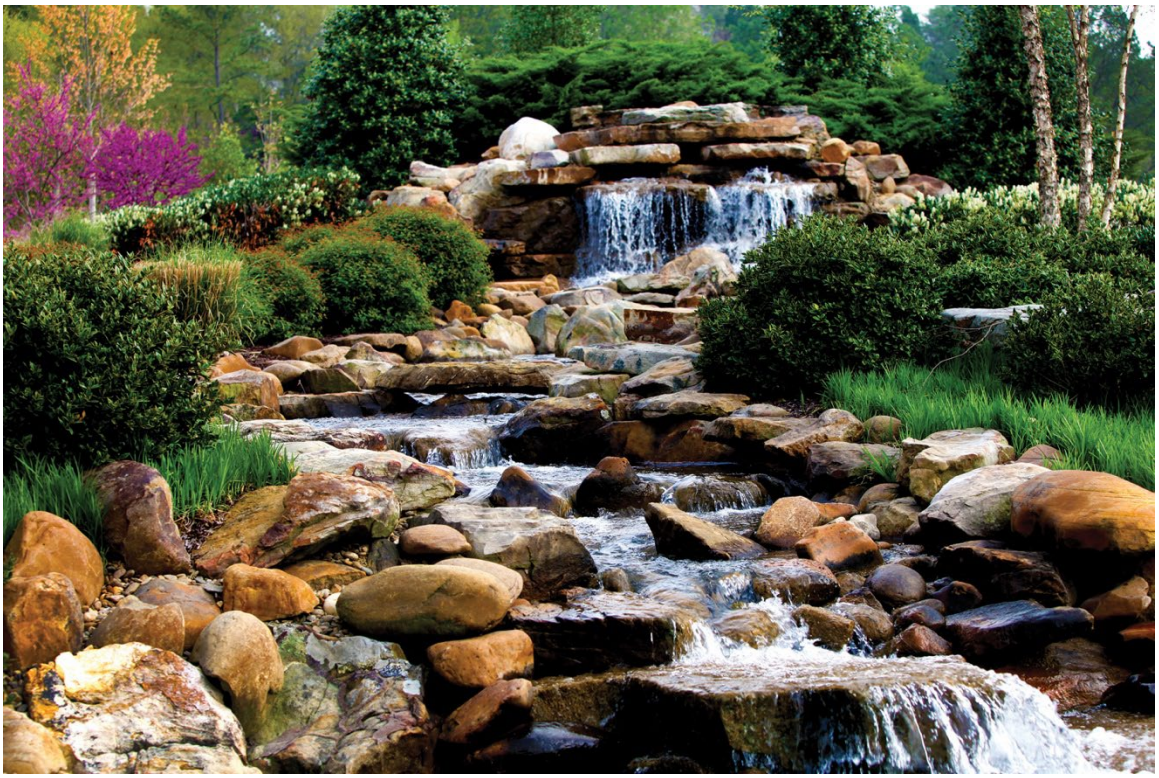
ORNL/SPR-2022/2400
PUB ID 174593

CTSA MHRI Datasets Curation Documentation



Christian Blair
Hilda B. Klasky
Kevin Sparks
Alina Peluso
Deeksha Rastogi
Joe Tuccillo
Hong-Jun Yoon
Rochelle Watson

March 2022

OAK RIDGE National Laboratory

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

ORNL/SPR-2022/2400
PUB ID 174593

Computational Sciences & Engineering Division

CTSA MHRI DATASETS CURATION DOCUMENTATION

Christian Blair
Hilda B. Klasky
Kevin Sparks
Alina Peluso
Deeksha Rastogi
Joe Tuccillo
Hong-Jun Yoon
Rochelle Watson

March 2022

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

TABLES

1. INTRODUCTION

Oak Ridge National Laboratory (ORNL) has collaborated with MedStar Health Research Institute (MHRI) to develop, test, and validate health outcomes using electronic health records (EHRs) from hospitals associated with participating Clinical and Translational Science Awards (CTSA). MHRI, the research organization of MedStar Health (MSH), has a history of initiating projects, both in the laboratory and in the field, that serve the needs of medically underserved and disenfranchised groups.

