Journal of the American Medical Informatics Association, 26(4), 2019, 286–293
doi: 10.1093/jamia/ocy172
Advance Access Publication Date: 31 January 2019
Research and Applications

Research and Applications

Incorporating a location-based socioeconomic index into a de-identified i2b2 clinical data warehouse

Bret J. Gardner,[1] Jay G. Pedersen,[2] Mary E. Campbell,[3] and James C. McClay[1]

[1]Department of Emergency Medicine, University of Nebraska Medical Center, Omaha, Nebraska, USA, [2]Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, Nebraska, USA, and [3]Department of Sociology, Texas A&M University, College Station, Texas, USA

Corresponding Author: Bret J. Gardner, PhD, 987810 NE Medical Center, Omaha, NE 68198-1150, USA (bret.gardner@unmc.edu)

Received 26 January 2018; Revised 29 September 2018; Editorial Decision 19 November 2018; Accepted 27 November 2018

ABSTRACT

Objective: Clinical research data warehouses are largely populated from information extracted from electronic health records (EHRs). While these data provide information about a patient's medications, laboratory results, diagnoses, and history, her social, economic, and environmental determinants of health are also major contributing factors in readmission, morbidity, and mortality and are often absent or unstructured in the EHR. Details about a patient's socioeconomic status may be found in the U.S. census. To facilitate researching the impacts of socioeconomic status on health outcomes, clinical and socioeconomic data must be linked in a repository in a fashion that supports seamless interrogation of these diverse data elements. This study demonstrates a method for linking clinical and location-based data and querying these data in a de-identified data warehouse using Informatics for Integrating Biology and the Bedside.

Materials and Methods: Patient data were extracted from the EHR at Nebraska Medicine. Socioeconomic variables originated from the 2011-2015 five-year block group estimates from the American Community Survey. Data querying was performed using Informatics for Integrating Biology and the Bedside. All location-based data were truncated to prevent identification of a location with a population <20 000 individuals.

Results: We successfully linked location-based and clinical data in a de-identified data warehouse and demonstrated its utility with a sample use case.

Discussion: With location-based data available for querying, research investigating the impact of socioeconomic context on health outcomes is possible. Efforts to improve geocoding can readily be incorporated into this model.

Conclusion: This study demonstrates a means for incorporating and querying census data in a de-identified clinical data warehouse.

Key words: social determinants of health, i2b2, census, American Community Survey (ACS), socioeconomic status

BACKGROUND AND SIGNIFICANCE

Clinical research data warehouses are often populated with de-identified patient data extracted from an electronic health record (EHR). With the continuing advancement and adoption of EHRs, the amount of information available for reuse in clinical research continues to rise.[1–4] However, for complete patient characterization, these data need to be linked to other sources. For instance, while a patient's race, gender, and smoking status are often well-documented in the EHR, other elements of socioeconomic status (SES) and a patient's social context are often unstructured or absent from the clinical record and unavailable for incorporation into a research data warehouse. These nonclinical elements describing a

patient's social, economic, and environmental determinants of health (healthypeople.gov) are a major contributing factor in readmission, morbidity, and mortality[5] and EHRs provide an important opportunity to integrate such data into research.[6]

With the paucity of data in the EHR related to SES, researchers have relied on insurance type as a proxy for this measure.[7] Insurance type has been demonstrated to be potentially related to area deprivation, however, these estimates are not synonymous.[8] Additionally, state and gender differences may exist in Medicaid and other insurance coverage, hampering studies crossing state lines. Data elements related to a patient's neighborhood SES may reliably be obtained from extra-EHR sources such as American Community Survey (ACS) data from the U.S. Census Bureau. Neighborhood resources have robust effects on health[9–12] because of their correlation with individual SES and as an independent source of influence. The demographic composition of residential areas also has important links to health behaviors and health outcomes.[13–16] Linking measures of the local residential context to clinical data from the EHR can provide insights into these socioeconomic and demographic correlates of health for researchers.

