



## OPEN ACCESS

## EDITED BY

Brian D. Davison,  
Lehigh University, United States

## REVIEWED BY

Chengxi Zang,  
Cornell University, United States  
Sachi Nandan Mohanty,  
College of Engineering Pune, India

## \*CORRESPONDENCE

Melody L. Greer  
mlgreer@uams.edu  
Sudeepa Bhattacharyya  
sbhattacharyya@astate.edu

## SPECIALTY SECTION

This article was submitted to  
Data Mining and Management,  
a section of the journal  
Frontiers in Big Data

RECEIVED 11 March 2022

ACCEPTED 07 July 2022

PUBLISHED 01 August 2022

## CITATION

Greer ML, Zayas CE and  
Bhattacharyya S (2022) Repeatable  
enhancement of healthcare data with  
social determinants of health.  
*Front. Big Data* 5:894598.  
doi: 10.3389/fdata.2022.894598

## COPYRIGHT

© 2022 Greer, Zayas and  
Bhattacharyya. This is an open-access  
article distributed under the terms of  
the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution  
or reproduction in other forums is  
permitted, provided the original  
author(s) and the copyright owner(s)  
are credited and that the original  
publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or  
reproduction is permitted which does  
not comply with these terms.

# Repeatable enhancement of healthcare data with social determinants of health

Melody L. Greer<sup>1\*</sup>, Cilia E. Zayas<sup>1</sup> and  
Sudeepa Bhattacharyya<sup>1,2\*</sup>

<sup>1</sup>Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, United States, <sup>2</sup>Department of Biological Sciences, Arkansas Biosciences Institute, Arkansas State University, Jonesboro, AR, United States

**Background:** Social and behavioral aspects of our lives significantly impact our health, yet minimal social determinants of health (SDOH) data elements are collected in the healthcare system.

**Methods:** In this proof-of-concept study we developed a repeatable SDOH enrichment and integration process to incorporate dynamically evolving SDOH domain concepts from consumers into clinical data. This process included SDOH mapping, linking compiled consumer data to patient records in Electronic Health Records, data quality analysis and preprocessing, and storage.

**Results:** Consumer compilers data coverage ranged from ~90 to ~54% and the percentage match rate between compilers was between ~21 and 64%. Our preliminary analysis showed that apart from demographic factors, several SDOH factors like home-ownership, marital-status, presence of children, number of members per household, economic stability and education were significantly different between the COVID-19 positive and negative patient groups while estimated family-income and home market-value were not.

**Conclusion:** Our preliminary analysis shows commercial consumer data can be a viable source of SDOH factor at an individual-level for clinical data thus providing a path for clinicians to improve patient treatment and care.

## KEYWORDS

data quality, social determinants of health (SDOH), electronic health records (EHR), health informatics, healthcare, data curation

## Introduction

Socioeconomic and behavioral aspects of our lives significantly impact our health, yet minimal social determinants of health (SDOH) data elements are collected in the healthcare system. Information of this type is needed for quality healthcare research and patient care because it is associated with the full-spectrum of health outcomes from acute to chronic disorders. Studies indicate cancer (Alcaraz et al., 2020), cardiovascular disease (Tamura et al., 2019), dementia (Nicholas et al., 2021), mental health and substance-abuse disorders (Galea and Vlahov, 2002; Alegría et al., 2018), viral infection (Greer et al., 2021), and sleep (Grandner and Fernandez, 2021) are among a long list of health problems (Kivimäki et al., 2020) which are linked to social risk factors not frequently

or consistently collected for patients. The combined effect of missing, inconsistent, or inaccurate data also leads to bias in machine learning, algorithms underlying clinical decision support, predictive analytics or other healthcare processes (Obermeyer et al., 2019; Cottrell et al., 2020; Seker et al., 2022). To avoid these problems as well as to gain rich insights from healthcare data we must be cognizant about diverse data collection, veracity and data-quality.