Health outcomes are typically modeled as a function of genetics and environment, where environment refers to air quality, access to transportation and food, homelessness status, etc. Individual health outcomes are considered to be associated with multiple stressors that fall under a variety of categories – socioeconomic, economic, physical environment. Understanding the relationships between these stressors, covariates, and health outcomes require curated, standardized data that can be input into the health outcomes model. ORNL's unique human dynamics datasets along with MHRI's EHR create a robust infrastructure to explain variation in health outcomes in Washington, DC.

ORNL and MHRI are part of the Georgetown-Howard Universities Center for Clinical and Translational Science (GHUCCTS). The formation of GHUCCTS was inspired by the National Center for Advancing Translational Sciences (NCATS) CTSA initiative to improve how academic medical centers perform clinical research and translational science. This initiative empowers researchers to develop new treatments and cures for diseases more quickly and make them available to the patients and communities more efficiently. GHUCCTS was funded by NCATS on July 1, 2010, thereby becoming one of just 64 CTSA-funded centers in the United States, and the five-year program has been renewed twice, most recently in 2020.

In this document, the ORNL team is providing data curation documentation for the following datasets, which are publicly available:

1. Area Deprivation Index (ADI) 2015 block group
2. Child Opportunity Index 2015 tract
3. Low food access 2017 block group
4. Neighborhood deprivation index 2017 tract
5. Social Capital Index 2014 county
6. Social Vulnerability Index 2014 tract

2. ADI 2015 BLOCKGROUP

The ORNL team is providing the 2015 Block Group ADI Files version 3.0. These files contain a linkage between the Census block group and the ADI score for the following states: DC, MD, PA, and VA.

The files contain four relevant fields, as follows:

1. GISJOIN: Key linkage field to the block group shapefile served by NHGIS
2. FIPS: The block group Census ID
3. ADI_NATRANK: National percentile of block group ADI score
4. ADI_STATERNK: State-specific decile of block group ADI score

NOTE: DO NOT USE EXCEL TO OPEN THE FILES! - Excel has a record limit of roughly 1.4 million and does not warn you if you exceed the limit. Truncation of the data may occur!

Upon download of the block group ADI files, please verify the record number count in your data against the chksum record number below:

Table 1. ADI 2015 Block-group Checksum Records for DC, MD, PA, and VA.

State_full	#_Records	File_size_KB
Washington D.C.	450	19
Maryland	3926	169
Pennsylvania	9740	430
Virginia	5332	235

3. CHILD OPPORTUNITY INDEX 2015 TRACT

The child opportunity index 2015 tract data set source is the following:

https://data.diversitydatakids.org/dataset?vocab_Topic=Child%20Opportunity%20Index&_ga=2.1221029 5.275131148.1644261846-649572742.1644261846.

Table 2 presents the child opportunity index 2015 tract dataset' data dictionary.

Table 2. Child Opportunity Index 2015 Tract Data Dictionary.

Column	Type	Label	Description
geoid	text	2010 census tract FIPS codes	
year	numeric	Year of observation (2010 or 2015)	
in100	numeric	In one of the 100 largest metro areas	This variable is equal to 1 if the census tract is located in one of the 100 largest metro areas (as of 2015), zero if located in a metro- or micropolitan area that does not belong to the 100 largest metro areas, and blank if it is located outside a metro- or micropolitan area.
msaid15	text	Metro/Micro Area (2015) FIPS Code	
msaname15	text	Metro/Micro Area (2015) Name	
countyfips	text	County FIPS Code	
statefips	text	State FIPS Code	
stateusps	text	State USPS Code	
pop	numeric	Number of children aged 0-17, ACS	

