

Research Methods and Technology Research Article

Cite this article: Tilmon S, Nyenhuis S, Solomonides A, Barbarioli B, Bhargava A, Birz S, Bouzein K, Cardenas C, Carlson B, Cohen E, Dillon E, Furner B, Huang Z, Johnson J, Krishnan N, Lazenby K, Li K, Makhni S, Miller D, Ozik J, Santos C, Sleiman M, Solway J, Krishnan S, and Volchenboum S. Sociome Data Commons: A scalable and sustainable platform for investigating the full social context and determinants of health. *Journal of Clinical and Translational Science* 7: e255, 1–12. doi: [10.1017/cts.2023.670](https://doi.org/10.1017/cts.2023.670)

Received: 1 August 2023

Revised: 27 September 2023

Accepted: 27 October 2023

Keywords:

Asthma; health disparities; Chicago; data commons; SDOH

Corresponding author:

S. Tilmon, MS, MPH;




Email: stilton@bsd.uchicago.edu

[†]The online version of this article has been updated since original publication. A notice detailing the change has been published at <https://doi.org/10.1017/cts.2024.537>.

© The Author(s), 2023. Published by Cambridge University Press on behalf of The Association for Clinical and Translational Science. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use.



Sociome Data Commons: A scalable and sustainable platform for investigating the full social context and determinants of health[†]

Sandra Tilmon¹ , Sharmilee Nyenhuis^{1,2}, Anthony Solomonides³ , Bruno Barbarioli⁴, Ankur Bhargava⁵, Suzi Birz¹, Kathryn Bouzein¹, Celine Cardenas⁶, Bradley Carlson⁷, Ellen Cohen¹, Emily Dillon⁸, Brian Furner¹, Zhong Huang⁷, Julie Johnson⁹, Nivedha Krishnan¹⁰, Kevin Lazenby⁷ , Kaitlyn Li¹¹, Sonya Makhni⁵, Doriane Miller², Jonathan Ozik¹², Carlos Santos¹³, Marc Sleiman⁷, Julian Solway², Sanjay Krishnan⁴ and Samuel Volchenboum¹

¹Pediatrics, University of Chicago, Chicago, IL, USA; ²Medicine, University of Chicago, Chicago, IL, USA; ³NorthShore University Health System, Research Institute, Evanston, IL, USA; ⁴Computer Science, University of Chicago, Chicago, IL, USA; ⁵Chicago Medicine, Chicago, IL, USA; ⁶Wake Forest University, Winston-Salem, NC, USA; ⁷Pritzker School of Medicine, University of Chicago, Chicago, IL, USA; ⁸Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, IL, USA; ⁹Clinical Research Informatics, University of Chicago, Chicago, IL, USA; ¹⁰University of Illinois at Chicago, Chicago, IL, USA; ¹¹University of Chicago, Chicago, IL, USA; ¹²Decision and Infrastructure Sciences Division, Argonne National Laboratory, Lemont, IL, USA and ¹³Internal Medicine, Rush University Medical Center, Chicago, IL, USA

Abstract

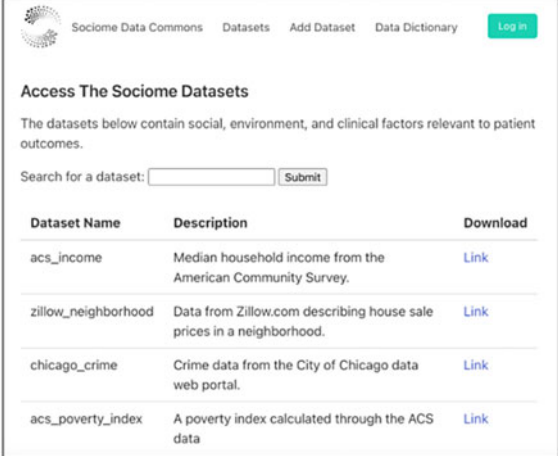
Background/Objective: Non-clinical aspects of life, such as social, environmental, behavioral, psychological, and economic factors, what we call the sociome, play significant roles in shaping patient health and health outcomes. This paper introduces the Sociome Data Commons (SDC), a new research platform that enables large-scale data analysis for investigating such factors. **Methods:** This platform focuses on “hyper-local” data, i.e., at the neighborhood or point level, a geospatial scale of data not adequately considered in existing tools and projects. We enumerate key insights gained regarding data quality standards, data governance, and organizational structure for long-term project sustainability. A pilot use case investigating sociome factors associated with asthma exacerbations in children residing on the South Side of Chicago used machine learning and six SDC datasets. **Results:** The pilot use case reveals one dominant spatial cluster for asthma exacerbations and important roles of housing conditions and cost, proximity to Superfund pollution sites, urban flooding, violent crime, lack of insurance, and a poverty index. **Conclusion:** The SDC has been purposefully designed to support and encourage extension of the platform into new data sets as well as the continued development, refinement, and adoption of standards for dataset quality, dataset inclusion, metadata annotation, and data access/governance. The asthma pilot has served as the first driver use case and demonstrates promise for future investigation into the sociome and clinical outcomes. Additional projects will be selected, in part for their ability to exercise and grow the capacity of the SDC to meet its ambitious goals.

Introduction

Non-clinical aspects of life, such as social, environmental, behavioral, psychological, and economic factors, play significant roles in shaping patient health and health outcomes. These are broadly studied as Social Determinants of Health, which the World Health Organization defines as “conditions in which people are born, grow, live, work and age” and “fundamental drivers of [health] [1].” Including sociome datasets is often a burdensome data problem, both in finding and integrating disparate datasets, where clinical patient data have to be integrated with other data sources to characterize a patient’s life outside of their clinical interactions. We refer to the entirety of these non-clinical or social factors as a patient’s “sociome.” Due to the diversity of data sources and file types that sociome research has to consider, key bottlenecks in scaling such research to large patient populations include data integration [2], data harmonization [3], uneven data quality [4], and statistical modeling of multimodal datasets [5]. Consequently, studies often focus on one factor, a composite index, or a set of highly related factors [6], where potentially crucial nuances and interactions among factors can be lost.

Here, we report on the design and implementation of the Sociome Data Commons (SDC). Leveraging the expertise of the Pediatric Cancer Data Commons in collecting, harmonizing, and

Sociome Data Commons Guiding Principles	
1. Governance	To ensure that researchers can ethically and accurately use the data stored in the SDC to construct a variety of exposure models.
2. Data Sustainability	To ensure that the accuracy and ease-of-use of the SDC does not degrade over time.
3. Dataset Inclusivity	To store, represent, and serve a variety of data types and structures that go beyond traditional health data.
4. Disaggregation	To focus on storing primary measurements of exposures and avoid derived indices.
5. FAIR principles	To ensure that all datasets and code comply with FAIR principles: Findable, Accessible, Interoperable, Reusable.
6. Multidisciplinary Management	To include clinicians, informatics researchers, computer scientists, social scientists, and community members in the research design.



Dataset Name	Description	Download
acs_income	Median household income from the American Community Survey.	Link
zillow_neighborhood	Data from Zillow.com describing house sale prices in a neighborhood.	Link
chicago_crime	Crime data from the City of Chicago data web portal.	Link
acs_poverty_index	A poverty index calculated through the ACS data	Link

Figure 1. Sociome Data Commons (SDC): guiding principles (left), interface showing four datasets for illustration (right).

sharing data [7], we created a repository of pre-harmonized, geospatial sociome datasets that can be used in concert with clinical data to predict a variety of outcomes. To this end, we:

- Assembled and integrated publicly available geocoded datasets about social, environmental, behavioral, psychological, and economic exposures.
- Developed a data governance framework using a structured, standardized metadata model that conforms to FAIR (findable, accessible, interoperable, reusable) [8] principles.
- Established a statistical methodology for analyzing sociome datasets of varying scope and quality, and for scaling and sustaining such analysis over large populations, environments, and diverse data sources.