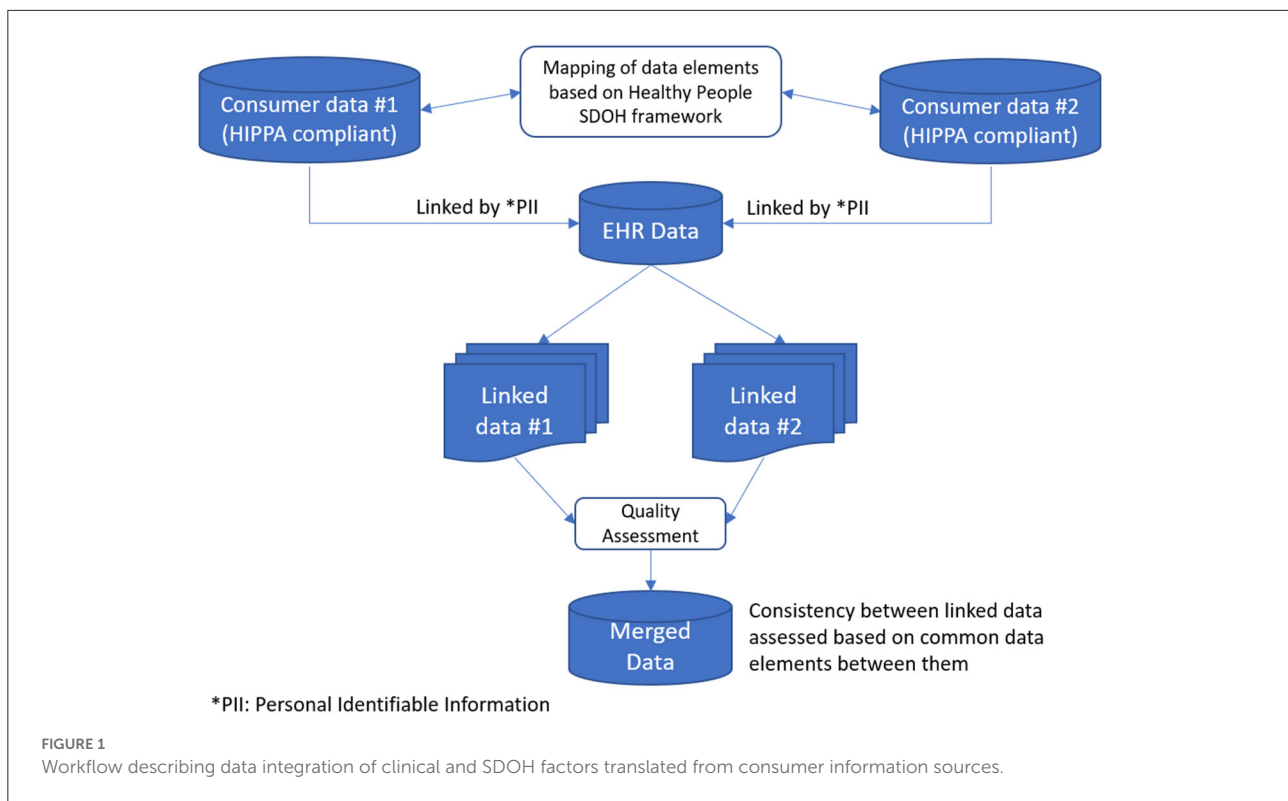
Electronic health records (EHR) are assembled from clinical, insurance and basic demographic information during the course of patient care. EHRs are real time, digital patient records containing, medical history, diagnoses, medications, treatment plans, immunizations dates, allergies, medical images, laboratory and test results. Modern EHR systems are built to go beyond standard clinical data collected in a healthcare facility and can include a broader view of a patient's care to provide a holistic assessment of a patient's health (Office of Disease Prevention Health Promotion, 2022a). Although there is growing interest in including SDOH data into EHR, the capture and management mechanisms for this process are uncertain (Ancker et al., 2018; Gold et al., 2018; Feldman et al., 2020). Current possibilities include (1) paper-based or digitally collected social needs screening before a patient receives care, (2) during a visit with a clinician, or (3) publicly available data sets that provide social context. All of these methods have challenges. Screening data that is incorporated into the EHR appears fragmented to users, increases the staff workload plus adds a data entry step where paper is used (Gold et al., 2018). And although clinicians felt SDOH information was valuable they are already pressed for time, and adding to their list of clerical tasks is often not practical or even possible (Tong et al., 2018). These barriers are likely the reason for the findings of Frazee et al. (2019) that social needs screening for food, housing, utilities, transportation, and experience with interpersonal violence was present in 24.4% of hospitals, and 15.6% of physician practices. SDOH data documented within unstructured EHR fields during a visit needs further study and will require natural language processing tools to be integrated into healthcare practice (Hatef et al., 2019). Publicly available data, whether in raw form (i.e., Census), or aggregated into measures [i.e., Social Vulnerability Index (SVI), Area Deprivation Index (ADI)] (Centers for Disease Control Prevention, 2015; Kind and Buckingham, 2018) has been valuable in studying the relationships between socioeconomic and patient health status (Johnson-Lawrence et al., 2017; Tung et al., 2018; Chamberlain et al., 2020). However, appending community-level data introduces issues with averaging. In their 2020 work on community-level and patient-level social risk data Cottrell et al. (2020) observed that community-level data misses some patients that patient-level data would not. Further complicating the issue, there are no currently accepted standard SDOH data elements (Cantor and Thorpe, 2018). This means that even in cases

where social risk data is consistently collected it cannot be easily shared with other healthcare providers during transfers or referrals.

Compiled consumer data can address these issues. Consumer data is the trail of information that customers leave behind as a result of their purchases and internet usage. This data, collated from multiple sources, can comprise of personal information and are sourced from social media networks, marketing campaigns, customer service requests, call center communications, online browsing data, mobile applications, purchase history, preferences and many others. These data are constantly being collected and analyzed by companies as part of a broad customer relationship management strategy. Finance and marketing businesses have successfully used this data for over 20 years to find customers, understand their needs, and tailor financial products. Using these same strategies, healthcare researchers and providers can improve patient treatment and care. The consumer data includes individual-level SDOH data providing a holistic snapshot of an individual's lifestyle. It includes amongst others, income, education, lifestyle variables, language spoken, household size, smoking status, life events, hobbies, shopping activity etc. that are not available in the insurance claims data or majority of EHR data as shown in Table 1. A support system, for complex assessments (i.e., risk assessments) and calculations, is needed using information that helps to arrive at conclusions regarding patient's health risk and treatment. The development of an automated pipeline process for multisource healthcare data integration will provide this support. The presence of information integrated from multiple different streams of data will support tools for nurses, social workers, community health workers and patient navigators that supports decision making by providing the ability to

TABLE 1 Consumer data and EHR data element categories.

EHR	Consumer
Basic demographics	Broad demographics
Clinical elements	Income bins
Diagnosis	Employment
Vital sign measures	Economic status bins
Inpatient and outpatient encounter	Education
Insurance	Insurance type
	Home owner status
	Vehicle owner status
	Lifestyle factors (i.e., sports, diet, smoking, and alcohol)
	Hobbies
	Neighborhood
	Weather



consider multiple factors simultaneously for patients and clinicians. To begin the development of an automated pipeline process for multisource healthcare data integration we have conducted a pilot study integrating clinical and consumer information sources to evaluate multiple SDOH and clinical factors simultaneously.

## Methods

Our goal was to prototype a repeatable clinical data enhancement process to incorporate compiled consumer data into SDOH domain concepts. This process included social risk factor mapping, linking compiled consumer data, data quality assessment, preprocessing, and storage. Consumer data does not currently line up one-to-one with factors identified as SDOH, so purchased data elements were mapped to concepts identified in social needs screening. Once the mapping was complete, the compiled consumer data elements were linked to the patient population and stored in a Microsoft SQL Server. The resulting data was then accessible using SQL queries, and the quality was evaluated for completeness, consistency, and timeliness or temporal alignment. Throughout SDOH enhancement, a security and data privacy layer overlaid the entire process with security features included in each step

to ensure data privacy. [Figure 1](#) depicts a snapshot of the entire process.

