterly Census of Employment and Wages** (U.S. Bureau of Labor Statistics)
- **Point-in-Time Homelessness Count** (U.S. Dept. of Housing and Urban Development)
- **Low Income Housing Tax Credit Program** (U.S. Dept. of Housing and Urban Development)
- **US Residential Real Estate Data** (House Canary)
- **Center for Medicare and Medicaid Services - Dual Enrollment** (U.S. Dept. of Health & Human Services)
- **Medicare** (U.S. Dept. of Health & Human Services)
- **Health Professional Shortage Areas** (U.S. Dept. of Health & Human Services)
- **CDC Births Data Summary** (Centers for Disease Control)
- **COVID-19 Data Repository by CSSE at JHU** (Johns Hopkins University)
- **COVID-19 Mobility Impact** (Geotab)
- **COVID-19 Open Data** (Google BigQuery Public Datasets Program)
- **COVID-19 Vaccination Access** (Google BigQuery Public Datasets Program)



How to access Google hosted datasets

Big Query

The Google public datasets are **available for access on Terra using BigQuery**

- BigQuery is the Google Cloud storage solution for structured data
- It is easy to use, works with large amounts of data and offers fast data retrieval and analysis
- **Our instructional notebooks in the Analyses tab** provide code and instructions on using Big Query to access Google datasets

```
Jupyter 06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb
```

The following Python code will read a BigQuery table into a Pandas dataframe.

From <https://cloud.google.com/community/tutorials/bigquery-ibis>

ibis is a Python library for doing data analysis. It offers a Pandas-like environment for executing data analysis composable, and familiar replacement for SQL.

```
In [9]: # Connect to the dataset
conn = ibis.bigquery.connect(dataset_id='bigquery-public-data.broadstreet_adi')
```

```
In [10]: # Read table
ADI_table_2 = conn.table('area_deprivation_index_by_census_block_group')
ADI_table_2
```

```
Out[10]: BigQueryTable[table]
name: bigquery-public-data.broadstreet_adi.area_deprivation_index_by_census_block_group
schema:
  geo_id : string
  state_fips_code : string
  county_fips_code : string
  block_group_fips_code : string
  description : string
  county_name : string
  state_name : string
  state : string
  year : int64
  area_deprivation_index_percent : float64
```



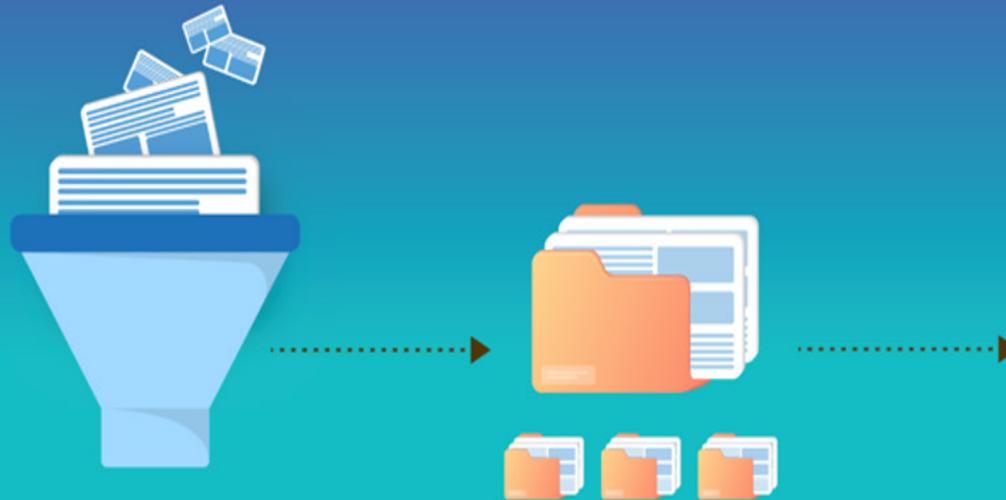
CDE benefits:

- Faster start-up for project
- Better data aggregation across projects
- Shared meaning
- Concept-focused to allow questions/answers variations
- Coding enables an URI approach for better data interoperability

A **Common Data Element (CDE)** is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection

Because Researchers use CDEs...

they can more quickly share data and get results faster, which ultimately can help make a **meaningful difference to our nation's health.**



For more information about how CDEs accelerate research discoveries, visit: cde.nlm.nih.gov/resources

ScHARe Core CDEs

PhenX Toolkit

- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex
- Gender
- Sexual Orientation
- Marital Status
- Education
- Annual Household Income
- Household Size

- English Proficiency
- Disabilities
- Health Insurance
- Employment Status
- Usual Place of Health Care
- Financial Security / Social Needs
- Self-Reported Health
- Health Conditions (and Associated Medications/Treatments)