To evaluate the SDC, we performed a pilot use case to identify sociome factors associated with pediatric asthma exacerbations on the South Side of Chicago. Pediatric asthma was selected as it is a community priority [9], has well-documented social disparities [10], and is known to have many sociome influences, including housing and environmental conditions [11,12]. In addition, clinical factors alone or models with limited variables have lacked sufficient predictive power for asthma outcomes [6,13,14].

Materials and Methods

In this manuscript, we use the following terms as defined below:

- **Sociome Data Commons (SDC):** A cloud-based repository of datasets characterizing a variety of local social, environmental, behavioral, psychological, and economic exposures.
- **Metadata:** Standardized descriptions of the content, quality, ownership, lineage, and scope of a dataset in the SDC.
- **Model:** A statistical or machine learning analysis that associates factors in the SDC to an outcome, often derived from the electronic health record (EHR) or a clinical study.
- **Data Governance:** The overall management and control of the assets in SDC, encompassing the policies, procedures, and frameworks that ensure data quality, accessibility to researchers, ethics, and privacy throughout its lifecycle.
- **Generalizability:** 1. The ability of the SDC to store and serve multiple types of data and multiple types of models.

2. Purposefully, only generalizable data is included in the SDC, whether probability-based survey data, modeled environmental metrics, direct measurements, or surveillance records.

- **Sustainability:** The software infrastructure and organizational processes that govern the SDC will persist beyond initial pilot studies. The SDC is intended to be a persistent platform that can be meaningfully engaged by researchers across disciplines. It is built with the guiding principles described in Fig. 1. Each of the guiding principles is elaborated in the supplement.

Software Implementation

The SDC team has assembled and integrated diverse datasets into a simple, well-documented interoperable format. Researchers can use an application programming interface (API) to access these datasets directly from code or via an interactive website (Fig. 1). The datasets are categorized with a structured, standardized metadata model that conforms to FAIR principles. These sociome datasets can be pulled into a protected enclave where they can be joined to clinical data (protected health information or a limited data set). Deployments behind an appropriate firewall can simplify privacy and security requirements for PHI, while a cloud-based multi-tenant solution could facilitate larger-scale collaborative research projects. Each dataset is documented with metadata describing its scope, quality, and units of measure. The project utilizes a Python toolkit (that will be made available with an open-source license) to aid with common data integration and harmonization steps that researchers using the data might encounter. Researchers can identify the types of sociome factors they wish to investigate and easily build an integrated profile for a certain region.

Governance and Sustainability

The SDC establishes standards for dataset quality, dataset inclusion, metadata annotation, and data access. These standards will help promote trust in the included data and any derived conclusions. Details are included in the supplement. Novel contributions are presented in Table 1.

Table 1. Sociome Data Commons (SDC) standards

Establishment of data documentation standards	Each dataset in the SDC is annotated with a comprehensive data dictionary that adheres to FAIR principles. The structure of this dictionary is derived from the Data Documentation Initiative [58].
Establishment of data quality standards	The project establishes data quality norms to help researchers understand data veracity. Each dataset is assigned a data quality score on a scale of 1 (worst) to 5 (best). The score includes any errors in the dataset as well as missing, malformed, or obvious outlier data using an error taxonomy developed in prior work [59]. Next, sampling biases in the dataset are evaluated to determine how representative the dataset is of the true underlying population (methodology described in detail in the supplement).
Multi-disciplinary research review	The SDC is managed by a multidisciplinary team that includes clinicians, informatics researchers, computer scientists, social scientists, and community members. The SDC is designed to be both a data and code repository where research artifacts can be reviewed by a multidisciplinary team to ensure reliable, reproducible, and ethical research.
Differential access	Where needed, authorization can be tailored to the user's role, allowing certain users access to different data, depending on privacy requirements.

FAIR = findable, accessible, interoperable, reusable.

Asthma Pilot Methodology

We conducted a pilot use case of the SDC for demonstration and to test workflows, beginning with a period of discovery. The pilot investigated sociome factors potentially related to pediatric asthma exacerbations. Clinical data was extracted from the University of Chicago Hospital's EHR for all pediatric visits (age < 18). Data management and analysis of the extracted data occurred mainly in Python [15]. All clinical data (address history, demographics, diagnoses, and encounters) were stored and analyzed on University of Chicago HIPAA-compliant compute and storage infrastructure. Some potentially important clinical data, including allergy testing, asthma control test score, and overweight status, were not available at the time of this pilot. This study was approved by the University of Chicago BSD IRB, #21-1920, and a waiver of consent was granted for this retrospective study.

Geocoding

To adhere to privacy requirements for PHI, an on-site geocoder was preferred. To test and show robustness between available geocoder platforms (both cloud-based and on-premises), a test was performed using 1,000 randomly selected publicly available Chicago addresses [16] as well as systematic misspellings of Chicago's city hall address (a public landmark). We tested batch geocoding with Decentralized Geomarker Assessment for Multi-Site Studies (DeGAUSS, locally-hosted geocoding software) [17] against industry standards: OpenStreetMap, GoogleV3 (both via GeoPy [18]), and the Census geocoder [19]. DeGAUSS performed as well as GoogleV3 and the Census, and all outperformed OpenStreetMap. Details are in the supplement.

Missing Data

The level of missingness in the clinical data was low. Insurance was 5% missing, race/ethnicity 2%, and gender < 1%. Missingness was resolved in two phases. First, by patient and sorted by date, values were filled forward and backward. Second, remaining missingness (2% for race/ethnicity, < 1% for insurance and gender) was resolved with multiple imputations [6,20].

Outcome Definition

Visits for asthma and asthma exacerbations were categorized by the encounter text description including "asthma" and

"exacerbation," respectively. Using text captured 2,010 additional asthma encounters (out of 3.3 million total visits, 2006–2021) than using ICD codes alone. Both terms were required in the visit text to qualify for a visit for an acute asthma exacerbation.

Spatial Clustering

Spatial clustering is based on the assumption that "location matters." Events near each other are often related more than events far apart. Clustering combines geographic areas (here, census tracts) together to maximize *similarity among* the census tracts and maximize *dissimilarity between* the clusters. This clustering is performed to find *meaningful* spatial commonalities. Further, clustering reduces the size of large location datasets. For example, over 300 census tracts on the South Side of Chicago can be reduced to fewer than 10 spatial clusters. Different clusters can have different characteristics, summary descriptions, and, especially of interest here, risk profiles.

Clustering itself is an unsupervised learning problem in that the researcher does not know which areas will be grouped together (though the researcher does need to decide on the *number* of clusters). We used the skater algorithm [21] in R [22], which creates a connectivity graph of tracts' central points as well as edges from tracts' contiguous borders. The "cost" of each edge is established from the dissimilarity between neighboring tracts, and edges with greater costs are pruned. Statistical significance is assessed with Moran's I statistic, which measures spatial autocorrelation (observations at locations that are either contiguous or not) [23].

Spatial clustering was conducted with exacerbation visits as a proportion of all asthma visits and only for tracts with at least 10 visits. The exacerbation percentage was then mean-centered and standardized. Since we did not know how many clusters would be meaningful, we output five different spatial clustering variables. These variables had different *numbers* of cluster assignments – between three and seven – and were designed to be tested in our later model. That is, the model would determine whether three total clusters or up to seven total were more important (described below).