## Mapping

Various social determinants of health frameworks have been created to assist communities, healthcare professionals, and others in better identifying and managing an expansive range of factors influencing health outcomes ([World Health Organization, 2010](#); [Office of Disease Prevention Health Promotion, 2022a](#); [Rural Health Information Hub, 2022](#)). Our mapping strategy was developed based on the Healthy People (HP) 2030 SDOH Framework for this study ([Office of Disease Prevention Health Promotion, 2022a,b](#)). HP 2030 not only continued the HP initiative, which set national health targets for 2020 through 2030, but also designed a framework to organize SDOH into five domains: (1) economic stability, (2) education, (3) social and community context, (4) health and healthcare access, and (5) the neighborhood and built environment. HP 2030 outlined essential SDOH within each of these domains ([Office of Disease Prevention Health Promotion, 2022b](#)). Employment, food insecurity, housing instability, and poverty, for example, all fall under the domain of economic stability. We mapped as many elements as possible

TABLE 2 Data elements mapped from consumer databases to SDOH categories with coverage percentages listed for each individual compiler.

SDOH domains	Element category	Compiler 1 coverage	Compiler 2 coverage
Economic stability	<b>Employment</b>		
	• Occupation	48.23%	54.31%
	<b>Food insecurity</b>		
	• Health natural foods	Not reported	54.31%
	• Food and drink	Not reported	54.31%
	• Grocery	Not reported	54.31%
	<b>Housing instability</b>		
	• Homeowner/renter	98.98%	54.31%
	• Length of residence	98.98%	54.31%
	<b>Poverty</b>		
	• Estimated income	98.98%	54.31%
	• Net worth indicator	85.98%	54.31%
	• Economic stability indicator	85.98%	Not reported
	• Estimated discretionary income %	Not reported	54.31%
	• Estimated household debt level	Not reported	54.31%
	• Loan to value ratio	Not reported	54.31%
	• Public housing	Not reported	54.31%
Education social and community context	<b>Early childhood education and development</b>		
	• Education level	Not reported	54.31%
	<b>Enrollment in higher education</b>		
	• Occupation student	Not reported	54.31%
	• Presence of college graduate	Not reported	54.31%
	<b>High school graduation</b>		
	• Education level	Not reported	54.31%
	<b>Language and literacy</b>		
	• Country of origin	85.80%	Not reported
	• Hispanic language preference	85.80%	Not reported
	• Likes to read	Not reported	54.31%
	<b>Civic participation</b>		
	• Activism social issues	Not reported	54.31%
	• Community civic activities	Not reported	54.31%
	• Charitable volunteer	Not reported	54.31%
	• Registered voter indicator	Not reported	54.31%
	<b>Discrimination</b>	Not reported	Not reported
	<b>Incarceration</b>	Not reported	Not reported
	<b>Social cohesion</b>		
	• Community groups	Not reported	54.31%
• Community and Family	Not reported	54.31%	
• Recreation	Not reported	54.31%	
• Sports	Not reported	54.31%	
• Travel family vacations	Not reported	54.31%	
• Caregiver in home	Not reported	48.89%	
Health and healthcare	<b>Access to healthcare</b>		
	• Insurance	Not reported	54.31%
	• Percent healthcare uninsured	Not reported	54.31%
	• Prescription: number of drugs	Not reported	54.31%
	• Long term care insurance index	Not reported	37.46%

(Continued)

TABLE 2 Continued

SDOH domains	Element category	Compiler 1 coverage	Compiler 2 coverage
	<ul style="list-style-type: none"> <li>• Medicare supplement insurance buyer index</li> <li>• Single service plan vision</li> <li>• Single service plan dental</li> <li>• Single service plan disability</li> </ul>	Not reported	54.31%
	<b>Access to primary care</b>		
	<ul style="list-style-type: none"> <li>• Health rank number of physicians</li> <li>• Health rank doctor visits</li> </ul>	Not reported	48.89%
	<b>Health literacy</b>		
	<ul style="list-style-type: none"> <li>• Reading cooking or culinary</li> <li>• Reading medical or health</li> <li>• Reading natural health remedies</li> </ul>	Not reported	54.31%
<b>Neighborhood and built environment</b>	<b>Access to food that support healthy eating patterns</b>	Not reported	Not reported
	<b>Crime and violence</b>		
	<ul style="list-style-type: none"> <li>• Concealed weapons</li> </ul>	Not reported	54.31%
	<b>Environmental conditions</b>		
	<ul style="list-style-type: none"> <li>• Census percent mobile homes</li> <li>• Census average number of automobiles</li> <li>• Digital neighborhoods</li> </ul>	Not reported	54.31%
	<b>Quality of housing</b>		
	<ul style="list-style-type: none"> <li>• Home market value, estimated</li> <li>• Home building repair</li> </ul>	94.16%	54.31%

based on the HP 2030 framework from the two consumer data sources.

## Clinical data

Clinical data is collected in the EHR then selected elements are imported into the clinical data warehouse (AR-CDR) for research (University of Arkansas for Medical Sciences Translational Research Institute, 2022). The clinical data set used in our work was selected from the AR-CDR and contained over 54,000 individuals with comorbidities linked to high COVID-19 severity (e.g., diabetes and heart disease). The demographic characteristics included name, address, gender (55% female and 45% male), race (59% white, 34% Black or African-American, 6% were missing or unknown, and 1% were either American Indian, Alaskan native, native Hawaiian, Pacific Islander, or Asian), age (45% >65, 49% 31–65, and 6% <30). Clinical features were also collected for this study including diagnosis, procedures, inpatient status, and vital signs. The appended consumer data added 21 features from Compiler 1, and 842 features from Compiler 2 that were manually mapped to existing SDOH concepts as shown in Table 2.

We requested electronic health records from the Clinical Data Repository (University of Arkansas for Medical Sciences Translational Research Institute, 2022) for all patients with chronic conditions (i.e., asthma, diabetes, heart disease, congestive heart failure, coronary artery disease, heart attack, and stroke), or contagious respiratory illness (i.e., influenza or COVID-19) between 2015 and 2020. All data received was stored on one of the following secure devices: institute supported controlled access server, institute supported password protected desktop computer, encrypted password protected laptop. The data used for linking social determinants information was name, address, DOB only. These demographics were transmitted to the selected data compiler vendors *via* SFTP. As per the requirement of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) for protection of patient health information these 3rd party vendors signed a Business Associate Addendum (BAA) with our medical institution prior to accessing the patient identification. Following the addition of the SDOH, the data was de-identified to increase protection of the participants from any negative consequences in the event of a data breach. De-identification was accomplished by deleting full name and address and replacing them with a random identification number and RUCCA code (USDA Economic Research Service, 2020). The DOB was deleted and replaced with age. Data was stored in a secure database server behind a firewall.