		(2008-12 for 2010, 2013-17 for 2015)	
z_ED_nat	numeric	Z-scores, education domain, nationally-normed	Weighted average of education domain component indicator z-scores, nationally normed.
z_HE_nat	numeric	Z-scores, health and environment domain, nationally-normed	Weighted average of health and environment domain component indicator z-scores, nationally normed.
z_SE_nat	numeric	Z-scores, social and economic domain, nationally-normed	Weighted average of social and economic domain component indicator z-scores, nationally normed.
z_COI_nat	numeric	Z-scores, overall COI, nationally-normed	Weighted average of three domain averaged z-scores (z_ED_nat, z_HE_nat, z_SE_nat), nationally normed.
c5_ED_nat	text	Child Opportunity Levels, education domain, nationally-normed	Nationally-normed Child Opportunity Levels (from "very low" to "very high") for the education domain.
c5_HE_nat	text	Child Opportunity Levels, health and environment domain, nationally-normed	Nationally-normed Child Opportunity Levels (from "very low" to "very high") for the health and environment domain.
c5_SE_nat	text	Child Opportunity Levels, social and economic domain, nationally-normed	Nationally-normed Child Opportunity Levels (from "very low" to "very high") for the social and economic domain.
c5_COI_nat	text	Child Opportunity Levels, overall COI, nationally-normed	Nationally-normed Child Opportunity Levels (from "very low" to "very high") for the overall COI.
r_ED_nat	numeric	Child Opportunity Scores, education domain, nationally-normed	Nationally-normed Child Opportunity Scores (from 1 to 100) for the education domain.
r_HE_nat	numeric	Child Opportunity Scores, health and environment domain, nationally-normed	Nationally-normed Child Opportunity Scores (from 1 to 100) for the health and environment domain.
r_SE_nat	numeric	Child Opportunity Scores, social and economic domain, nationally-normed	Nationally-normed Child Opportunity Scores (from 1 to 100) for the social and economic domain.

r_COI_nat	numeric	Child Opportunity Scores, overall COI, nationally-normed	Nationally-normed Child Opportunity Scores (from 1 to 100) for the overall COI.
c5_ED_stt	text	Child Opportunity Levels, education domain, state-normed	State-normed Child Opportunity Levels (from "very low" to "very high") for the education domain.
c5_HE_stt	text	Child Opportunity Levels, health and environment domain, state-normed	State-normed Child Opportunity Levels (from "very low" to "very high") for the health and environment domain.
c5_SE_stt	text	Child Opportunity Levels, social and economic domain, state-normed	State-normed Child Opportunity Levels (from "very low" to "very high") for the social and economic domain.
c5_COI_stt	text	Child Opportunity Levels, overall COI, state-normed	State-normed Child Opportunity Levels (from "very low" to "very high") for the overall COI.
r_ED_stt	numeric	Child Opportunity Scores, education domain, state-normed	State-normed Child Opportunity Scores (from 1 to 100) for the education domain.
r_HE_stt	numeric	Child Opportunity Scores, health and environment domain, state-normed	State-normed Child Opportunity Scores (from 1 to 100) for the health and environment domain.
r_SE_stt	numeric	Child Opportunity Scores, social and economic domain, state-normed	State-normed Child Opportunity Scores (from 1 to 100) for the social and economic domain.
r_COI_stt	numeric	Child Opportunity Scores, overall COI, state-normed	State-normed Child Opportunity Scores (from 1 to 100) for the overall COI.
c5_ED_met	text	Child Opportunity Levels, education domain, metro-normed	Metro-normed Child Opportunity Levels (from "very low" to "very high") for the education domain.
c5_HE_met	text	Child Opportunity Levels, health and environment domain, metro-normed	Metro-normed Child Opportunity Levels (from "very low" to "very high") for the health and environment domain.
c5_SE_met	text	Child Opportunity Levels, social and economic domain, metro-normed	Metro-normed Child Opportunity Levels (from "very low" to "very high") for the social and economic domain.
c5_COI_met	text	Child Opportunity Levels, overall COI, metro-normed	Metro-normed Child Opportunity Levels (from "very low" to "very high") for the overall COI.

r_ED_met	numeric	Child Opportunity Scores, education domain, metro-normed	Metro-normed Child Opportunity Scores (from 1 to 100) for the education domain.
r_HE_met	numeric	Child Opportunity Scores, health and environment domain, metro-normed	Metro-normed Child Opportunity Scores (from 1 to 100) for the health and environment domain.
r_SE_met	numeric	Child Opportunity Scores, social and economic domain, metro-normed	Metro-normed Child Opportunity Scores (from 1 to 100) for the social and economic domain.
r_COI_met	numeric	Child Opportunity Scores, overall COI, metro-normed	Metro-normed Child Opportunity Scores (from 1 to 100) for the overall COI.