Sociome Data Commons

For SDC datasets, a breadth of high-quality data types were chosen for this pilot study. All data were consistently aggregated at the

Table 2. University of Chicago asthma visits for Chicago pediatric patients, 2017–2019: A. All patients residing in Chicago; B. All patients residing on the South Side of Chicago; and C. All patients residing in spatial cluster 1

	A. Chicago 2017–2019	B. South Side 2017–2019	South Side % of Chicago	C. Cluster 1 2017–2019	Cluster 1 % of South Side
Asthma visits	12,392	11,871	96%	10,501	88%
2017	3,867	3,697	96%	3,232	87%
2018	3,964	3,791	96%	3,369	89%
2019	4,561	4,383	96%	3,900	89%
Outpatient	5,353	5,035	94%	4,258	85%
Emergency	5,554	5,414	97%	4,976	92%
Inpatient	1,485	1,422	96%	1,267	89%
Median age at visit (range)	8 (0–17)	8 (0–17)		7 (0–17)	
Asthma exacerbations	3,602	3,497	97%	3,256	93%
Unique asthma patients	5,826	5,585	96%	4,665	84%
Tracts per patient: Median (range)	1 (1–3)	1 (1–3)		1 (1–3)	
Insurance					
Public: Medicaid	4,511	4,361	97%	3,633	83%
Private	991	913	92%	760	83%
Private with Medicaid	95	90	95%	75	83%
Miscellaneous and unknown	223	215	96%	192	89%
Public: Medicare	6	6	100%	5	83%
Race/ethnicity					
Hispanic	294	271	92%	83	31%
Non-Hispanic					
American Indian or Alaska Native	2	2	100%	2	100%
Asian/Mideast Indian	46	36	78%	33	92%
Black/African-American	5,197	5,034	97%	4,366	87%
Native Hawaiian/Other Pacific Islander	1	1	100%	1	100%
White	183	149	81%	112	75%
More than one race	61	56	92%	39	70%
Unknown	42	36	86%	29	81%
Gender					
Female	2,476	2,228	90%	2,001	90%
Male	3,350	3,007	90%	2,664	89%
Number of visits: median (range)	1 (1–92)	1 (1–92)		1 (1–92)	
Days between visits (gt 0): median (range)	397 (1–1,078)	398 (1–1,078)		402 (1–1,073)	

census tract level to match the included American Community Survey (ACS) 2015–2019 planning database [24] and to avoid the modifiable areal unit problem of using different geographic levels in the same study [25]. As available, data were reduced to years 2017–2019 to match the clinical data. Chicago crime [26] was characterized via FBI guidelines [27] as violent or not and rates for all crime, violent crime, and homicide (as a subset of violent crime) were created. Rates were also created for Chicago building code violations [26]. ChiVes, the Chicago data collaborative and community mapping application, assembled a novel dataset including tree cover, biodiversity, summer PM (particulate matter)

2.5 estimates, traffic levels, and housing cost burden, among others [28]. We developed a housing dataset, including building age and repair condition, sourced from the Cook County Tax Assessor's Office [29]. The Environmental Protection Agency (EPA)'s Environmental Justice mapping and screening tool (EJ Screen) data was also included [30].

Select variables in the ACS planning database [24] are relative to the population (census tract averages, percentages, or median values). ACS average and percentages variables with less than a 10% range were also excluded, as data needs to vary to find meaningful differences. Census tract race/ethnicity distributions

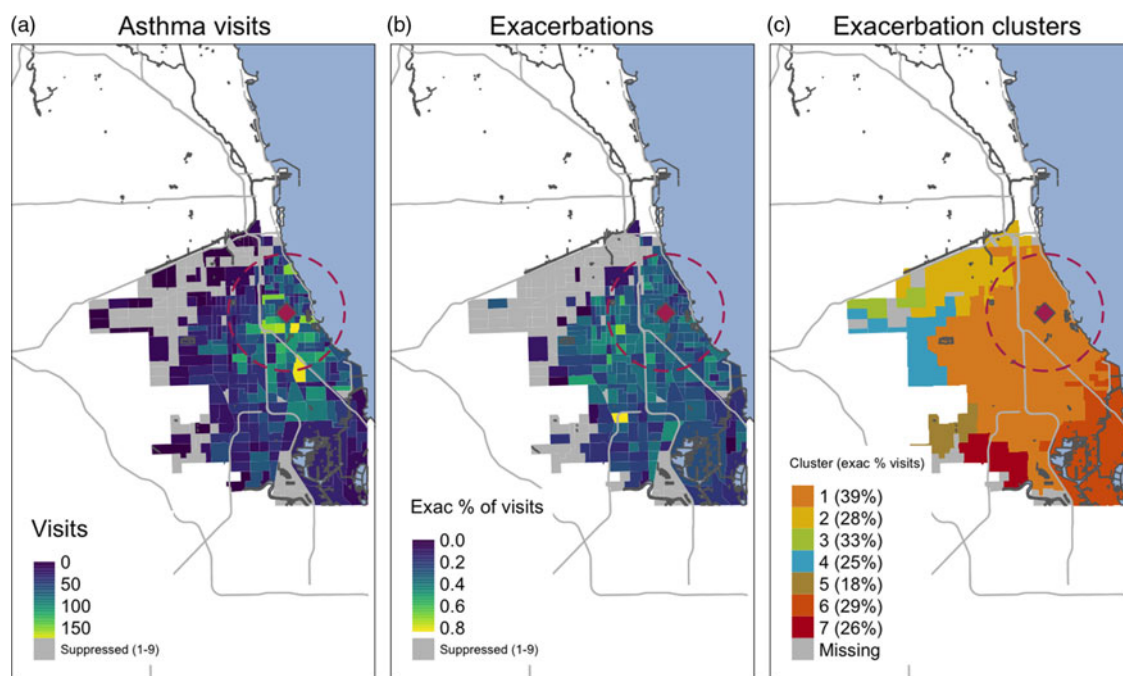


Figure 2. Asthma visits 2017–2019 by census tract. **a.** Asthma visit counts (continuous); **b.** Exacerbations as a proportion of all asthma visits; and **c.** Spatial clustering for exacerbations. The University of Chicago hospital is in red and its 5-mile perimeter is represented with a dashed red line.

were also excluded to prevent overfitting on race. Reciprocal measures (e.g., percentage of male and female) were reduced to one variable.

To account for poverty and still identify other components of neighborhood conditions, we collapsed all ACS poverty-related variables together. Correlations were assessed relative to the percentage of persons living below the poverty level. The inclusion threshold was set at [0.5]. Both Pearson (linear relationships) and Spearman (monotonic) methods were used to maximize inclusion. Principal component analysis (PCA), which combines variables together with a linear orthogonal transformation, was conducted for feature reduction for the poverty-correlated variables (hereafter “poverty PCA”) [31,32].

Model

Data were modeled on the visit level and restricted to asthma visits. To avoid a UChicago-specific EHR artifact that occurred in late 2016 as well as COVID-19 pandemic complications, data were restricted to visits from 2017 to 2019. Because of sparse data outside of the South Side of Chicago, this pilot was limited to census tracts on the South Side (Table 2 and Fig. 2). Patient race/ethnicity and insurance variables were excluded: race/ethnicity to prevent overfitting on race as well as to permit later bias testing and insurance status because public health insurance is a proxy for poverty. All 5 spatial cluster variations were included to determine which provided the most information to the model.

One nonlinear machine learning algorithm, a boosted decision tree [33], was piloted. Decision trees are non-parametric and do not make assumptions about the form of the data, can manage correlated data, have high dimensionality, and tolerate missingness. Boosted trees are an ensemble meta-algorithm, converting weak learning decision trees to strong ones in an iterative manner, such that each new tree improves upon the previous one. The

strengths of decision tree-based modeling provide the researcher the ability to include data elements not previously identified as risks in the literature, allowing for discovery of novel influences upon an outcome. Diverse datasets (including factors known to be related to asthma in the literature and, purposefully, elements that are not in the literature) were included in the model to allow for novel factor discovery. After this pilot period of discovery, more specific modeling will be undertaken.

Decision tree models allow more flexibility and use of more complex datasets but also risk overfitting to the noise or randomness in the data. To assess overfitting, we split the data into train and test sets. The model is designed (“fit”) on train data only, and then solely “run” on the test data to predict outcomes. Model prediction accuracy, defined below, is compared between train and test datasets to determine if the model is overfitting on the initial train data.

The outcomes were slightly imbalanced in that there were 2.4 times as many routine asthma visits as for exacerbations; this can affect prediction accuracy, especially for the minority group (exacerbations). The XGBoost model allows for rebalancing to prevent this via adjusting weights as a model hyperparameter. Other model hyperparameters were optimized with Hyperopt [34], and manually adjusted after fit assessment with a 70/30 train/test split; with the final hyperparameters, shuffled 5-fold cross-validation was conducted.