## Non-clinical data integration and refresh process

Commercial data is updated monthly and is made up of hundreds of different sources, including consumer surveys, public records, purchase transactions, real estate data, offline and online buying behavior, and warranty information. Wherever possible, compilers compare values from multiple data sources to check accuracy for each element. Vendors that compile data for commercial purposes (i.e., marketing) were identified and interviewed, and then costs were negotiated. SDOH data points were appended by commercial data compilers or other external data sources using the identifiers, patient name, date of birth, and address. These processes occur entirely within a database system using fully HIPAA compliant vendors. All data was encrypted while in transit and immediately destroyed at the compiler location after the completion of the processing. Not all clinical data will be matched to existing consumer data during this process. This is the problem of coverage which refers to the number of patients who could be linked to compiled data. Coverage varies by commercial compiler, but the reasons for coverage variation may also be associated with varying aspects of the patient's lifestyle (e.g., people who use cash exclusively are less likely to have a substantial digital imprint in consumer databases).

Commercial compilers link information about consumers from multiple sources using individual or sets of identifying information like Social Security number, name, address, telephone number, and age or date of birth when available. This is done using proprietary matching algorithms that use statistical and rules-based methodologies. Often, elements are weighted within the algorithm based on their uniqueness within a larger population. In some cases, data elements are missing, making it impossible to link to the commercial data. In this work we have linked at the individual and household level in all cases where it was possible. Typically, studies have linked at the ZIP Code level using public data sources. In [Table 2](#) we can see that Compiler 2 was unable to link almost 50% of the clinical data set while Compiler 1 reported values for fewer elements but had better match rates on those that were reported.

## Data quality analysis

Data quality is a constellation of factors essential for data collections. The quality of the linked clinical and commercial data (henceforth called merged-data) was evaluated for conformance, completeness, consistency, plausibility and temporal alignment ([Kahn et al., 2016](#)). As was mentioned above, accuracy, and consistency, were evaluated by compilers before data was purchased. After data was purchased and linked with EHR data we further measured consistency of common

data elements between compilers from the different sources. In the merged-data there were 5 data elements that were common between the compilers. We matched the data records based on these 5 data elements to measure consistency between the compilers as an added level of data-quality assessment. Each of the data elements differed in categories or levels, between the compilers.

Our first step was to collapse the categories in each element into identical categories so that they could be compared for consistency. In some cases, the element values are inconsistent across clinical and compiler data sources as shown in [Table 3](#). Although we preserved all of the values even if they were inconsistent this problem can be addressed by selecting values that occur most often across multiple sources, by giving preference to sources known to have higher quality data, or by using a gold standard such as direct patient contact. These techniques can be used individually or together depending on the needs of the end-user. For example, the data element "Home Market Value, Estimated" in compiler 1 mapped to "Home Value Range" in compiler 2. The categories of each these feature variables were however not the same. "Home Market Value, Estimated" in compiler 1 had the following 20 categories; "\$1,000–24,999," "\$25,000–49,999," "\$50,000–74,999," "\$75,000–99,999," "\$100,000–124,999," "\$125,000–149,999," "\$150,000–174,999," "\$175,000–199,999," "\$200,000–224,999," "\$225,000–249,999," "\$250,000–274,999," "\$275,000–299,999," "\$300,000–349,999," "\$350,000–399,999," "\$400,000–449,999," "\$450,000–499,999," "\$500,000–749,999," "\$750,000–999,999," "\$1,000,000 Plus," "NA". The "Home Value Range" in compiler 2 had the following 18 categories; "Under \$50 k," "\$50–100 k," "\$100–150 k," "\$150–200 k," "\$200–250 k," "\$250–300 k," "\$300–350 k," "\$350–400 k," "\$400–450 k," "\$450–500 k," "\$500–550 k," "\$550–600 k," "\$600–650 k," "\$650–700 k," "NA," "\$700–750 K," "\$750 K +," "Unknown". In order to assess consistency we first collapsed the categories in each variable in an intuitive and meaningful fashion and made them identical. The new collapsed categories in both variables were: "Less 100 K," "100–200 K," "200–300 K," "300–400 K," "400–500 K," "500 K plus". After discarding the "NAs" or "Unknowns", the consistency between the two variables were calculated.

Missing data are a pervasive problem in any source of data also referred to as data "completeness". Lack of complete data can significantly affect a study outcome by introducing unwanted bias. This is why it was important for us to obtain consumer marketing data from a diverse selection of sources to ensure that we have a collection of data that is complete and deep enough to provide meaningful information about majority of the study participants.

During preprocessing we also examined conformance and plausibility of the data elements, thus comparing the actual format of the data against the expected, and evaluating the feasibility of multiple existing values of the data elements. Measuring the persistence of the data was not possible in this

work since to analyze changes in the data over time multiple batches of consumer data would be needed which was cost-prohibitive.

## Bias

Human bias exists. As we collect, analyze and take actions based on data our biases are perpetuated. This bias pervades healthcare in machine learning, decision support, operations and logistics planning in health systems. Applications to guide clinical practice are not exempt. Real-world clinical data is important for clinical decision-making but it has everyday biases imprinted within it and can preserve or even amplify health disparities. To address this issue requires detection and correction. Sensitive attributes (i.e., race, gender, etc.) are evaluated against classifications to first detect bias that may be present. Once uncovered biased data can be rebalanced using class labels. While this method is not foolproof, biases that are not expected may remain hidden, and it does allow for the mitigation of known issues and for the discovery of unknown issues.

## Data analysis

After the data was linked and preprocessed we explored the data statistical software packages that were linked to the SQL Server hosting the merged-data. We determined the summary statistics of a subset of the variables, from both EHR and compiled SDOH sources, that characterized the patients. We are also currently in the process of building classifiers that can predict disease risk based on patients' clinical, demographic and SDOH factors.

