# Supplementary materials

## Supplement 1. Governance and Metadata

### Supplement 1a. Metadata Standard

Data governance refers to the overall management and control of an organization's data assets, encompassing the policies, procedures, and frameworks that ensure data integrity, privacy, security, and compliance throughout its lifecycle. There are numerous data governance considerations that go beyond traditional informatics questions of patient reidentification and data standardization. The SDC establishes standards for dataset quality, dataset inclusions, metadata annotation, and data access.

|  |
| --- |
| **Attribute: DataCategory**A classification of original into broad categories of mechanisms. e.g., neighborhood economic activity v.s. environmental readings. This attribute is populated with the following values:* **EnvironmentalData**: Data which refer to measurements of environmental contaminants/conditions such as air pollution, light, noise, and temperature.
* **DemographicData**: Data which refer to commonly used demographic variables such as race/ethnicity, income, educational level, and household make up.
* **PropertyData**: Data which refers to condition, values, or age of individual properties in a region.
* **SafetyData**: Data which refers to observations about crime, safety, and other related risk factors in a region.
* **AccessData**: Data which refers to observations about proximity and access to certain neighborhood resources.
* **EconomicActivityData**: Data which refers to measurements of economic activity in a certain region.

**Attribute: DataType**A classification of original data format which has implications on how the data can be stored, processed, and interpreted. In essence, it defines the main population unit of interest for this data This attribute is populated with the following values:* **IndividualData**: Data that are linked to an individual patient.
* **OrganizationData**: Data that are linked to associations, institutions, agencies, businesses, political parties, schools, etc..
* **HouseholdData**: Data that are linked to a person or a group of persons who share the same dwelling unit and common living arrangements.
* **EventData**: Data that are linked to any type of incident, occurrence, or activity. Events are linked to GeographicUnits
* **LocationData**: Data that are linked to a some GeographicalUnit.

**Attribute: GeographicUnit**Any entity that can be spatially defined as a geographic area, with either natural (physical) or administrative boundaries. This attribute is populated with the following values:* **NoneUnit**: Data that are not linked to a particular geographic unit, e.g., tied to the individual/household.
* **HousingUnit**: Data that are linked to a housing unit which is a house, an apartment, a mobile home, a group of rooms, or a single room that is occupied (or if vacant, is intended for occupancy) as separate living quarters.
* **LatLongUnit**: Data that are linked to a specific latitude and longitude.
* **CensusBlockUnit**: Data that are linked to census blocks. A census block is the smallest geographic unit used by the United States Census Bureau for tabulation of 100-percent data.
* **CensusTractUnit**: Data that are linked to census tracts. In the U.S., census tracts are designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions and average about 4,000 inhabitants
* **ZipCodeUnit**: Data that are linked to zipcodes. A ZIP Code is a postal code used by the United States Postal Service (USPS).
* **OtherNeighborhoodUnit**: Data that are linked to some well-defined boundary of neighborhood but does not fall into one of the above categories.

**Attribute: TimeUnit**Any period of time: year, week, month, day, or bimonthly or quarterly periods, etc. This attribute is populated with the following values:* **NoneUnit**: Observation is not tied to a particular time instant or period.
* **InstantUnit**: Capture one-time, individual occurrences, with a limited, or short duration.
* **MonthlyUnit**: Capture on going processes with a one-month update cycle.
* **QuarterlyUnit**: Capture on going processes with a one-quarter update cycle.
* **YearlyUnit**: Capture on going processes with a one-year update cycle.

**Attribute: CollectingOrganization**The name of the original organization that collected the data. The data may have been subsequently processed, but this captures the original data source. This attribute is a string field.**Attribute: ProcessingCode**A link to a computational notebook that converts the source data into the final format. This includes cleaning, reformatting, and aggregating. This attribute is a url field.**Attribute: TermsOfUse**A description of any stipulations of how this data can be used This attribute is a string field.**Attribute: WhenAccessed**The date at which the artifact was accessed This attribute is a date field.**Attribute: WhenCollected**The date at which the artifact was collected.**Attribute: ProcessingOrganization**The organization that processed the data into its current form. This attribute is a string field.**Attribute: Missingness**A description of any incompleteness in this dataset. This attribute is populated with the following values:* **InRecordMissing**: The collected records have known missing cell values marked by a common indicator such as None, NaN, etc.
* **CompletelyMissing**: The collected records are known to be incomplete where known units are missing from the dataset.