4. LOW FOOD ACCESS 2017 BLOCK GROUP

The Low Food Access 2017 Block Group dataset source is the following: https://opendata.dc.gov/datasets/DCGIS::low-food-access-areas/explore?location=38.890908%2C-77.026467%2C12.47

The District of Columbia's low food access neighborhoods are assessed to be more than a 10-minute walk from the nearest full-service grocery shop.

We provide the following data, for more information see: https://opendata.dc.gov/datasets/DCGIS::low-food-access-areas/about:

1. FIPS: Federal Information Processing System code numbers which uniquely identify geographic areas.
2. WARD : Polygon
3. PARTPOP2: The total population estimated to live within the low food access area polygon (derived from Census tract population, assuming even distribution across the polygon after removing non-residential areas, followed by the removal of population living within a grocery store radius.)
4. PRTOVR185: The portion of PartPop2 which is estimated to have household income above 185% of the federal poverty line (the food secure population)
5. PRTUND185: The portion of PartPop2 which is estimated to have household income below 185% of the federal poverty line (the food insecure population)
6. PERCENTUND185: A calculated field showing PrtUnd185 as a percent of PartPop2. This is the percent of the population in the polygon which is food insecure (both living in a low food access area and below 185% of the federal poverty line).

5. NEIGHBORHOOD DEPRIVATION INDEX 2017 TRACT

The Neighborhood Deprivation Index 2017 Tract dataset source is the following:

https://www.gis.cancer.gov/research/files.html#soc-dep

5.1 CODEBOOK – NEIGHBORHOOD DEPRIVATION INDEX DATA

A Neighborhood Deprivation Index (NDI) for each Census tract in the U.S. was created using factor analysis to identify key variables from 13 measures in the following dimensions of socioeconomic (SES) status: wealth and income, education, occupation, and housing conditions.[1] These 13 variables were obtained from the Census Bureau's 5-year American Community Survey (ACS) data for 2013-2017.

Neighborhood Deprivation Index data tables are available in both Excel and comma delimited text (CSV) file formats. There is a row for each census tract in the 50 U.S. states and the District of Columbia. In addition to the NDI value, we include the original 13 socioeconomic variables to allow researchers to generate NDI values for their particular study are if desired. Data tables include the following variables for each tract (see Table 3).

Table 3. Neighborhood Deprivation Index (NDI) Metadata

Variable	Format	Description
TractID	Char 11	The fully qualified census tract ID. Includes the state FIPS code (2 chars), the county FIPS code (3 chars) and the tract ID (6 characters). The tract ID has an implied decimal before the last two characters. For example "010102" is referred to in Census tables and descriptions as tract 101.02.
StCoFIPS	Char 5	The state and county FIPS code. Useful for selecting data for a particular county or set of counties.
StAbbr	Char 2	The alphabetic state postal abbreviation. Useful for selecting data for a particular state or set of states.
NDI	Numeric	The Neighborhood Deprivation Index computed using data from all U.S. census tracts. Values range from -2.5 to +1.9. Higher values indicate more neighborhood deprivation (lower socioeconomic status)
NDIQuint	Char 24	Quintiles for the Neighborhood Deprivation Index. Possible values are: "1-Least deprivation": the tract is in the first NDI quintile"2-BelowAvg deprivation": the tract is in the second NDI quintile "3-Average deprivation": the tract is in the third NDI quintile"4-AboveAvg deprivation": the tract is in the fourth NDI quintile "5-Most deprivation": the tract is in the highest NDI quintile "9-NDI not avail": the NDI value is missing for this tract
MedHHInc	Numeric	Median household income (dollars) SES dimension: wealth and income ACS table source: B19013

[1] Diez Roux A V, Mair C, Roux AVD, Mair C, Diez Roux A V, Mair C. Neighborhoods and health. *Ann N Y Acad Sci* 2010; 1186: 125–45.