## Results

### Mapping and linking consumer data with EHR at an individual-level

Table 1 lists SDOH domains, elements, and coverage percentages from two consumer data databases that were used in our study and named here as compiler 1 and compiler 2. There were 55,422 patients in the initial set of EHR records. After linking with two commercial compilers, 30,895 and 54,880 patients with SDOH data remained, respectively. As the table shows, compiler 1 had fewer mapped data elements compared to compiler 2. However, compiler 1 had, on an average, ~90% coverage on the mapped data elements while compiler 2 had many more elements mapped but the coverage was much lower, ~54% on an average. Thus, consumer data collection from at least two sources ensured that all categories of SDOH

domains, based on Healthy People 2020 SDOH Framework were covered in our merged-data (Office of Disease Prevention Health Promotion, 2022a). The connection between unlinked patients needs further study to determine what they have in common aside from a minimal individual digital footprint.

## Data quality assessment and preprocessing

To measure consistency between the compiler data collected from different sources we matched the data records based on elements that were common between the compilers as part of data-quality assessment. Table 3 shows the randomly picked data elements and their percent matches. The percent matches of compiler 2 ranged between ~21 and 64% with compiler 1. This underscores the importance of data collected from multiple, trusted and standardized sources.

## Bias

In healthcare as in other areas existing data is sometimes used to power algorithmic prediction. However, when predictions are based on biased information we can proliferate existing biases (Seyyed-Kalantari et al., 2021). Comorbidity indices are common clinical data used for risk adjustment based on patient characteristics (Alonso-Morán et al., 2015) and comorbidity rises with age (Boersma et al., 2020). Yet rural populations in Arkansas are less likely to be diagnosed with multiple chronic conditions as they age than their urban counterparts (Seker et al., 2022). To study this problem of bias in healthcare data, we have measured the amount of bias due to home location and preprocessed a real-world data set of patients with chronic conditions from geographically disparate locations. Bias in data that will be used for modeling can be addressed at one of three different time points: (1) before it is used for modeling, (2) during the modeling process, or (3) after modeling is complete. We have chosen to correct bias as a preprocess because it prepares the data for downstream

TABLE 3 Percent matches between overlapping elements of the two compilers.

Variable name			
Compiler 1	Compiler 2	No of categories	Percent match
Home market value estimated	Home value range	6	55.68%
Marital status	Marital status	2	64.26%
Presence of children	Presence of children	2	41%
Income estimate household	Income range	9	20.85%

TABLE 4 Summary statistics of demographics and individual-level SDOH factors in the merged data with COVID-19 status as indicated by International Classification of Diseases (ICD) codes.

Characteristics	N	COVID-19 positive		p-value <sup>b</sup>
		No, N = 54,330 <sup>a</sup>	Yes, N = 1,092 <sup>a</sup>	
<b>Gender</b>	55,422			0.032
F		29,611 (55%)	638 (58%)	
M		24,706 (45%)	454 (42%)	
U		13 (<0.1%)	0 (0%)	
<b>Age</b>	55,422	62 (18)	46 (18)	<0.001
<b>Race</b>	46,869			<0.001
African American		10,220 (22%)	298 (32%)	
Asian		551 (1.2%)	25 (2.7%)	
Hispanic		1,445 (3.1%)	110 (12%)	
White/other		33,726 (73%)	494 (53%)	
<b>Home owner renter</b>	54,068			<0.001
Home owner		34,558 (65%)	605 (56%)	
Renter		18,420 (35%)	485 (44%)	
<b>Marital status</b>	54,068			<0.001
Married		26,140 (49%)	391 (36%)	
Single		26,838 (51%)	699 (64%)	
<b>Has children</b>	54,068	14,826 (28%)	377 (35%)	<0.001
<b>Member household</b>	54,068			<0.001
1		20,473 (39%)	487 (45%)	
2		15,603 (29%)	224 (21%)	
3		7,772 (15%)	173 (16%)	
4		5,201 (9.8%)	112 (10%)	
5		3,305 (6.2%)	71 (6.5%)	
6		393 (0.7%)	13 (1.2%)	
7		145 (0.3%)	3 (0.3%)	
8		61 (0.1%)	6 (0.6%)	
8 Plus		25 (<0.1%)	1 (<0.1%)	
<b>Estimated family income</b>	34,792			0.4
<30 K		0 (0%)	0 (0%)	
\$30–50 K		14,070 (41%)	271 (43%)	
\$50–75 K		9,228 (27%)	165 (26%)	
\$75–100 K		4,448 (13%)	90 (14%)	
\$100–125 K		2,567 (7.5%)	36 (5.7%)	
\$125 K plus		3,848 (11%)	69 (11%)	
<b>Estimated home market value</b>	51,438			0.9
<100 K		19,231 (38%)	405 (39%)	
100–200 K		21,183 (42%)	434 (42%)	
200–300 K		5,719 (11%)	124 (12%)	
300–400 K		2,153 (4.3%)	39 (3.7%)	
400–500 K		878 (1.7%)	20 (1.9%)	
500 K plus		1,230 (2.4%)	22 (2.1%)	
<b>Economic stability indicator</b>	37,135			<0.001
10–15		7,126 (20%)	117 (15%)	
16–20		7,696 (21%)	118 (15%)	
21–25		10,940 (30%)	226 (28%)	
26–30		10,573 (29%)	339 (42%)	

(Continued)