**Attribute: Incorrectness (Can have many values)**A description of any incorrect values in this dataset. This attribute is populated with the following values:* **SensorFailure**: The collected records have apparently incorrect values due to failures in the collection device
* **ResponseBias**: The collected records have apparently incorrect values due to survey subjects misrepresenting (either intentionally or unintentionally) the data
* **EstimationError**: The collected records have apparently incorrect values due to estimation
* **CodingError**: The collected records have apparently incorrect values due to inconsistent data coding
 |

### Supplement 1b. Data Quality Scores

.Data quality scores are essential for a data repository as they serve as critical indicators of the reliability and usability of the stored information. These scores provide a quantifiable measure of the overall quality of the data, helping users and stakeholders assess the trustworthiness of the repository's contents. By evaluating various aspects such as accuracy, completeness, consistency, and timeliness, data quality scores assist in identifying potential issues, errors, or inconsistencies in the dataset. This information empowers data consumers to make informed decisions and select appropriate data for analysis, research, or decision-making purposes. Additionally, data quality scores enable data custodians to prioritize data cleansing and enhancement efforts, ensuring the repository's content remains up-to-date and valuable. Ultimately, by incorporating data quality scores, a data repository can enhance its overall credibility, encourage data sharing and collaboration, and foster the development of reliable insights and knowledge based on trustworthy data.

When developing such a metric, it is important to accurately taxonomize the types of errors manifest in such data. We leverage established prior work from the database community to develop this taxonomy.

**Syntactic Error.** A syntactic error is a data value that does not match the expected format of its data attribute. For example, a particular address might be formatted differently than the others. Or, a value might be outright missing and represented by some sentinel symbol indicating its missingness. Syntactic errors are characterized by their effects on downstream analysis code -- the errors need special handling since they do not match the expected format. Syntactic errors are "self-evident" in the sense that no outside information is needed to determine their presence.

**Semantic Error.** A semantic error is a factual inaccuracy or implausibility in the data. For example, a traffic dataset may indicate a speed limit on a particular road of 29 miles per hour. While syntactically correct (the speed is properly a number), it is semantically incorrect (the 29 mile per hour speed limit is likely wrong). In a sense, semantic errors are more insidious that syntactic errors as they might cause "silent" failures, where the analysis code works but the conclusions are biased. Semantic errors require outside knowledge to identify and rectify.

**Systemic Error.** A systemic error is when the choice of data representation or measurement methodology is systematically biased or incomplete. While almost all datasets are biased in one way or another, we define a systemic error as a dataset that ignores "a significant confounding variable". In such cases, it would be much better to store and measure the confounding variable itself. As a principle, the SDC tries to store the most direct measurement of exposures.

**Data Quality Score**

Based on this taxonomy, we can create a framework to score datasets in the repository. The scoring is further based on our judgment of the statistical implications of these errors, i.e., whether they can be safely ignored during analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5 (best)** |
| Substantial systemic errors present in the dataset.  | Syntactic and/or semantic errors are present and might be correlated with covariates.  | Minor syntactic and/or semantic errors in the dataset that can be ignored during analysis. | Minor syntactic errors in the dataset that can be ignored during analysis. | No known errors in the dataset |

In some cases, data errors are correlated with covariates of interest such as neighborhoods, demographics, or other latent factors. For example, the analysis below shows how missing values are distributed through a dataset of business listings in the city of Chicago. We see that missing values in the dataset are unevenly distributed. The figure below colorizes census tracts by the number of complete business records. The plot is formatted as follows: 0.0 marks the region average per census tract, 0.3 marks 30% above average completeness, -0.3 marks 30% below average. The more brown the tract is, the less “clean” the data is.

****

### Supplement 1c. Data Sustainability

Sustainability refers to the ability to maintain or support a process continuously over time. An objective for the SDC is what we term “data sustainability”, namely, that the software infrastructure and organizational processes that govern it will persist beyond initial pilot studies – maintaining both its ease of use and accuracy. The goal is to develop a community of researchers who have a vested interest in maintaining and supporting the SDC. Through experience in prior projects, we have identified a number of key threats to sustainability:

1. Dataset Decay. Sociome factors are often highly dynamic and evolve due to social, economic, or environmental changes. Datasets that are not refreshed periodically leading to stale data and inaccurate conclusions.
2. Utility Decay. There is also the concern that the data repository is overly tailored to a few Sociome use-cases. Long-term interest in the repository can wane if it does not contain enough useful datasets beyond a few pilot projects.
3. Documentation Decay. The loss of code or domain knowledge means that certain datasets cannot be easily reused in the future.
4. Few Experts. If only a small number of researchers know how to use the data in the commons, the interest in maintaining the commons is tied to their tenure in the project.