Using geographic information system (GIS) software, a patient's physical address can be linked to a variety of location-based datasets such as the Environmental Protection Agency's Air Quality System[17] or the ACS (https://www.census.gov/programs-surveys/acs/). While efforts are being made to integrate these elements directly into the EHR, to date, no EHR has demonstrated widespread integration of such "community vital signs."[18] Implementing this linkage for clinical research with an EHR-agnostic approach introduces additional challenges related to patient privacy and data standardization. The Health Information Portability and Accountability Act (HIPAA) privacy rule, the Health Information Technology for Economic and Clinical Health Act of 2009, and the Federal Policy for the Protection of Human Subjects are designed to safeguard protected health information (PHI), including street address and geocodes.[19] These safeguards may prohibit researchers from sharing the information required for geocoding with an academic or business partner third party, hampering the ability to link clinical and SES data within an institutional review board (IRB)–approved process. Additionally, data that would identify a patient's location with too much granularity may not be displayed in a de-identified data warehouse. For instance, HIPAA requires zip codes be obfuscated if the population is below 20 000 for that area. Many details are available from the ACS for significantly smaller populations, requiring obfuscation before being made available in a de-identified system.

One approach to maximize sample population while maintaining patient privacy is to participate in a distributed research network. PCORnet (National Patient-Centered Clinical Research Network) and its participating Clinical Data Research Networks illustrate how patient data may be stored locally and federated queries may be shared across the network.[20–22] However, interoperability of research queries across a CDRN is challenging, as variability may be introduced into a collaborative study if geocoding is performed independently at each site with disparate methods.[23–25] This variability may affect analysis and conclusions in healthcare studies.[26] As clinical and socioeconomic data are linked, successful collaboration and data analysis is dependent on a means of querying these data from each site for a variety of studies.

OBJECTIVE

In this manuscript, we provide a model for combining socioeconomic and clinical data while maintaining patient privacy and allowing rapid querying in a de-identified data warehouse. We describe an algorithm for extracting neighborhood socioeconomic data from the ACS, geocoding patient data without involving a third party, and combining these data within an Informatics for Integrating Biology and the Bedside (i2b2) framework for interrogation. Due to the volume and variety of data within the ACS, we extracted only elements to calculate a validated socioeconomic index,[5] similar to many neighborhood SES indices. We demonstrate not only the data integration, but, also the metadata development requisite to allow querying of the data within the i2b2 framework. The utility of the resulting data in the i2b2 context is demonstrated with a sample use case evaluating the correlation of SES and emergency department (ED) utilization previously described in the literature.[27] The approach we describe may be fully deployed at other sites and will allow for collaborative research and federated queries while keeping PHI secure.[28]

MATERIALS AND METHODS

Clinical data

Patient data were extracted from a research copy of the EHR data warehouse at the University of Nebraska Medical Center. This system contains data originating from multiple hospitals and clinics in urban, suburban, and rural Nebraska. Clinical and demographic data are extracted, standardized based on Office of the National Coordinator–recommended vocabularies, and transformed for use within an i2b2 clinical data warehouse.[29,30] Data are transformed and staged on an identified server and then de-identified and made accessible to researchers via i2b2 in a fully de-identified database on a separate server (Figure 1). This data extraction and use in a de-identified data warehouse was approved by the IRB at the University of Nebraska Medical Center (IRB #132-14-EP).

Geocoding process

Current and historic patient address information was extracted from the EHR and stored on a secure server. TIGER/Line Shapefiles and other location-based files needed for geocoding were obtained via File Transfer Protocol from the U.S. Census Bureau and loaded onto the server alongside patient address data. These files were for year 2017 and for the states of Nebraska and Iowa. Using these data, PostGIS version 2.4[31] geocoding software running on PostgreSQL version 10.1 (https://www.postgresql.org/) was used to identify the longitude and latitude for each patient address. Subsequently, the U.S. Census block group for each successfully geocoded address was determined via PostGIS. For this study, we geocoded only patient addresses for Nebraska and Iowa. In the extraction, we eliminated post office box addresses and those addresses with null or invalid street addresses (Figure 2). Invalid street addresses were defined as those consisting of all alpha or all numeric characters, consisting of a single character, or containing variants of the words *unknown* or *invalid*.