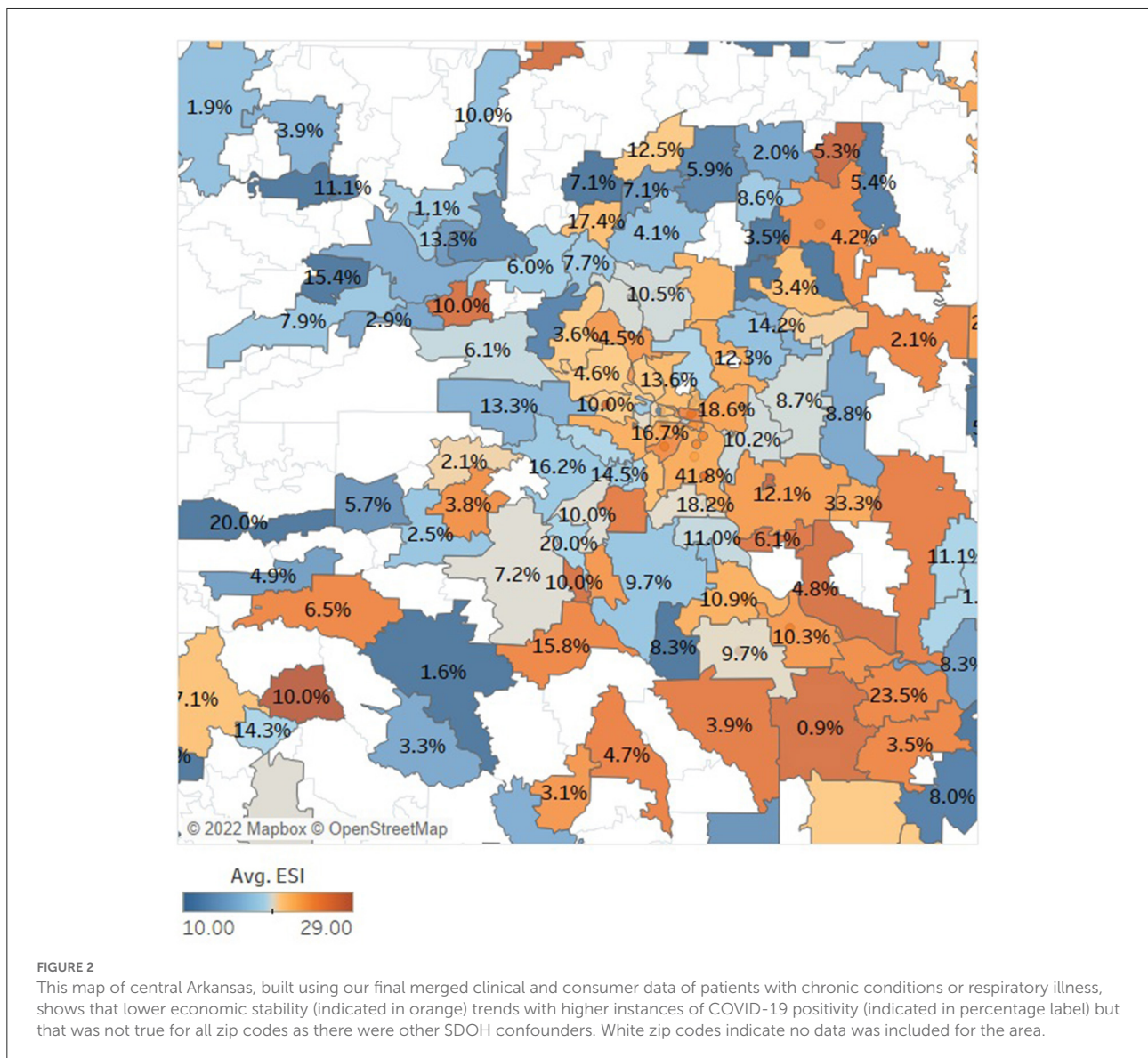


TABLE 4 Continued

Characteristics	N	COVID-19 positive		p-value <sup>b</sup>
		No, N = 54,330 <sup>a</sup>	Yes, N = 1,092 <sup>a</sup>	
<b>Education</b>	34,295			<0.001
Attended vocational/technical		361 (1.1%)	5 (0.7%)	
Completed college		11,392 (34%)	187 (28%)	
Completed graduate school		4,008 (12%)	45 (6.6%)	
Completed high school		17,857 (53%)	440 (65%)	

<sup>a</sup>Statistics presented: n (%); mean (SD).

<sup>b</sup>Statistical tests performed: chi-square test of independence; Wilcoxon rank-sum test.



modeling and does not need to be repeated for each new model. Irrespective of the timepoint when bias processing occurs, it must not damage the integrity of the data and thereby negatively

impact modeling. To monitor this potentiality, we also tested addressing bias by removing the biased data element, the location of the home residence from models. Each of these tests

produced similar AUC results to the original model constructed with unaltered biased data indicating that data preprocessed for bias still performs well when used for modeling as described previously (Seker et al., 2022).

## Preliminary analysis and visualization of the merged data

The SDOH data merged with EHR provided insights into the social risk factors of disease both at patient level as well as the population level. We first explored the disparities in demographics and SDOH factors between the 55,422 patients who were COVID-19 positive compared to those who were not. Table 4 provides a summary of a subset of patient characteristics that were compared between these two groups. Our preliminary analysis showed that apart from demographic factors, several SDOH factors like home-ownership, marital-status, presence of children, number of members per household, Economic Stability Indicator (ESI) and education were significantly different between the two patient groups while estimated family-income and home market-value were not. Figure 2 shows a map constructed with the merged clinical and consumer data. In this figure the top COVID-affected zip codes in Arkansas are overlaid on a heatmap image of average ESI of those zip codes. ESI is a proprietary consumer data element provided by compiler 1 and shown in Tables 2, 4. It is commonly used in marketing which is constructed to function like a credit score but generated without credit data. The darker orange counties reflected patients from our data that were less economically stable and also had higher percentages of COVID-19 positive patients.

## Discussion

In this proof-of-concept study our main objective was to evaluate the viability of consumer marketing data, purchased from 3rd party HIPPA-compliant vendors, as a source of SDOH factors that are largely missing from the EHR data. We purchased in-depth patient-level consumer data from two different vendors, mapped a wide array of data elements to the 5 broad SDOH domains as defined by the Healthy People 2030 SDOH framework, linked the consumer data to EHR patient level data, stored the data in a SQL server linked with several statistical and data exploration tools, evaluated data quality and preprocessed the data, and lastly completed a preliminary analysis of a subset of the SDOH elements that characterized the patients. To our knowledge this is the first study to explore the viability of consumer marketing data as a source of patient-level SDOH data.