Evaluation

Metrics center on comparing model predictions to actual values. “Positive” indicates the outcome of interest, here, an asthma exacerbation, while “negative” indicates no exacerbation. “True” indicates that the actual outcome in the data matches the model-predicted outcome, while “false” indicates a mismatch between real and predicted outcomes. The metrics are: true positive (TP); true negative (TN); false positive (FP); and false negative (FN).

Accuracy is the proportion of correct predictions $((TP + TN) / (TP + FP + TN + FN))$. Test accuracy is the accuracy of the model on only the test dataset. Recall is the proportion of actual positives identified correctly $(TP / (TP + FN))$.

Variable importances were determined by gain, which is the relative improvement in accuracy contributed by a particular feature. Gain provides a *ranking* of all variables in the model to indicate which are most important. It is vital to note that decision trees are not regression lines. If two variables are correlated, the decision tree will rank as more important whichever variable provides better accuracy. For example, we included five different spatial clustering variables to test which model ranked highest, that is, which of the five provided the most gain. This served as *variable selection* for spatial clusters for future efforts.

A baseline model included clinical and ACS data only. This was followed by a model with all SDC datasets and clusters to determine any improvement in predictive power. Further, the model would decide which of the five cluster variations provided the most gain to the model and, additionally, if clusters were at all important *relative* to other variables.

Protocol Testing

Two data-driven inclusion protocols were tested as future optional tools for users, and each used lasso-regularized logistic regression [35] and added SDC datasets one by one to the clinical data (serially, not cumulatively). The first protocol assessed including full datasets via the AIC metric (Akaike information criterion; the lower the value, the better the model quality) [36]. The second protocol used lasso as variable selection. Each variable that reached statistical significance ($p < .05$) was included. Lasso regularization compresses coefficients to reduce bias and is a useful technique for variable selection. (See the supplement.)

Challenges

Challenges for this pilot include clinical data availability and the relatively low predictive power of variables (“signal”) with the high variability of individual patients (“noise”). There is also a sampling bias in that the University of Chicago sees a distinct patient population – demographically, socio-economically, and geographically – which likely does not generalize to other populations.

Results

Sociome Data Commons

We report initial technical metrics for the SDC.

SDC Scope and Quality

The initial data repository consists of 22 total datasets and 375 individual metadata entries documenting their quality, provenance, and scope. Dataset content categories include environmental exposures ($n = 16$), public safety ($n = 3$), demographics ($n = 2$), access and mobility ($n = 2$), property ($n = 1$), and economic activity ($n = 1$). The geographic levels include street addresses ($n = 14$), census tract ($n = 7$), census block ($n = 1$), and latitude/longitude points ($n = 3$).

In the initial pilot use case, only high-quality datasets were included. All 22 datasets to date have a data quality score of four or higher (see the supplement) and three of the datasets have a score of five. Furthermore, we found that datasets had varying

geographic granularities. The fine-grained data, i.e., data at a finer precision than census tract, were in varying formats and scope. This requires significant efforts to harmonize and align with the other datasets using our software toolkit. In all, the datasets include 1906 total variables.

Usability Metrics

We evaluated the incremental cost of adding new datasets to the SDC by measuring the time required by a data engineering intern. These datasets had already been assessed by the team for relevance and quality. The metrics measure the time needed to use the automated harmonization software to reformat the data and enter the metadata into the commons. Over 13 random datasets, the average time to find and add to the SDC was 25 minutes with a wide standard deviation of 21 minutes; this does not include data cleaning or any integration activities.

CONSORT and Clinical Trial Ethos

We leveraged the deep expertise of the clinicians on the SDC team to develop a dataset inclusion pipeline that resembles a clinical trial flowchart [37]. Datasets are assessed through a review process and progress through a number of stages until inclusion. Fig. 3 shows the current status of this inclusion process.

Asthma Pilot Results

Clinical data

Using DeGauss, 93% of all clinical data were successfully geocoded from patient address to latitude and longitude, census block, and census tract. The other 7% of addresses were post office boxes (0.4%), non-address text (0.2%), and “imprecise” (6.9%), where “the address was geocoded, but results were suppressed because the precision was intersection, zip, or city and/or the score was less than 0.5 [38].” Table 2 is restricted to asthma visits among residents of Chicago between 2017 and 2019.

The UChicago pediatric asthma patient population is 77% Medicaid and 89% non-Hispanic Black (Table 2.A). Given the hospital’s location and consequent patient catchment, restricting the data to the South Side of Chicago captured most asthma visits (Table 2.B). However, population changes from this restriction included reductions in the non-Hispanic Asian/Mideast Indian (78%) and White (81%) patients, while 97% of non-Hispanic Black and 92% of Hispanic patients remained, reflecting the segregation by race/ethnicity in Chicago. Moving between census tracts was not common, with a median count of census tract per patient of 1 (range 1–3).

Spatial clustering of census tract exacerbations as percentage of asthma visits was significant (Moran’s $I = 0.5958$, $p < .0001$) and also showed a dominant and geographically large cluster 1 where 39% of asthma visits were for exacerbations (Fig. 3c). In contrast, only 18% of visits were for exacerbations in cluster 5. The University of Chicago hospital itself resides within cluster 1, though the cluster extends far to the south and west (Fig. 3c). Restricting clinical data to patients within cluster 1, most asthma visits were retained. However, important changes include a dramatic reduction in the South Side Hispanic population (31%, Table 2.C).

Sociome Data Commons datasets

For the poverty PCA, Spearman added four additional variables not included in the Pearson results, while Spearman alone would have excluded eight identified from Pearson. Twenty-nine