Our strategy for long-term sustainability involves two key insights. First, we open-source all of our data harmonization code in a Python library. This code allows independent researchers to quickly contribute new datasets and ideas into the SDC in an interoperable format. From the beginning, we have envisioned the SDC as less of a walled garden of curated datasets and more of a “wiki for datasets” that are maintained by a community of researchers within and outside our organizations. Our governance structure ensures that such data are always properly annotated with metadata and their quality explained to other researchers.

## Supplement 2: SDC Software Details

The SDC is a cloud-based data repository with a variety of datasets describing social, environmental, behavioral, and psychological exposures/circumstances connected with locations. Each dataset is stored in a geocoded format, called GeoJSON, where measurements are associated with latitude and longitude regions. This format is supported by all major GIS frameworks. The SDC makes it easy to run *integrative queries* across the datasets, i.e., to identify all of the exposures associated with a particular geographical unit ranging from an individual location to an entire zip code.

These exposure profiles can be pulled into a protected enclave where these geocoded non-clinical data can be joined to clinical data (be that limited data sets or full PHI data). Each dataset is documented with metadata describing its scope, quality, and units of measure. The repository is searchable and datasets can be directly accessed by researchers through a programmatic interface. Researchers can identify the types of exposures they wish to investigate and easily build an integrated profile for a certain region.

The SDC will store and harmonize social context of disease data at its finest granularity available, i.e., at a housing parcel level or at an individual level. It will further avoid storing derived indices and focus on storing primary measurements of Sociome factors. The authors of this paper believe that aggregated Sociome indices, while designed to absorb as much variance as possible, can hide meaningful variations of individual factors among patient subpopulations and thereby obfuscate potential causal mechanisms.

## Supplement 3: Dataset search protocol

### Finding relevant datasets

#### Purpose

Datasets include any aspect of the social context in the United States. We aim to cover traditional SDOH but aim to expand beyond traditional SDOH into the full social context of disease.

#### Inclusion/exclusion criteria

|  |  |  |
| --- | --- | --- |
|  | Inclusion | Exclusion |
| Content | Any aspect of the social context | No restriction on generalizable content |
| Coverage area | The whole United States OR the Chicago metropolitan area (Cook County, other collar counties, Skokie, Evanston, etc.) | Does not cover the United States OR covers a subset of the United States which is not the Chicago metropolitan area (e.g., covers another metropolitan area like Detroit, MI or Boulder, CO) |
| Age | At least as current as 2010 for survey data, but unlimited for pollution or other environmental exposure data | Older than 2010 for survey data only |
| Geographic level | Minimum geographic level is the census tract. Acceptable levels:* Latitude/longitude (aka “point”)
* Census block
* Census tract
 | Geographic level is larger than the census tract. Examples include:* Zip code
* Community area
* Health equity zone
* City
* County
* Metropolitan statistical area
* State
* Region of the country (e.g., South, West)
 |

#### Tracking search, inclusion, exclusion

Dataset tracking google sheet

* By
* SearchID
* Date
* Resource (pubmed, google)
* Search term
* Included from this search #
* Excluded from this search #
	+ Does not include the United States
	+ Last updated before 2010
	+ Geographic area too large
	+ Covers US subset besides the Chicago metropolitan area
	+ Vendor selling data services
	+ Other (add reason)
* For google (or other web engine) searches:
	+ No sponsored ads
	+ First *n* pages looked through
	+ Stop when you reach saturation / not learning anything new

## Supplement 4: Geocoding testing summary

From a random sample of 1,000 Cook County addresses, the testing results were:

|  |  |  |
| --- | --- | --- |
| Method | Correctly geocoded | Time |
| DeGAUSS (local installation) | 994 (99.4%) | 4:23 minutes |
| Nominatum via GeoPy | 870 (87%) | 19:05 minutes |
| GoogleV3 via GeoPy | 997 (97%) | 16:67 minutes |
| Census via the website | 972 (97.2%) | n/a |

Testing of Chicago’s City Hall address (121 N La Salle St, Chicago, IL 60602) included 24 deliberate changes to or misspellings of the street number, direction, street name, suffix, city, state, and zip. The results were:

|  |  |  |
| --- | --- | --- |
| Method | Correctly geocoded | Notes |
| DeGAUSS (local installation) | 22/24 (92%) | 1 was missed (no zipcode) and 1 was geocoded to incorrect points (incorrect zip code, “60604” instead of “60602”). |
| Nominatum via GeoPy | 1/24 (4%) | 22 were missed entirely and 1 was geocoded to incorrect points (incorrect direction, “S” instead of “N”). |
| GoogleV3 via GeoPy | 21/24 (88%) | 3 were geocoded incorrectly: Altered direction (“No” instead of “N”), no street direction, and incorrect street direction (“S” instead of “N”). |
| Census via the website | 22/24 (92%) | 2 were geocoded incorrectly: incorrect zipcode (“60602” instead of “60604”), and incorrect direction (“S” instead of “N”). |