With recent upsurge in research solidifying the significant relationship between SDOH and population health, an increasing number of healthcare stakeholders are exploring

the use of public databases for community-level information in order to identify those patients that are most vulnerable to SDOH. For example, census tracts data have been used to identify areas associated with socio-economic risks and poor health outcomes (Liaw et al., 2018). But a recent study by Cottrell et al. (2020) showed that only about 48% of the times community-level data can accurately identify social risks at the patient level. Thus, healthcare decisions on individual patients based on community-level data may fall short on providing adequate care to a significant number of patients. This may give rise to the problem of “ecologic fallacy” where incorrect assumptions can be made about a patient based on aggregate-level information from community-level data (Garg et al., 2016; Cottrell et al., 2020). In this study we have attempted to address this problem by partnering with companies/vendors that are honed consumer market researchers.

We have developed a repeatable process to incorporate commercially compiled data into EHR data. The added value has been demonstrated based on a published paper (Greer et al., 2021). We have also identified opportunities in data quality research areas that need further study as part of this work. The curated data are being used to support several healthcare analytics applications, including descriptive analytics, and predictive modeling. During this work, we have developed the first mapping scheme of commercial data elements with SDOH elements. This is a fascinating aspect of this work because the healthcare community has not reached a consensus on a standard set of social determinants of health concepts demanding that this process be agile and flexible.

While building an enrichment process for EHR data we had to address important issues related to temporal alignment, data dictionary and coverage, legal requirements, and security requirements. Initially, the legal, research and business processes required were complex and time-consuming. Patient data must be kept private and secure at all times, and all parties must be bound by a contract to minimize the possibility of a data breach. Fortunately, once contracts and transfer processes are in place, they remain active and available for repeated consumer data collection. This is important because continued collection will be necessary. Compiled data becomes stale over time, and EHR data is collected only at the time of each encounter. Aligning these time windows is necessary for elements that must be current while is less critical for elements that are more likely to remain stable over time. As the process iterates and newly compiled data is integrated into the EHR data, the data dictionary must also be updated. We discovered that the data dictionaries provided by compilers vary in quality and detail. In addition, compilers are continuously adding, removing, and updating elements resulting in multiple versions of the dictionary documentation. Integrating data from these dynamic systems also impacted the mapping of compiled elements onto SDOH concepts, resulting in a mapping component for each iteration. If kept current, the SDOH mapping will require minimal effort to maintain. Throughout all of these components, it was also necessary to

tackle practical information technology issues such as storage, tools, permissions, and access which will need to be customized to each institution.

Our study had several limitations. Due to budgetary constraints we restricted our SDOH data sources to two different vendors only. One had more mapped data elements while the other had more coverage, thus highlighting the need for data collection from multiple, trust worthy and reputable vendors with standardized methods of data collection. There were significant missing data in each data set. Also, among the overlapping data elements, the concordance between the data was not very high which underscores the need for good quality data sources.

In conclusion, we have developed a repeatable SDOH enhancement process to incorporate dynamically evolving SDOH domain concepts from consumers into clinical data. The literature provides early and rapidly growing evidence that integrating individual-level SDOH into EHRs can assist in risk assessment and predicting healthcare utilization and health outcomes, which further motivates efforts to collect and standardize patient-level SDOH information. This study highlights one potential means to incorporate individual-level patient data into EHR, thus opening up possibilities for predictive analytics and enhanced solutions for providers, payers and healthcare organizations to enable them to address the social needs of patients.

## Data availability statement

The datasets presented in this article are not readily available because it contains clinical information. Requests to access the datasets should be directed to [mlgreer@uams.edu](mailto:mlgreer@uams.edu).

## Ethics statement

The studies involving human participants were reviewed and approved by University of Arkansas for Medical Sciences IRB. Written informed consent for participation was not required for

this study in accordance with the national legislation and the institutional requirements.

## Author contributions

MG developed the concept. MG and SB worked on concept development analysis and manuscript writing. CZ performed SDOH mapping and writing. All authors contributed to the article and approved the submitted version.

## Funding

The research reported in this publication was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Numbers TL1 TR003109 and UL1 TR003107.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Author disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

- Arcaraz, K. I., Wiedt, T. L., Daniels, E. C., Yabroff, K. R., Guerra, C. E., Wender, R. C., et al. (2020). Understanding and addressing social determinants to advance cancer health equity in the United States: a blueprint for practice, research, and policy. *CA Cancer J. Clin.* 70, 31–46. doi: 10.3322/caac.21586
- Alegria, M., NeMoyer, A., Falgàs Bagué, I., Wang, Y., and Alvarez, K. (2018). Social determinants of mental health: where we are and where we need to go. *Curr. Psychiatry Rep.* 20, 1–13. doi: 10.1007/s11920-018-0969-9
- Alonso-Morán, E., Nuño-Solinis, R., Onder, G., and Tonnara, G. (2015). Multimorbidity in risk stratification tools to predict negative outcomes in adult population. *Eur. J. Internal Med.* 26, 182–189. doi: 10.1016/j.ejim.2015.02.010
- Ancker, J. S., Kim, M.-H., Zhang, Y., Zhang, Y., and Pathak, J. (2018). The potential value of social determinants of health in predicting health outcomes. *J. Am. Med. Inform. Assoc.* 25, 1109–1110. doi: 10.1093/jamia/ocy061
- Boersma, P., Black, L. I., and Ward, B. W. (2020). Peer reviewed: prevalence of multiple chronic conditions among US adults, 2018. *Prevent. Chronic Dis.* 17, 130. doi: 10.5888/pcd17.200130
- Cantor, M. N., and Thorpe, L. (2018). Integrating data on social determinants of health into electronic health records. *Health Affairs* 37, 585–590. doi: 10.1377/hlthaff.2017.1252

Centers for Disease Control and Prevention (2015). *Planning for an Emergency: Strategies for Identifying and Engaging At-risk Groups. A Guidance Document for Emergency Managers*. Atlanta: Centers for Disease Control and Prevention.