We compared demographic data for the geocoded population relative to those patients we excluded from analysis. Rural versus urban location was based on U.S. Department of Agriculture Rural Urban Commuting Area (RUCA) codes mapped to zip codes.[32,33] Because no direct measure of the patient's financial class was available, we used the primary insurance category listed on the patient's account in the EHR as a proxy. Age was calculated based on the date of the data extract (December 2017).

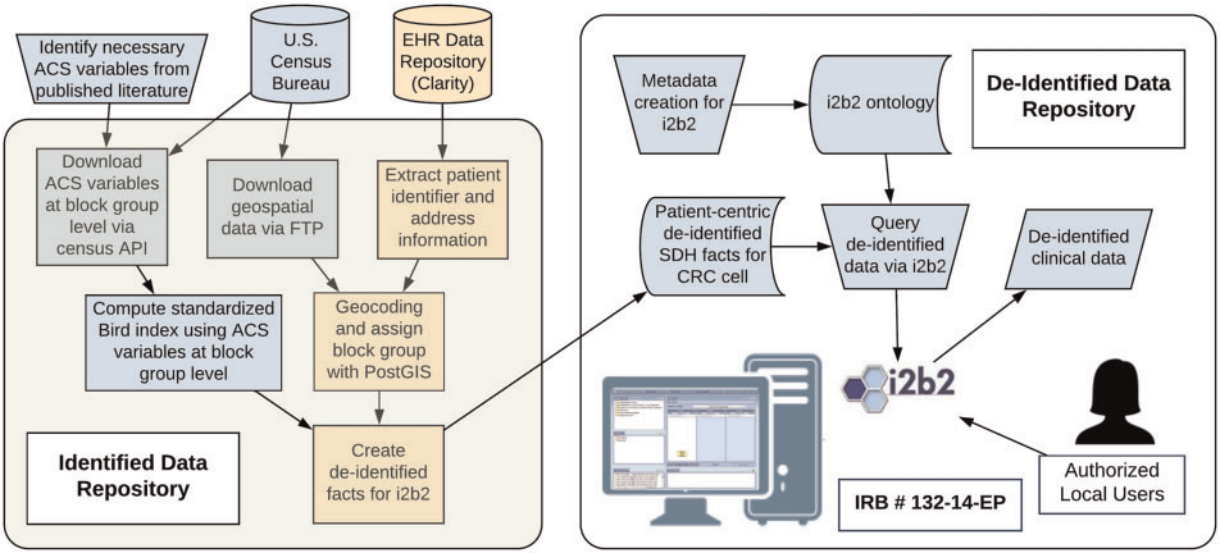

Figure 1. Overview of integration of clinical and location-based data. Patient data was extracted from Nebraska Medicine's electronic health record (EHR). Location-based data for geocoding and for socioeconomic variables were obtained from the U.S. Census Bureau. All identified data were stored on a secure server. De-identified data and the Informatics for Integrating Biology and the Bedside (i2b2) platform to allow researcher querying of the data were housed on a separate, de-identified server. ACS: American Community Survey; API: application program interface; CRC: Clinical Research Chart or Data Repository Cell; FTP: File Transfer Protocol; IRB: institutional review board; SDH: Social Determinants of Health.

Census variable extraction and socioeconomic index calculation

Many similar contextual measures of socioeconomics are available. We selected a parsimonious validated index (1) with all variables available at the block group level of geography, (2) that did not include the demographic makeup of the local area (to keep the racial composition of the neighborhood conceptually distinct from the socioeconomic characteristics), and (3) that was validated on a nationally representative sample of adults. This method could be used with a wide variety of similar indices with little modification. Many indices have been tested (see Messer et al[34] for an example of the range of contextual variables in use) that cover similar domains to those in the Bird index (education, employment, poverty, public assistance, family structure, and income).