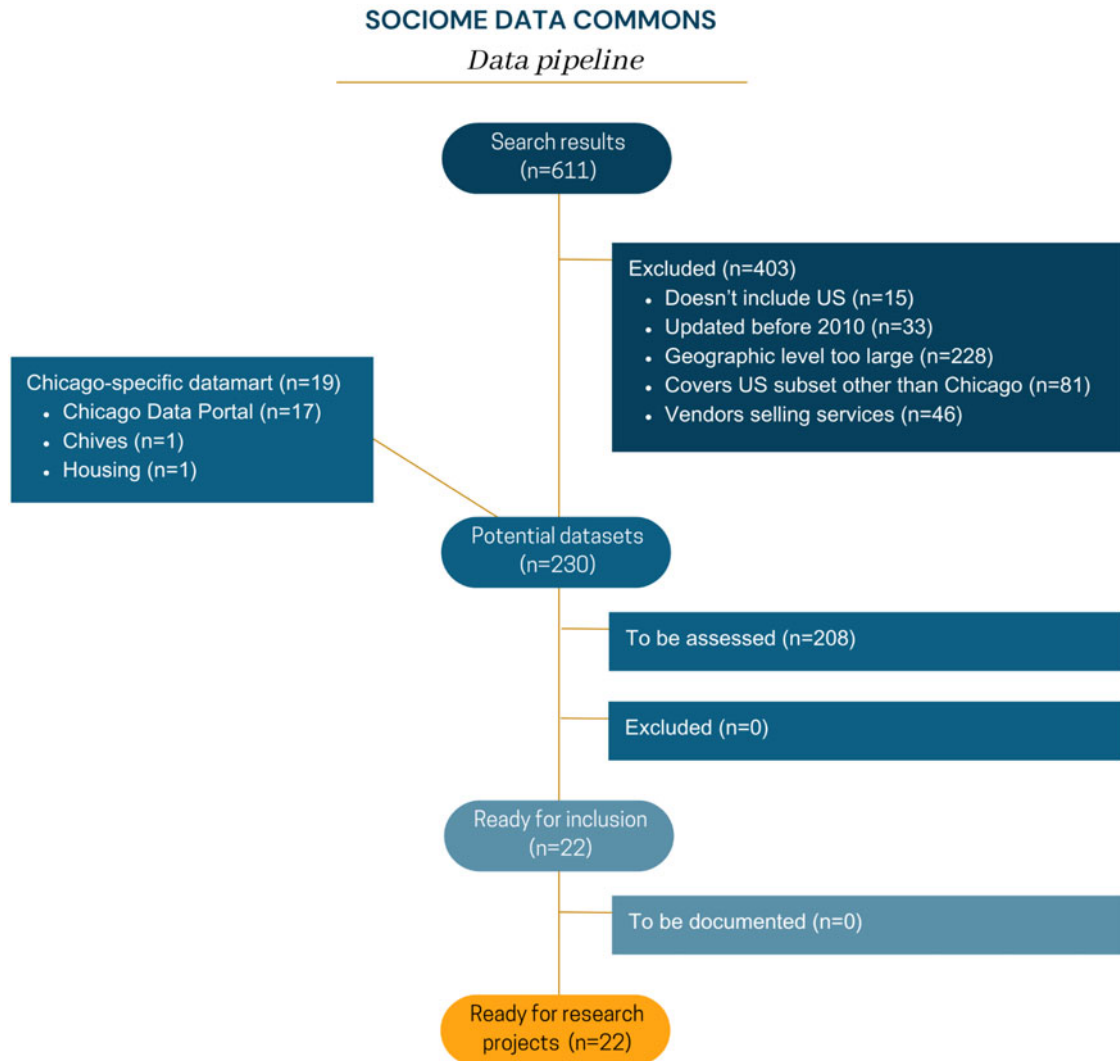


Figure 3. Sociome Data Commons (SDC): data pipeline.

variables correlated with percent below the poverty level at ≥ 0.5 or above (see the supplement) and, along with percent below the poverty level, were mean-centered and standardized [31]. These variables were reduced to one PCA loading, which explained 56% of variance.

Model

A baseline boosted tree was run with only clinical and ACS data (which included the poverty PCA) with a test accuracy of 58%. With all datasets for the entire South Side, accuracy was increased to 62%, with recall for one (an asthma exacerbation) at 65%.

Feature importance for the South Side (Fig. 4a) revealed clear outliers for gain around 21 and again at 28 (Fig. 4c). Most features were at the lowest end of gain. Spatial clustering appeared twice in the top features. Seven clusters and the average age of housing provided the most gain in accuracy, followed by age at visit, proximity to Superfund pollution sites (marked for decontamination by the EPA), median rent, and the violent crime rate. The proportion of residential housing, urban flood

susceptibility, and three spatial clusters followed in the top 10 variables.

Given the geographic dominance of spatial cluster 1 and its high proportion of asthma exacerbations, we ran a model only on those patients residing within that cluster. Cross-validated accuracy was 57%, accuracy was 61%, and recall for one was 60%.

For cluster 1, there is a clear grouping of top feature importances at gain above 25. Otherwise, there is an accumulation of features around 8. (Fig. 4d) The average age of housing units, the percentage of those under age 19 with no health insurance, the visit month, and the patient's age at visit were dominant variables in the model (Fig. 4b). These were followed by a second grouping of features in the top 10 of those 65 and over with no health insurance, housing cost burden, foreign born residents, median cost of rent, the poverty PCA, and several variables indicating a lack of health insurance.

Variable importances changed between the model including all of the South Side (Fig. 4a) and the model restricted to cluster 1 (Fig. 4b). Housing age (Fig. 5a) remained a top variable, as did patient age and median rent. However, moving from the full



Figure 4. Variable importance: **a.** Top 10 variables for the full south side; **b.** Top 10 variables for cluster 1; **c.** Histogram of gain for all variables, full south side; and **d.** Histogram of gain for all variables, cluster 1 only.

South Side to only cluster 1, housing cost burden (Fig. 5b) appeared in the top variables while urban flood susceptibility (Fig. 5c), the violent crime rate (Fig. 5d), and proximity to Superfund sites (Fig. 5e) left the top 10 important variables. Instead, lack of health insurance variables and the poverty PCA (Fig. 5f) gained importance.

Conclusion

Platform Discussion

The SDC has been purposefully designed to support and encourage extension of the platform into new data sets as well as the continued development, refinement, and adoption of SDC

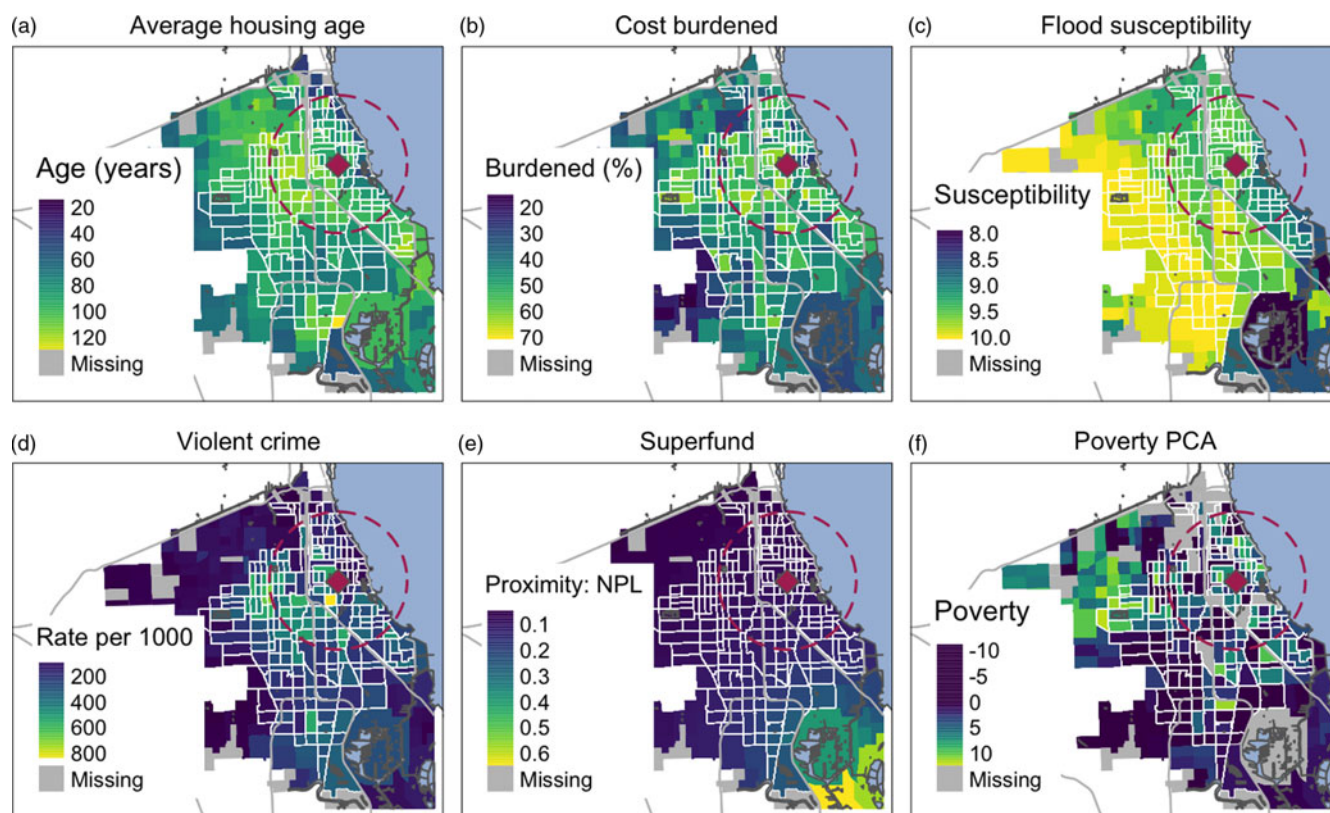


Figure 5. Select south side maps: **a.** Average housing age; **b.** Median rent; **c.** Urban flood susceptibility; **d.** Violent crime rate; **e.** Proximity to superfund sites; and **f.** Poverty principal component analysis (PCA). Cluster 1 census tracts are outlined in white, and the University of Chicago hospital is in red with its 5-mile perimeter represented with a dashed red line.

standards for dataset quality, dataset inclusion, metadata annotation, and data access/governance. The asthma pilot has served as the first driver use case for the SDC. Additional projects will be selected, in part for their ability to exercise and grow the capacity of the SDC to meet its ambitious goals. The purpose of this study is twofold: (1) to further the understanding of sociome factors in a variety of pathologies and (2) to understand principles of sustainable data commons design.