Chamberlain, A. M., Finney Rutten, L. J., Wilson, P. M., Fan, C., Boyd, C. M., Jacobson, D. J., et al. (2020). Neighborhood socioeconomic disadvantage is associated with multimorbidity in a geographically-defined community. *BMC Public Health* 20, 13. doi: 10.1186/s12889-019-8123-0

Cottrell, E. K., Hendricks, M., Dambrun, K., Cowburn, S., Pantell, M., Gold, R., et al. (2020). Comparison of community-level and patient-level social risk data in a network of community health centers. *JAMA Network Open* 3, e2016852. doi: 10.1001/jamanetworkopen.2020.16852

Feldman, S. S., Davlyatov, G., and Hall, A. G. (2020). Toward understanding the value of missing social determinants of health data in care transition planning. *Appl. Clin. Inform.* 11, 556–563. doi: 10.1055/s-0040-1715650

Fraze, T. K., Brewster, A. L., Lewis, V. A., Beidler, L. B., Murray, G. F., Colla, C. H., et al. (2019). Prevalence of screening for food insecurity, housing instability, utility needs, transportation needs, and interpersonal violence by US physician practices and hospitals. *JAMA Network Open* 2, e1911514–e1911514. doi: 10.1001/jamanetworkopen.2019.11514

Galea, S., and Vlahov, D. (2002). Social determinants and the health of drug users: socioeconomic status, homelessness, and incarceration. *Public Health Rep.* 117, S135.

Garg, A., Boynton-Jarrett, R., and Dworkin, P. H. (2016). Avoiding the unintended consequences of screening for social determinants of health. *JAMA* 316, 813–814. doi: 10.1001/jama.2016.9282

Gold, R., Bunce, A., Cowburn, S., Dambrun, K., Dearing, M., Middendorf, M., et al. (2018). Adoption of social determinants of health EHR tools by community health centers. *Ann. Fam. Med.* 16, 399–407. doi: 10.1370/afm.2275

Grandner, M. A., and Fernandez, F.-X. (2021). The translational neuroscience of sleep: a contextual framework. *Science* 374, 568–573. doi: 10.1126/science.abj8188

Greer, M. L., Sample, S., Jensen, H. K., McBain, S., Lipschitz, R., Sexton, K. W., et al. (2021). COVID-19 is connected with lower health literacy in rural areas. *Stud. Health Technol. Inform.* 281, 804. doi: 10.3233/SHTI210286

Hatef, E., Rouhizadeh, M., Tia, I., Lasser, E., Hill-Briggs, F., Marsteller, J., et al. (2019). Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Med. Inform.* 7, e13802. doi: 10.2196/13802

Johnson-Lawrence, V., Zajacova, A., and Sneed, R. (2017). Education, race/ethnicity, and multimorbidity among adults aged 30–64 in the National Health Interview Survey. *SSM-Popul. Health* 3, 366–372. doi: 10.1016/j.ssmph.2017.03.007

Kahn, M., Callahan, T., Barnard, J., Bauck, A., Brown, J., Davidson, B., et al. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS* 4, 1244. doi: 10.13063/2327-9214.1244

Kind, A. J., and Buckingham, W. R. (2018). Making neighborhood-disadvantage metrics accessible—the neighborhood atlas. *N. Engl. J. Med.* 378, 2456. doi: 10.1056/NEJMp1802313

Kivimäki, M., Batty, G. D., Pentti, J., Shipley, M. J., Sipilä, P. N., Nyberg, S. T., et al. (2020). Association between socioeconomic status and the development of mental and physical health conditions in adulthood: a multi-cohort study. *Lancet Public Health* 5, e140–e149. doi: 10.1016/S2468-2667(19)30248-8

Liaw, W., Krist, A. H., Tong, S. T., Sabo, R., Hochheimer, C., Rankin, J., et al. (2018). Living in “cold spot” communities is associated with poor health and health quality. *J. Am. Board Fam. Med.* 31, 342–350. doi: 10.3122/jabfm.2018.03.170421

Nicholas, L. H., Langa, K. M., Bynum, J. P., and Hsu, J. W. (2021). Financial presentation of Alzheimer disease and related dementias. *JAMA Internal Med.* 181, 220–227. doi: 10.1001/jamainternmed.2020.6432

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

Office of Disease Prevention and Health Promotion (2022a). *Healthy People 2030 Framework, Health People 2030*. Washington, DC: U.S. Department of Health and Human Services.

Office of Disease Prevention and Health Promotion (2022b). *Social Determinants of Health, Health People 2030*. Washington, DC: U.S. Department of Health and Human Services.

Rural Health Information Hub (2022). *Frameworks to Address Social Determinants of Health*. North Bethesda: Rural Health Information Hub.

Seker, E., Talburt, J. R., and Greer, M. L. (2022). “Preprocessing to address bias in healthcare data,” in *Accepted for Publication in the Medical Informatics Europe (MIE) 2022 Conference Proceedings*.

Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I. Y., and Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* 27, 2176–2182. doi: 10.1038/s41591-021-01595-0

Tamura, K., Langerman, S. D., Ceasar, J. N., Andrews, M. R., Agrawal, M., Powell-Wiley, T. M., et al. (2019). Neighborhood social environment and cardiovascular disease risk. *Curr. Cardiovasc. Risk Rep.* 13, 1–13. doi: 10.1007/s12170-019-0601-5

Tong, S. T., Liaw, W. R., Kashiri, P. L., Pecsok, J., Rozman, J., Bazemore, A. W., et al. (2018). Clinician experiences with screening for social needs in primary care. *J. Am. Board Fam. Med.* 31, 351–363. doi: 10.3122/jabfm.2018.03.170419

Tung, E. L., Wroblewski, K. E., Boyd, K., Makelarski, J. A., Peek, M. E., Lindau, S. T., et al. (2018). Police-recorded crime and disparities in obesity and blood pressure status in Chicago. *J. Am. Heart Assoc.* 7, e008030. doi: 10.1161/JAHA.117.008030

University of Arkansas for Medical Sciences Translational Research Institute (2022). *Clinical Data Repository, (AR-CDR)*. Little Rock: University of Arkansas for Medical Sciences Translational Research Institute.

USDA Economic Research Service (2020). *Rural-Urban Commuting Area Codes, Economic Research Service*. Washington, DC: U.S. Department of Agriculture.

World Health Organization (2010). *A Conceptual Framework for Action on the Social Determinants of Health, Social Determinants of Health Discussion Paper 2*. Geneva: World Health Organization.