Using the index and variables described by Bird et al,[5] a set of equivalent variables which could be computed from the U.S. Census Bureau's annual ACS were identified. Table 1 identifies the Bird variables from the 2000 decennial census and the field and computation employed from the 2011-2015 five-year estimates from the ACS. Using the U.S. Census Bureau application programming interface, ACS estimates for each variable for each block group in the United States were extracted and stored locally. These raw estimates were transformed and normalized for all Nebraska and Iowa block groups to have a mean of 0 and an SD of 1 as described by Bird et al.[5] The Bird index is computed as the sum of the standardized values for each of the 6 variables, where the standardized values are multiplied by −1 for variables in which a higher positive value indicates lower SES. This method results in higher Bird index values corresponding to a higher SES. There were 88 missing values across the 6 variables included in the scale. Missing values on any of the 6 variables were imputed in Stata 15 (StataCorp, College Station, TX), using an imputation model that predicts the value of the missing variable based on the other, nonmissing values. We did not impute values for the 2 block groups that had missing values for all 6 variables and an estimated population reported as 0. Standardized values for the 6 variables as well as the Bird index were stored for all block groups within Nebraska and Iowa.

Identifying patient SES and i2b2 fact creation

Patient addresses from Nebraska and Iowa were linked to census block groups.[35] Patient identifiers were linked to the Bird index and standardized ACS variables associated with their block group. Data were de-identified and loaded into the database on the de-identified server used by i2b2. De-identification included using a randomly generated patient number, shifting all dates for each patient randomly by −1 to −365 days, and excluding any HIPAA identifiers. For each block group, we used the block group estimated population reported in the ACS to ensure the level of granularity of the reported SES index could not be used to identify a population smaller than 20 000 individuals. For each patient, 7 records were inserted into the database: the Bird index and the 6 standardized variables necessary for its computation.

Metadata creation and i2b2 querying

The demographics portion of the i2b2 ontology cell was updated to support interrogation of ACS-based facts (Figure 3). A folder for neighborhood SES was added with subfolders for the Bird SES index and the 6, standardized variables of interest. The c_basecode in the metadata table used to reference rows in the fact table was based on the field name specified in the ACS for each variable. The c_name and c_tooltip, specifying what will be visible in the user interface, contained both a human readable description and a standardized reference for each ACS variable. For the Bird index and each component variable, metadata XML was created.

Example use case

ED Utilization: All patients with a computed Bird neighborhood SES index who were seen in any Nebraska Medicine affiliated hospital or clinic between January 2013 and December 2017 were

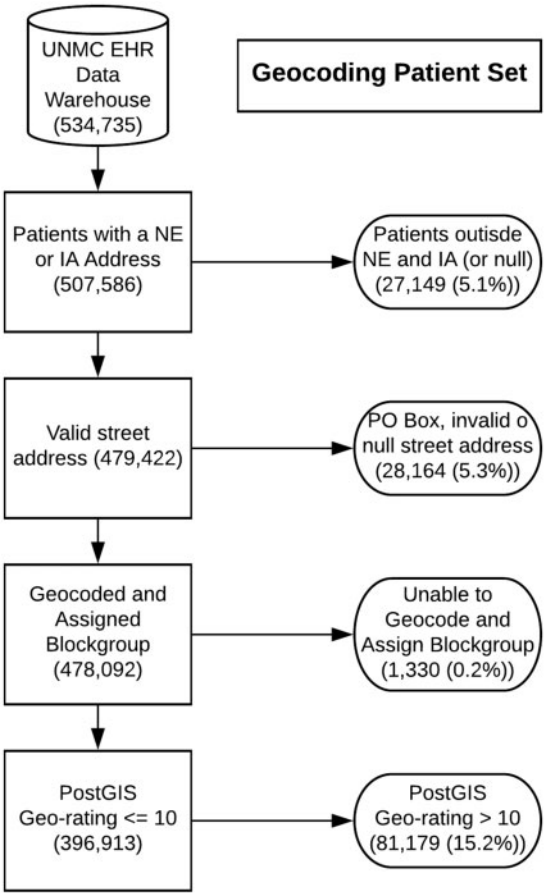

Figure 2. Overview of patients with successfully geocoded addresses linked to a census block group. Patients with missing or invalid addresses, post office (PO) boxes, and addresses that could not be geocoded with high confidence were excluded. EHR: electronic health record; IA: Iowa; NE: Nebraska; UNMC: University of Nebraska Medical Center.