The first purpose has parallels to other similarly-intentioned ongoing efforts. The National Neighborhood Data Archive (NaNDA) at the University of Michigan [39], the Health Equity Explorer (H2E) at Boston Medical Center [40], Exposomics from the University of Utah [41], the City Health Dashboard from New York University [42], PopHR at the University of Tennessee [43], and SDOH and Place [44] are among many groups aggregating important data expressly to enable researchers' use of sociome factors.

We believe the second purpose – to understand the principles of sustainable data commons design – is unique to this project. The SDC is studying the design and implementation of such a data commons as a scientific problem including establishing quantitative metrics for assessing data quality, ease of use, researcher adoption, and sustainability. This manuscript contributes a framework for evaluating such criteria and we believe this to be an important contribution to this research area.

This pilot and manuscript have been tailored to Chicago, but we are collecting national datasets and can build additional location-specific datamarts as needed. However, local understanding provides critical context to properly using the data, and our group holds high knowledge of Chicago and available datasets. Key

informant interviews with community members and other content experts are underway, and the Institute for Translational Medicine's Community and Collaboration core is currently conducting community outreach activities. Notably, not all cities have tools such as the City of Chicago's Data Portal [26].

In ongoing work, the team plans to make the entire SDC infrastructure publicly available and open source. It will include the software artifacts designed as a part of the project, such as the data harmonization and analysis code. It will further open-source the management infrastructure of how to host and serve these datasets. Furthermore, we will release policy templates for data governance and quality assurance. Once we finish integration of datasets, the SDC will serve as a reference implementation for data commons across a variety of social contexts of health research problems.

Analysis Discussion

The pilot use case reported above was helpful in providing direction for our future SDC and analytic efforts. Spatial clustering of exacerbations, housing conditions, proximity to Superfund sites, violent crime, urban flood susceptibility, lack of health insurance, and the poverty PCA all contributed importantly to prediction of asthma exacerbation. Some of these reflect current literature on risks for asthma, such as housing conditions and violence [45,46]. Further, housing variables appeared in the top 20 variables, such as the building violation rate and mobile home percentage (see the supplement).

Of course, these housing quality indicators are only proxy estimates for each individual patient's actual housing conditions and exposures, such as indoor air quality, first, second, or third-

hand smoke exposure, and mold or pest exposure [47–50]. Given a subset of patients' actual housing and indoor air quality, we could work to identify which, if any, SDC datasets could serve as the best proxy [51]. The addition of personal exposure data might increase predictive accuracy, and the extent to which this occurs would inform how well (or not) generalized survey data like the sociome datasets perform in their stead.

Proximity to Superfund sites merits further exploration of these and other pollution sites such as landfills and risk management plan sites. Exploring patient-level distance to these, rather than census tract estimates, is a next step. Notably, known risks such as PM_{2.5} exposure and traffic proximity [48] did not appear in the top 10 variables, though they are in the top 20 (see the supplement).

Other findings, such as the violent crime rate and lack of insurance, also replicate the literature [46,52]. Rarely-seen findings [53] to be further explored and include urban flood susceptibility, which could indicate poorer housing quality and perhaps indicate susceptibility to damp housing and mold growth. Further exploration of flooding is needed.

In this pilot, spatial clustering proved to be important. Model stratification was only conducted for the dominant cluster 1. A comparison of top variable importances demonstrates the promise of providing geography-specific risks (Fig. 4), though further work is needed to clarify the reasons for the cluster differences, as well as a comparison model for the full South Side excluding clustering variables. Still, future work exploring cluster-specific models could inform geographically tailored interventions.

Limitations

This study suffers from several data set limitations. For example, some probability surveys (such as NHANES [54]) contain important sociome data but are not publicly available at the census tract level.

The models resulted in rather weak signals. While we anticipated that the importance might lie in an aggregation of multiple weak signals, predictive improvement is still needed. By choosing just a few datasets, we might not have yet included the most important datasets. We anticipate that broadening the range of sociome factors in an expanded SDC (which we are currently building) may increase model performance.

The University of Chicago catchment area does not generalize to all pediatric asthma patients in Chicago. Varied data are needed to appreciate differences, and we need EHR data from other metropolitan Chicago health systems to provide greater patient heterogeneity. Efforts to expand the data are underway with our partners from the Institute for Translational Medicine [55].

The spatial clustering of exacerbations might be affected by proximity to the UChicago hospital, as exacerbations are often urgent or emergency events. However, cluster 1 does extend far to the south and southwest of the hospital. Adding other hospitals' data should clarify clustering of exacerbations against hospital proximity.

Analysis needs to be expanded. We piloted data as cross-sectional, and future efforts will use longitudinal methods and correct for this temporal bias. Including both temporality-based regression discontinuity and high-dimensionality mediation analyses will also enable causal exploration. Patient-level, rather than visit-level, analysis (as a multilevel generalized estimating equation) might also increase predictive performance, as many of the asthma exacerbation visits could have been from a particular subset of patients. By matching patients to specific locations in our

housing dataset, we can move to sub-census tract metrics. Future efforts should include patient-specific distance to the nearest hospital and also investigate asthma exacerbations in more detail, for example, if a visit to the emergency department was needed. Modeling itself needs to be expanded. With machine learning, multiple models can be combined, leveraging the models' strengths and balancing their respective deficiencies.

For this pilot study, we were missing data on many traditional clinical phenotypes (disease severity, atopic status, comorbidities) and all biological (genetic, epigenetic) information about the pediatric asthma patients included in this study. As a next step, via the structured flow sheet, we will include the asthma control test in future efforts. Asthma exacerbations will also be defined via medication prescriptions, though pharmacy dispensing information is not available. Phenotypes (atopic/non, obesity related, etc.) will be assigned. If we are able to obtain biomarker data for select patients, that could allow identification of gene-environment interactions.

The users were restricted to the project team with insights and guidance provided by the team. A robust governance framework will be crafted. In addition to the data governance discussed here, we will address appropriate use, responsible use, and community impact including a flexible regulatory module to meet varied and changing requirements. We will activate a multi-disciplinary work group composed of technical, research, legal experts, ethicists, and community representatives.

In sum, the pilot study reported here demonstrates promise for future analyses of the complex interactions of the sociome and clinical health factors using the SDC. We expect that its further development, including accounting for dataset quality metrics, will facilitate the accounting for sociome factors in a wide range of clinical research topics, from analysis of response to cancer immunotherapy to pandemic preparedness [56], to common complex diseases like diabetes mellitus [57], and more.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/cts.2023.670>.

Acknowledgments. The authors thank Lila Midyett and Krish Modi for their help reviewing the literature.

Author contributions. BB: Software, Data Curation; AB: Writing-Original Draft; SB: Writing-Review & Editing; BC: Data Curation; CC: Resources; KB: Project Administration; EC: Conceptualization; ED: Writing – Review & Editing; BF: Writing-Review & Editing; JJ: Methodology, Investigation, Resources, Data Curation; HZ: Resources; NK: Resources; SK: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing Original-Draft, Writing- Review & Editing, Supervision; KL: Data Curation; KL: Software, Data Curation, Writing-Original Draft; SM: Data Curation; DM: Conceptualization; SN: Writing – Original Draft; JO: Conceptualization, Writing-Review & Editing; CS: Writing – Original Draft, Writing – Review & Editing; MS: Data Curation, Writing - Review & Editing; AS: Conceptualization, Investigation, Supervision, Writing – Original Draft, Writing- Review & Editing; JS: Conceptualization, Writing – Review & Editing; ST: Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization; SV: Conceptualization, Resources, Writing – Original Draft, Writing – Review & Editing.