## Supplement 5: ACS poverty PCA variables

[ 'pct\_Prs\_Blw\_Pov\_Lev\_ACS', 'pct\_Children\_in\_Pov\_ACS', 'avg\_Agg\_HH\_INC\_ACS', 'pct\_Civ\_unemp\_16p\_ACS', pct\_Pop\_NoCompDevic\_ACS','pct\_Civ\_emp\_25\_44\_ACS', 'Med\_House\_Value\_ACS', 'pct\_Not\_HS\_Grad\_ACS', 'pct\_MrdCple\_HHD\_ACS', 'pct\_Vacant\_Units\_ACS', 'pct\_Pop\_w\_BroadComp\_ACS', 'pct\_Civ\_emp\_16p\_ACS', 'pct\_Civ\_unemp\_25\_44\_ACS', 'pct\_Not\_MrdCple\_HHD\_ACS', 'avg\_Agg\_House\_Value\_ACS', 'pct\_Female\_No\_SP\_ACS', 'pct\_HHD\_NoCompDevic\_ACS', 'pct\_Renter\_Occp\_HU\_ACS', 'pct\_HHD\_w\_Computer\_ACS', 'pct\_NoHealthIns1964\_ACS', 'pct\_Civ\_emp\_45\_64\_ACS', 'pct\_College\_ACS', 'pct\_Civ\_unemp\_45\_64\_ACS', 'pct\_Tot\_Occp\_Units\_ACS', 'pct\_PUB\_ASST\_INC\_ACS', 'pct\_Owner\_Occp\_HU\_ACS', 'pct\_HHD\_No\_Internet\_ACS', 'pct\_HHD\_w\_Broadband\_ACS', 'pct\_HHD\_w\_OnlySPhne\_ACS', 'Med\_HHD\_Inc\_ACS']

## Supplement 6: Sociome analytic pipeline



**Figure**: Sociome analytic pipeline

#

## Supplement 7: Data-driven selection protocols

### Protocol 1: Datasets compared with AIC

|  |  |
| --- | --- |
| Model | AIC (lower is better) |
| Baseline: Clinical only | 49803.62 |
| Spatial clustering | 49576.24 |
| American Community Survey | 44674.91 |
| Building code violations | 49594.28 |
| Chives | 49335.16 |
| Crime  | 49539.90 |
| EPA EJscreen | 49534.28 |
| Housing dataset | 49286.87 |
| **Saturated model (all datasets)** | 44379.12 |

**Figure:** AIC-driven dataset inclusion protocol

Only ACS was a clear dataset addition via the AIC-based inclusion protocol.

### Protocol 2: Lasso-selected features

|  |  |
| --- | --- |
| Dataset | Variable |
| ACS | Poverty PCA |
| ACS | pct\_HHD\_Moved\_in\_ACS |
| ACS | pct\_Schl\_Enroll\_3\_4\_ACS |
| ACS | pct\_Sngl\_Prns\_HHD\_ACS |
| Building violations | Building violation rate per 100 |
| Chives | Cost burdened owners/renters |
| Chives | Population density |
| Clinical | Age at visit |
| Clinical | Visit month |
| Housing | Average age of housing |
| Spatial clustering | 7 clusters |

**Figure:** Data-driven lasso-based variable-level inclusion protocol

 Using lasso as variable selection, the above variables were selected from their respective datasets.