Table 1. Description of neighborhood socioeconomic status computations

Description	ACS Variable	Computation
Percent of adults older than 25 with less than a high school education	B15003	$\Sigma_{i=2}^{16} \frac{\text{HD01 VD}_i}{\text{HD01\_VD01}}$
Percent male unemployment	B23022	$\frac{\text{HD01 VD26}}{\text{HD01\_VD02}}$
Percent of households with income below the poverty line	B17017	$\frac{\text{HD01 VD02}}{\text{HD01\_VD01}}$
Percent of households receiving public assistance	B19057	$\frac{\text{HD01 VD02}}{\text{HD01\_VD01}}$
Percent of female-headed households with children	B23007	$\frac{\text{HD01 Vd}_i}{\text{HD01\_VD01}}$
Median household income	B19013	HD01_VD01

Note: The variables reported by Bird et al were mapped to the American Community Survey (ACS). Variables to compute the Bird index were calculated from the ACS 5-year estimates at the block group level

identified. We estimated a logit model where visiting an ED during this time was the dependent variable (compared with 0 visits) and the key independent variable was the blurred Bird neighborhood

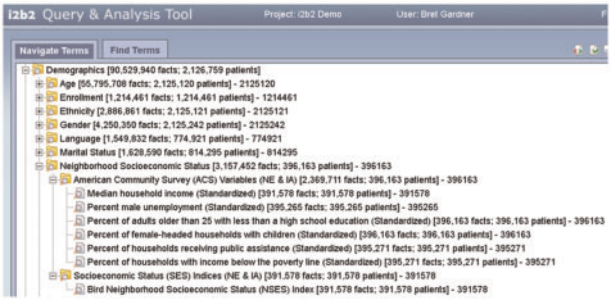

Figure 3. Incorporation of American Community Survey socioeconomic variables and summary index into the demographics folder of an Informatics for Integrating Biology and the Bedside (i2b2) hierarchy.

SES index. ED visits included encounters with any resulting discharge disposition, including hospital admission or expiration. The model was estimated in Stata 15, and controlled for the age, race, gender, and insurance type of the patients, as well as how rural their residential area was. In addition, for patients with at least 1 ED encounter, we estimated a negative binomial regression of the number of visits, also controlling for the same variables. A total of 64 patients were dropped from the analyses due to missing data on gender.

RESULTS

Metadata creation

The demographics folder of the i2b2 hierarchy was updated to allow for integration of SES variables and indices. A subfolder for ACS variables was added. This currently contains the 6 variables necessary to compute the Bird SES index. These variables are referenced both by a human readable description and the ACS variable reference to allow for interoperability. A second subfolder was added for SES. This currently contains only the Bird index. Each of these folders may be expanded in the future to allow for integration of additional variables and further indices. For all included items in the i2b2 hierarchy, metadata XML was developed and deployed to allow users to specify values or ranges of interest to query.

Geocoding

We geocoded only patients with a Nebraska or Iowa address, representing the majority (507 586 of 534 735 [94.9%]) of Nebraska Medicine patients. Patients were excluded from analysis if (1) The patient address failed to geocode or have a block group assigned (1330), (2) the patient street address was unknown or invalid (28 893), or (3) the geo-rating assigned by PostGIS was >10, indicating a low confidence in the geocode assignment (81 179). The final geocoded population consisted of 396 913 patients who have had an encounter at Nebraska Medicine, a well-geocoded address, and a block group assigned (Figure 2).