Funding statement. This work was supported by NIH award number UL1TR002389-07.

Competing interests. S.N. served on an advisory board for Avillion/Astra Zeneca, receives royalties from Wolters-Kluwer and Springer, and research funding from NIH and Asthma Allergy Foundation of America.

C.S. served on advisory boards for Gilead and Merck, receives royalties from Wolters-Kluwer, and research funding from CDC.

A.S. Holds voluntary positions in the American Medical Informatics Association and is an equity investor in healthcare companies and other industries.

J.S. reports a potential financial interest in PulmOne Advanced Medical Devices, Ltd, Israel, and research grant funding from NIH, NSF, and Respiratory Health Association of Metropolitan Chicago.

S.V. is co-founder and Chief Medical Officer for Litmus Health, Inc., and receives consulting royalties from CVS Accordant.

References

1. **WHO Commission on Social Determinants of Health, World Health Organization.** *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health: Commission on Social Determinants of Health Final Report.* Geneva, Switzerland: WHO Press; 2008. (https://play.google.com/store/books/details?id=zc_VfH7wfv8C).
2. **Mazilu L, Paton NW, Konstantinou N, Fernandes AAA.** Fairness-aware data integration. *J Data and Information Quality.* 2022;**14**(4):1–26. doi: [10.1145/3519419](https://doi.org/10.1145/3519419).
3. **Krasikov P, Legner C.** A method to screen, assess, and prepare open data for use. *J Data and Information Qual.* 2023;**15**(4):1–25. doi: [10.1145/3603708](https://doi.org/10.1145/3603708).
4. **Batini C, Cappiello C, Francalanci C, Maurino A.** Methodologies for data quality assessment and improvement. *ACM Comput Surv.* 2009;**41**(3):1–52. doi: [10.1145/1541880.1541883](https://doi.org/10.1145/1541880.1541883).
5. **Storås AM, Strümke I, Riegler MA, Halvorsen P.** Explainability methods for machine learning systems for multimodal medical datasets: research proposal. In: *Proceedings of the 13th ACM Multimedia Systems Conference. MMSys '22*, Athlone, Ireland: Association for Computing Machinery. June 14–17, 2022:347–351.
6. **Patel D, Hall GL, Broadhurst D, Smith A, Schultz A, Foong RE.** Does machine learning have a role in the prediction of asthma in children? *Paediatr Respir Rev.* 2021;**41**:51–60. doi: [10.1016/j.prrv.2021.06.002](https://doi.org/10.1016/j.prrv.2021.06.002).
7. **Volchenbom SL, Cox SM, Heath A, Resnick A, Cohn SL, Grossman R.** Data commons to support pediatric cancer research. *Am Soc Clin Oncol Educ Book.* 2017;**37**(37):746–752. doi: [10.1200/EDBK_175029](https://doi.org/10.1200/EDBK_175029).
8. **Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al.** The FAIR guiding principles for scientific data management and stewardship. *Sci Data.* 2016;**3**(1):160018. doi: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
9. **2019 Community Health Needs Assessment - UChicago Medicine.** (https://issuu.com/communitybenefit-ucm/docs/ucm-2019-chna?fr=xKAE9_zU1NQ). Accessed July 11, 2023.
10. **Dirksen JC, Prachand NG.** *Healthy Chicago 2.0: Partnering to Improve Health Equity 2016–2020.* City of Chicago, Chicago, IL; 2016.
11. **Grant T, Croce E, Matsui EC.** Asthma and the social determinants of health. *Ann Allergy Asthma Immunol.* 2022;**128**(1):5–11. doi: [10.1016/j.anai.2021.10.002](https://doi.org/10.1016/j.anai.2021.10.002).
12. **Sullivan K, Thakur N.** Structural and social determinants of health in asthma in developed economies: a scoping review of literature published Between 2014 and 2019. *Curr Allergy Asthma Rep.* 2020;**20**(2):5. doi: [10.1007/s11882-020-0899-6](https://doi.org/10.1007/s11882-020-0899-6).
13. **Khanam UA, Gao Z, Adamko D, et al.** A scoping review of asthma and machine learning. *J Asthma.* 2022;**60**(2):1–13. doi: [10.1080/02770903.2022.2043364](https://doi.org/10.1080/02770903.2022.2043364).
14. **Kothalawala DM, Kadalayil L, Weiss VBN, et al.** Prediction models for childhood asthma: a systematic review. *Pediatr Allergy Immunol.* 2020;**31**(6):616–627. doi: [10.1111/pai.13247](https://doi.org/10.1111/pai.13247).
15. **Van Rossum G, Drake FL.** *Python 3 Reference Manual: (Python Documentation Manual Part 2).* Scotts Valley, CA: CreateSpace Independent Publishing Platform, 2009. (<https://play.google.com/store/books/details?id=KlybQQAACAAJ>).
16. **Cook County Government.** Cook County Address Points. Cook County Open Data. Published June 1, 2016. (<https://datacatalog.cookcountyil.gov/>, <https://hub-cookcountyil.opendata.arcgis.com/datasets/5ec856ded93e4f85b3f6e1bc027a2472>). Accessed September 2, 2022.
17. **Brokamp C, Wolfe C, Lingren T, Harley J, Ryan P.** Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. *J Am Med Inform Assoc.* 2018;**25**(3):309–314. doi: [10.1093/jamia/ocx128](https://doi.org/10.1093/jamia/ocx128).
18. **Welcome to GeoPy's documentation! — GeoPy 2.3.0 documentation.** (<https://geopy.readthedocs.io/en/stable/>). Accessed July 11, 2023.
19. **U.S. Census Bureau. Census Geocoder Documentation.** United States Census Bureau. Published June 21, 2022. (<https://www.census.gov/programs-surveys/geography/technical-documentation/complete-technical-documentation/census-geocoder.html>). Accessed September 16, 2022.
20. **Sklearn. Impute.IterativeImputer.** scikit-learn. (<https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html>). Accessed July 11, 2023.
21. **Assunção RM, Neves MC, Câmara G, Da Costa Freitas C.** Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Geogr Inf Syst.* 2006;**20**(7):797–811. doi: [10.1080/13658810600665111](https://doi.org/10.1080/13658810600665111).
22. **R Core Team.** *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2020. (<https://www.R-project.org/>).
23. **Zhou X, Lin H.** Moran's I. In: Shekhar S, Xiong H, eds. *Encyclopedia of GIS.* Springer US; 2008:725–725. doi: [10.1007/978-0-387-35973-1_817](https://doi.org/10.1007/978-0-387-35973-1_817).
24. **U.S. Census Bureau.** 2015–2019 American Community Survey Planning Database Tract Data. Planning Database. Published 2021. (<https://www.census.gov/topics/research/guidance/planning-databases.2021.html>). Accessed August 23, 2022.
25. **Wong DWS.** The modifiable areal unit problem (MAUP). In: Janelle DG, Warf B, Hansen K, eds. *WorldMinds: Geographical Perspectives on 100 Problems: Commemorating the 100th Anniversary of the Association of American Geographers 1904–2004.* Springer Netherlands; 2004:571–575. doi: [10.1007/978-1-4020-2352-1_93](https://doi.org/10.1007/978-1-4020-2352-1_93).
26. **City of Chicago.** Chicago Data Portal. Published 2023. (<https://data.cityofchicago.org/>). Accessed October 29, 2020.
27. **Violent crime.** FBI. (<https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime>). Accessed July 11, 2023.
28. **Healthy Regions & Policies Lab (HeRoP), Department of Geography & GIScience, University of Illinois at Urbana-Champaign and Department of Geography, DePaul University.** Data - ChiVes. Uncover the nature of Chicago's environment. Published April 21, 2023. (<https://chichives.com/>). Accessed April 28, 2023.
29. **Kaegi F.** Cook County Assessor's Office. Published 2023. (<https://www.cookcountyassessor.com/>). Accessed 2022.
30. **United States Environmental Protection Agency. OP.** EJScreen: Environmental Justice Screening and Mapping Tool. Accessed September 3, 2014. (<https://www.epa.gov/ejscreen>).
31. **Pedregosa F, Varoquaux G, Gramfort A, et al.** Scikit-learn: machine learning in python. *J Mach Learn Res.* 2011;**12**, 2825–2830.
32. **SKLearn PCA decomposition.** scikit-learn. (<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>). Accessed July 25, 2023.
33. **Chen T, Guestrin C.** A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16.* New York, NY: Association for Computing Machinery; 2016:785–794. doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785).
34. **Bergstra J, Yamini D, Cox D.** Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2013, **28**, 115–123. (<https://proceedings.mlr.press/v28/bergstra13.html>).
35. **Seabold S, Perktold J.** Statsmodels: econometric and statistical modeling with python. In: *Proceedings of the 9th Python in Science Conference.* Austin, TX: SciPy; 2010. doi: [10.25080/majora-92bf1922-011](https://doi.org/10.25080/majora-92bf1922-011).
36. **Cavanaugh JE, Neath AA.** The akaike information criterion: background, derivation, properties, application, interpretation, and refinements. *Wiley Interdiscip Rev Comput Stat.* 2019;**11**(3):e1460. doi: [10.1002/wics.1460](https://doi.org/10.1002/wics.1460).
37. **Equator network: Enhancing the QUALity and Transparency Of health Research.** CONSORT 2010 Statement: updated guidelines for