## Supplement 8: Top 20 features

|  |  |  |
| --- | --- | --- |
| All spatial clusters |  | Cluster 1 only |
| Dataset | Variable description (name) | Gain |  | Dataset | Variable description (name) | Gain |
| Spatial clustering | 7 clusters (clus7r.groups) | 28.49 |  | Housing | Average age of housing units (age\_avg) | 27.81 |
| Housing | Average age of housing units (age\_avg) | 21.29 |  | ACS | Percentage of people under age 19 with no health insurance in the ACS (pct\_NoHealthIns\_U19\_ACS) | 26.87 |
| Clinical | Age at visit (Age) | 13.62 |  | Clinical | Visit month (VisitMonth) | 25.62 |
| EPA | Superfund: Count of proposed or listed NPL sites within 5 km (or nearest one beyond 5 km), each divided by distance in kilometers (PNPL) | 13.43 |  | Clinical | Age at visit (Age) | 25.61 |
| Chives | Median (middle value) cost of rent by census tract (median\_rent) | 11.75 |  | ACS | Percentage of people aged 65 and over with no health insurance in the ACS (pct\_NoHealthIns\_65P\_ACS) | 14.37 |
| Crime | Violent crime rate per 1000 persons (ViolRate1000) | 11.11 |  | Chives | Percentage of households paying over 30% of their gross monthly income for housing (cost\_burdened\_owners\_renters) | 14.29 |
| Chives | Proportion of tract which is residential (prop\_resdntl) | 10.3 |  | ACS | Percentage of the ACS population who were not a citizen of the United States at birth. This includes respondents who indicated that they were a US citizen by naturalization or not a US citizen (pct\_Born\_foreign\_ACS) | 11.85 |
| Chives | The urban flood susceptibility index identifies priority areas across the region for flood mitigation activities (urban\_flood\_suscep) | 9.33 |  | Chives | Describes the median (middle value) cost of rent by census tract (median\_rent) | 11.63 |
| Spatial clustering | 3 clusters (clus3r.groups) | 8.81 |  | ACS poverty PCA | Poverty PCA (PCA1) | 11.43 |
| ACS | Percentage of all ACS occupied housing units where a householder lives alone(pct\_Sngl\_Prns\_HHD\_ACS) | 7.73 |  | ACS | Percentage of the ACS population that have no health insurance, public or private (pct\_No\_Health\_Ins\_ACS) | 11.33 |
| ACS | Percentage of ACS occupied housing units that have more than 1.01 persons per room (pct\_Crowd\_Occp\_U\_ACS) | 7.53 |  | ACS | Percentage of ACS occupied housing units that have more than 1.01 persons per room (pct\_Crowd\_Occp\_U\_ACS) | 11 |
| ACS | Percentage of family-occupied housing units with a related child under 6 years old (pct\_Rel\_Under\_6\_ACS) | 7.32 |  | ACS | Percentage of the ACS population that have one type of health insurance coverage, including public or private (pct\_One\_Health\_Ins\_ACS) | 10.95 |
| ACS | Percentage of the ACS population that is male (pct\_Males\_ACS) | 7.19 |  | ACS | Percentage of all ACS occupied housing units where one or more people are ages 18 years or under (pct\_HHD\_PPL\_Und\_18\_ACS) | 9.35 |
| Clinical | Visit month (VisitMonth) | 7.13 |  | ACS | Percentage of all ACS households that contain amarried couple with child under 18. Includes married same-sex couples (pct\_MrdCple\_w\_child\_ACS) | 9.24 |
| ACS | Percentage of the ACS population who were not a citizen of the United States at birth. This includes respondents who indicated that they were a US citizen by naturalization or not a US citizen (pct\_Born\_foreign\_ACS) | 6.67 |  | Building violations | Building violation rate per 100 buildings (BldgViolRate100) | 9.18 |
| ACS | Percentage of the ACS population aged 5 years and over that speaks a language other than English at home (pct\_Othr\_Lang\_ACS) | 6.62 |  | ACS | Percentage of all ACS housing units that are considered mobile homes (pct\_Mobile\_Homes\_ACS) | 8.77 |
| Chives | Logged average annual average daily traffic counts by street segment, by census tract (logtraf) | 6.6 |  | Chives | The urban flood susceptibility index identifies priority areas across the region for flood mitigation activities (urban\_flood\_suscep) | 8.73 |
| ACS | Percentage of all ACS housing units that are considered mobile homes (pct\_Mobile\_Homes\_ACS) | 5.59 |  | ACS | Percentage of the ACS population that is between 18 and 24 years old (pct\_Pop\_18\_24\_ACS) | 7.99 |
| Chives | Population density, the number of people per square mile (pop\_density (sq.mi.)) | 5.51 |  | EPA EJScreen | Proximity to Risk Management Plan (RMP) facilities; these facilities use extremely hazardous substances and are required to develop a RMP by the EPA (PRMP)Count of RMP (potential chemical accident management plan) facilities within 5 km (or nearest one beyond 5 km), each divided by distance in kilometers | 7.88 |
| EPA | Traffic proximity and volume (PTRAF) | 5.26 |  | ACS | The percentage of ACS civilians between the ages of 16 and 24 in the labor force that are unemployed(pct\_Civ\_unemp\_16\_24\_ACS) | 7.78 |

#