PctRecvIDR	Numeric	Percent of households receiving dividends, interest, or rental income SES dimension: wealth and income ACS table source: B19054
PctPubAsst	Numeric	Percent of households receiving public assistance SES dimension: wealth and income ACS table source: B19058
MedHomeVal	Numeric	Median home value (dollars) SES dimension: wealth and income ACS table source: B25077
PctMgmtBusSciArt	Numeric	Percent in a management, business, science, or arts occupation SES dimension: occupationACS table source: C24060
PctFemHeadKids	Numeric	Percent of households that are female headed with any children under 18 SES dimension: housing conditions ACS table source: B11005
PctOwnerOcc	Numeric	Percent of housing units that are owner occupied SES dimension: housing conditions ACS table source: DP04
PctNoPhone	Numeric	Percent of households without a telephone SES dimension: housing conditions ACS table source: DP04
PctNComPlmb	Numeric	Percent of households without complete plumbing facilities SES dimension: housing conditions ACS table source: DP04
PctEducHSPlus	Numeric	Percent with a high school degree or higher SES dimension: education ACS table source: S1501
PctEducBchPlus	Numeric	Percent with a college degree or higher SES dimension: education ACS table source: S1501
PctFamBelowPov	Numeric	Percent of families with incomes below the poverty level SES dimension: wealth and income ACS table source: S1702
PctUnempl	Numeric	Percent unemployed SES dimension: occupation ACS table source: S2301

Vintage 2017 tract IDs are used. These including all tract and county coding changes made through 2017. For details, see https://www.census.gov/programs-surveys/acs/technical-documentation/table- and-geography-changes.html.

5.2 METHODS – NEIGHBORHOOD DEPRIVATION INDEX DATA

A Neighborhood Deprivation Index (NDI) for each Census tract in the U.S. was created using factor analysis to identify key variables from 13 measures in the following dimensions of socioeconomic (SES) status: wealth and income, education, occupation, and housing conditions.[21] The specific 13 measures are:

- Wealth and income
    - Median household income (dollars)
    - Percent of households receiving dividends, interest, or rental income
    - Percent of households receiving public assistance
    - Median home value (dollars)
    - Percent of families with incomes below the poverty level
- Education
    - Percent with a high school degree or higher
    - Percent with a college degree or higher
- Occupation
    - Percent in a management, business, science, or arts occupation
    - Percent unemployed
- Housing conditions
    - Percent of households that are female headed with any children under 18
    - Percent of housing units that are owner occupied
    - Percent of households without a telephone
    - Percent of households without complete plumbing facilities

These 13 variables were obtained from the Census Bureau's 5-year American Community Survey (ACS) data for 2013-2017. Factor analysis was then used to generate the NDI. This involved the following steps:

1. Log transform median household income and median home value
2. Reverse code percentages so that higher values represent more deprivation. For example, the percent of housing units that are owner occupied was converted to the percent of housing units that are not owner occupied.
3. Z-standardize the percentages
4. Run a factor analysis using Promax (oblique) rotation and a minimum Eigenvalue of 1
5. Calculate the factors using only variables with a loading score > 0.4 for the first factor **(this removed three variables: the percent of housing units that are owner occupied, the percent ofhouseholds without a telephone, and the percent of households without complete plumbing facilities)**
6. Calculate Cronbach's alpha correlation coefficient among the factors and verify values are above0.7.
7. Use the resulting calculation of the first factor as the Neighborhood Deprivation Index (NDI)

The final NDI calculation is based on 10 of the 13 variables included in the factor analysis. Those variables are median household income; percent of household receiving dividends interest or rental income; percent of households receiving public assistance; median home value; percent of families with incomes below the poverty level; percent with a high school degree or higher; percent with a college degree or higher; percent in a management, business, science, or arts occupation; percent unemployed;and