- reporting parallel group randomised trials. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Published May 3, 2023. (<https://www.equator-network.org/reporting-guidelines/consort/>). Accessed June 1, 2023.
38. **Brokamp C.** DeGAUSS: decentralized geospatial assessment for multi-site studies. *J Open Source Softw.* 2018;**3**(30):812. doi: [10.21105/joss.00812](https://doi.org/10.21105/joss.00812).
 39. **Social Environment and Health program at the University of Michigan Institute for Social Research.** National Neighborhood Data Archive (NANDA). National Neighborhood Data Archive (NaNDA) - Institute for Social Research. Published 2023. (<https://nanda.isr.umich.edu/>). Accessed May 1, 2023.
 40. **Adams WG, Gasman S, Beccia A, Cabral HJ.** Leveraging the OMOP common data model to support distributed health equity research, poster 518. In: *Pediatric Academic Societies Meeting, Washington, DC*, 2023.
 41. **Exposomics.** University of Utah School of Medicine. Published 2023. (<https://medicine.utah.edu/dbmi/expertise/exposomics>). Accessed June 1, 2023.
 42. **Betro J, Breslin S, Chen A, et al.** City Health Dashboard: Empowering cities to create thriving communities. Published July 27, 2023. (<https://www.cityhealthdashboard.com/>). Accessed July 30, 2023.
 43. **Shaban-Nejad A, Lavigne M, Okhmatovskaia A, Buckeridge DL.** PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data. *Ann N Y Acad Sci.* 2017;**1387**(1):44–53. doi: [10.1111/nyas.13271](https://doi.org/10.1111/nyas.13271).
 44. **Kolak M, Bhatt J, Park YH, Padrón NA, Molefe A.** Quantification of neighborhood-level social determinants of health in the continental united states. *JAMA Network Open.* 2020;**3**(1):e1919928. doi: [10.1001/jamanetworkopen.2019.19928](https://doi.org/10.1001/jamanetworkopen.2019.19928).
 45. **Bryant-Stephens TC, Strane D, Robinson EK, Bhambhani S, Kenyon CC.** Housing and asthma disparities. *J Allergy Clin Immunol.* 2021;**148**(5):1121–1129. doi: [10.1016/j.jaci.2021.09.023](https://doi.org/10.1016/j.jaci.2021.09.023).
 46. **Landeo-Gutierrez J, Forno E, Miller GE, Celedón JC.** Exposure to violence, psychosocial stress, and asthma. *Am J Respir Crit Care Med.* 2020;**201**(8):917–922. doi: [10.1164/rccm.201905-1073PP](https://doi.org/10.1164/rccm.201905-1073PP).
 47. **Kang I, McCreery A, Azimi P, et al.** Impacts of residential indoor air quality and environmental risk factors on adult asthma-related health outcomes in Chicago, IL. *J Expo Sci Environ Epidemiol.* 2023;**33**(3):358–367. doi: [10.1038/s41370-022-00503-z](https://doi.org/10.1038/s41370-022-00503-z).
 48. **Grant TL, Wood RA.** The influence of urban exposures and residence on childhood asthma. *Pediatr Allergy Immunol.* 2022;**33**(5):e13784. doi: [10.1111/pai.13784](https://doi.org/10.1111/pai.13784).
 49. **Rousseau JF, Oliveira E, Tierney WM, Khurshid A.** Methods for development and application of data standards in an ontology-driven information model for measuring, managing, and computing social determinants of health for individuals, households, and communities evaluated through an example of asthma. *J Biomed Inform.* 2022;**136**:104241. doi: [10.1016/j.jbi.2022.104241](https://doi.org/10.1016/j.jbi.2022.104241).
 50. **Tiotiu AI, Novakova P, Nedeva D, et al.** Impact of air pollution on asthma outcomes. *Int J Environ Res Public Health.* 2020;**17**(17):E6212. doi: [10.3390/ijerph17176212](https://doi.org/10.3390/ijerph17176212).
 51. **Bozigar M, Connolly CL, Legler A, et al.** In-home environmental exposures predicted from geospatial characteristics of the built environment and electronic health records of children with asthma. *Ann Epidemiol.* 2022;**73**:38–47. doi: [10.1016/j.annepidem.2022.06.034](https://doi.org/10.1016/j.annepidem.2022.06.034).
 52. **Hasegawa K, Stoll SJ, Ahn J, Kysia RF, Sullivan AF, Camargo CA Jr.** Association of insurance status with severity and management in ED patients with asthma exacerbation. *West J Emerg Med.* 2016;**17**(1):22–27. doi: [10.5811/westjem.2015.11.28715](https://doi.org/10.5811/westjem.2015.11.28715).
 53. **Larson PS, Gronlund C, Thompson L, et al.** Recurrent home flooding in Detroit, MI 2012–2020: results of a household survey. *Int J Environ Res Public Health.* 2021;**18**(14):7659. doi: [10.3390/ijerph18147659](https://doi.org/10.3390/ijerph18147659).
 54. **Centers for Disease Control and Prevention (CDC).** National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey (NHANES). National Health and Nutrition Examination Survey. Published May 31, 2023. (<https://www.cdc.gov/nchs/nhanes/index.htm>). Accessed July 25, 2023.
 55. **Institute For Translational Medicine (ITM).** Published September 19, 2017. (<https://chicagoitm.org/>). Accessed July 30, 2023.
 56. **Badalov E, Blackler L, Scharf AE, et al.** COVID-19 double jeopardy: the overwhelming impact of the social determinants of health. *Int J Equity Health.* 2022;**21**(1):76. doi: [10.1186/s12939-022-01629-0](https://doi.org/10.1186/s12939-022-01629-0).
 57. **Egede LE, Campbell JA, Walker RJ, Linde S.** Structural racism as an upstream social determinant of diabetes outcomes: a scoping review. *Diabetes Care.* 2023;**46**(4):667–677. doi: [10.2337/dci22-0044](https://doi.org/10.2337/dci22-0044).
 58. **DDI Alliance.** DDI Alliance. Document, Discover and Interoperate. Published 2023. (<https://ddialliance.org/>). Accessed July 1, 2023.
 59. **Chu X, Ilyas IF, Krishnan S, Wang J.** Data cleaning: overview and emerging challenges. In: *Proceedings of the 2016 International Conference on Management of Data. SIGMOD '16.* San Jose, CA: Association for Computing Machinery; 2016, 2201–2206. doi: [10.1145/2882903.2912574](https://doi.org/10.1145/2882903.2912574).