- **NIMHD Framework***
- **Health Disparity Outcomes***

* Project Level CDEs

NIH Endorsed



ScHARe has developed **Common Data Elements** to ensure consistent data collection across studies, facilitate interoperability, and link data from different sources

NIH CDE Repository:

cde.nlm.nih.gov/home

PhenX Toolkit:

www.nimhd.nih.gov/resources/phenx/

NIMHD Research Framework

		Levels of Influence*			
		Individual	Interpersonal	Community	Societal
Domains of Influence (Over the Lifecourse)	Biological	Biological Vulnerability and Mechanisms	Caregiver–Child Interaction Family Microbiome	Community Illness Exposure Herd Immunity	Sanitation Immunization Pathogen Exposure
	Behavioral	Health Behaviors Coping Strategies	Family Functioning School/Work Functioning	Community Functioning	Policies and Laws
	Physical/Built Environment	Personal Environment	Household Environment School/Work Environment	Community Environment Community Resources	Societal Structure
	Sociocultural Environment	Sociodemographics Limited English Cultural Identity Response to Discrimination	Social Networks Family/Peer Norms Interpersonal Discrimination	Community Norms Local Structural Discrimination	Social Norms Societal Structural Discrimination
	Health Care System	Insurance Coverage Health Literacy Treatment Preferences	Patient–Clinician Relationship Medical Decision-Making	Availability of Services Safety Net Services	Quality of Care Health Care Policies
Health Outcomes		 Individual Health	 Family/ Organizational Health	 Community Health	 Population Health



Project Level CCDEs – Framework

What NIMHD Research framework levels and domains of influence is your study targeting? (Select all that apply)

Levels of Influence

- Individual
- Interpersonal
- Community
- Societal

Domains of Influence

- Biological
- Behavioral
- Physical/Built Environments
- Sociocultural Environment
- Health Care Systems and Clinical Care

NIMHD Research Framework. <https://www.nimhd.nih.gov/about/overview/research-framework/nimhd-framework.html>



NIMHD's Mission: Improve Minority Health

Minority Health:
Distinctive health characteristics and attributes of racial and/or ethnic minority populations who are socially disadvantaged due in part to being subject to racist or discriminatory acts and are underserved in health care.

Minority Health Research

The scientific investigation of singular and combinations of attributes, characteristics, behaviors, biology, and societal and environmental factors that influence the health of minority racial and/or ethnic population(s), including within-group or ethnic sub-populations, with the goals of improving health and preventing disease.

Minority Health Populations

The OMB Directive 15 defines racial and ethnic minority populations as:

- American Indian or Alaska Native
- Asian
- Black or African American
- Hispanic or Latino American
- Middle Eastern or North African
- Native Hawaiian or Pacific Islander



NIMHD's Mission: Reduce Health Disparities

Health Disparity:

A health disparity is a health difference that adversely affects disadvantaged populations in comparison to a reference population, based on one or more health outcomes.

All populations with health disparities are socially disadvantaged due in part to being subject to racist or discriminatory acts and are underserved in health care.

Health Disparity Research

A multi-disciplinary field of study devoted to:

- Gaining greater scientific knowledge about the influence of health determinants.
- Understanding the role of mechanisms.
- Determining how this knowledge is translated into interventions to reduce or eliminate adverse health outcomes.

Populations with Health Disparities

Populations that experience health disparities include:

- Racial and ethnic minority groups
- People with lower socioeconomic status (SES)
- Underserved rural communities
- Sexual and gender minority (SGM) groups
- People with disabilities



Health Disparity Outcomes

Unfair disadvantages that people face in different aspects of life, like education, income, or opportunities **can lead to health disparities**

Some groups of people may experience poorer health outcomes than others as a result

Health Disparity Outcomes

The health outcomes are categorized as:

- Higher incidence and/or prevalence of disease, including earlier onset or more aggressive progression of disease.
- Premature or excessive mortality from specific health conditions.
- Greater global burden of disease, such as Disability Adjusted Life Years (DALY), as measured by population health metrics.
- Poorer health behaviors and clinical outcomes related to the aforementioned.
- Worse outcomes on validated self-reported measures that reflect daily functioning or symptoms from specific conditions.



Project Level CCDEs – Research Area Focus

Which of the following content areas of research is this study addressing, if any? Select all that apply.

- Minority health
- Health Disparity (select the focus area)
 - Higher incidence and/or prevalence of disease, including earlier onset or more aggressive progression of disease
 - Premature or excessive mortality from specific health conditions
 - Greater global burden of disease, such as Disability Adjusted Life Years (DALY), as measured by population health metrics
 - Poorer health behaviors and clinical outcomes related to the aforementioned
 - Worse outcomes on validated self-reported measures that reflect daily functioning or symptoms from specific conditions
- Other Health Outcomes / Health Delivery or care