[2] Diez Roux A V, Mair C, Roux AVD, Mair C, Diez Roux A V, Mair C. Neighborhoods and health. *Annals of the NewYork Academy of Sciences* 2010; 1186: 125–45.

percent of households that are female headed with any children under 18. The percent of housing units that are owner occupied, the percent of households without a telephone, and the percent of households without complete plumbing facilities loaded poorly during the factor analysis and thus, are excluded**.** NDI values range from -2.5 to +1.9. Higher values indicate more neighborhood deprivation (lower socioeconomic status). We also created a categorical variable representing NDI quintiles weighted by tract population (so that 20% of the population is in each quintile group).

The NDI is based on factor analysis of tract-level variables at the national level. For research involving a particular study area, there might not be as much variation in deprivation levels as there in nationally. Also, variation in one or more of the variables may be more pronounced regionally or locally than nationally. For example, in an area with a high cost of living, regional or local variations in household income and home value may be more significant than they are at a national level. Researchers could generate a specific index for their study area (or a larger area that includes the study area) to characterize regional or local variation in deprivation. The original 13 variables are included in the dataset for this purpose. Working with a subset of the national data, researchers could generate a custom version of the NDI using the steps described above. The generated index might include a different subset of the original variables and, thus, might emphasize slightly different aspects of neighborhood deprivation.

6. SCI 2014 COUNTY

ORNL produced a Social Capital Index for 2014 (county) based on [Rupasingha at el 2006 (https://www.sciencedirect.com/science/article/abs/pii/S1053535705000971)] (https://www.sciencedirect.com/science/article/abs/pii/S1053535705000971)), and an update to the Social Capital Index for the years 1997, 2005, 2009, and 2014. Social capital has had a powerful impact on the study of politics, policy, and social science at large. While the concept of social capital is valid universally, the measure of social capital varies by context. Much of what we know about the causes and effects of social capital, however, is limited by the nature of data used regularly by scholars working in this area. Principal Component Analysis is used to extract principal components from data variables and create a signal index that indicates the social capital. Data are used that represents relevant establishments, voter turnout, census response rates, and non-profit organizations. The implementation presented various challenges including missing and suppressed data and changing county names.

For more information on the Social Capital Index, the inspiration data products are archived at: (https://aese.psu.edu/nercrd/community/social-capital-resources).

The primary data sources used in the calculations for 2014 are as follows:

1) Establishments: County Business Patterns (https://www.census.gov/programs-surveys/cbp.html)
2) Population: US Census, Population and Housing Unit Estimates (https://www.census.gov/programs-surveys/popest.html)
3) Voter Turnout: MIT Election Lab (https://electionlab.mit.edu/)
4) Census response rate: US Census 2020 (https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-main.html)
5) Non-profit: National Center for Charitable Statistics (http://www.nccs.urban.org/)

Table 4. Social Capital Index Data Dictionary

variable name	variable label
FIPS	County FIPS code
sci	Social Capital index for 2019
civic	Number of establishments in civic and social associations
bowling	Number of establishments in bowling center
fitness	Number of establishments in fitness and recreational sports centers
golf	Number of establishments in golf courses and country clubs
religion	Number of establishments in religious organizations
sport	Number of establishments in sports teams and clubs
business	Number of establishments in business associations
political	Number of establishments in political organizations
professional	Number of establishments in professional organizations
labor	Number of establishments in labor organization
associations	Average of all 10 above variables divided by population per 10,000 (1st factor)
vote	Voter turnout (2nd factor)
response	Response rate (3rd factor)
nccs	Number of non-profit organizations divided by population per 10,000 (4th factor)
population	Population estimate
Putnam	Average of civic, bowling, fitness, golf, religion, and sport, divided by population per 10,000
Olson	Average of business, political, professional, and labor, divided by population per 10,000

7. SVI 2014 TRACT

Social Vulnerability Index (SVI) 2014 Tract dataset's documentation is found at https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2014.html

The ORNL team is providing the raw data as downloaded from the source.

8. ACKNOWLEDGEMENTS

Research reported in this document was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1-TR001409. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.