Table 2 illustrates the comparison of the included versus the excluded population for analysis. The percentage of the excluded population living in a rural zip code was 23.7% compared with only 9.8% of the included population. The racial composition of the included and excluded populations were very similar, with each demonstrating a majority white (78.9% and 80.8% of included and excluded populations, respectively) with a lower percentage of black (9.5% and 7.6% of included and excluded populations, respectively) and other races (11.7% and 11.6% of included and excluded populations, respectively) in both populations. The populations

Table 2. Comparison of included and excluded populations

		Included	%	Excluded	%
Sex	Male	179 750	45.3	52 675	47.3
	Female	217 098	54.7	58 677	52.7
	Unknown	65	0.0	50	0.0
Race	White	312 964	78.8	90 061	80.8
	Black	37 599	9.5	8425	7.6
	Other/Unknown	46 350	11.7	12 916	11.6
Age, y	0-18	72 167	18.2	17 794	16.0
	19-34	92 750	23.4	24 372	21.9
	35-49	73 325	18.5	19 318	17.3
	50-64	78 767	19.8	23 298	20.9
	65-79	55 955	14.1	18 300	16.4
	80+	23 949	6.0	8320	7.5
Financial Class	Private/commercial	190 509	48.1	50 529	45.4
	Medicare	65 246	16.5	22 349	20.1
	Medicaid	38 219	9.6	9626	8.7
	Self-pay/none	66 343	16.7	20 708	18.6
	Other/unknown	35 846	9.0	7997	7.2
Rurality	Urban (RUCA 1-6)	358 007	90.2	84 707	76.0
	Rural (RUCA 7-10)	38 900	9.8	26 392	23.7
	No zip code	6	0.0	392	0.4
Total		396 913		111 402	

Note: Data represent the demographic characteristics for patients with successfully geocoded addresses relative to those excluded as described in Figure 2.
RUCA: Rural Urban Commuting Area.

demonstrated little difference in gender proportion (54.7% and 52.7% women [inclusion and exclusion populations, respectively]). The included population had a higher percentage of private or commercial insurance (48.09% vs 45.44% excluded population) while having a slightly lower Medicare percentage (16.5% vs 20.1% excluded population). The included population had a lower age relative to the excluded population (mean 42.1 ± 23.6 years vs 44.6 ± 24.0 years).

Neighborhood SES variables

We computed a Bird index for 4261 of 4263 (99.96%) of all block groups in Nebraska and Iowa. The 2 block groups dropped from analysis had no data for any of the variables reported as well as having a population estimate reported as 0 within the ACS.

ED utilization

Using the i2b2 data warehouse, 360 947 patients were identified as being seen in any Nebraska Medicine hospital or clinic between January 2013 and December 2017 who also had an assigned Bird index. The first model in Table 3 shows that, controlling for the patient's age, gender, race, insurance status, and rurality, each unit increase in the Bird neighborhood SES index is associated with a 13% decline in their odds of visiting an ED during this time. Figure 4 shows how the predicted probability of visiting an ED declines as the SES index rises from the 10th percentile to the 90th percentile, holding all other variables in the model at their means.

For the 78 990 patients who visited an ED during this time, the number of visits is also significantly associated with neighborhood SES, as we see in the second model in Table 3. The negative binomial model estimates the count of visits, and shows that higher neighborhood SES is significantly associated with fewer ED visits.

Table 3. Results of logistic regression analysis for all patients and negative binomial analysis for all patients visiting the ED

	Logit: ED visit Odds ratio	Negative binomial Number of ED visits
Blurred bird neighborhood SES	0.874*	−0.063*
	(0.003)	(0.003)
Female	1.096*	0.101*
	(0.009)	(0.006)
Age in years	1.022*	0.018*
	(0.001)	(0.001)
Age^2	1.000*	−0.000*
	(0.000)	(0.000)
White (reference)		
Black	1.684*	0.064*
	(0.022)	(0.009)
Any other race	0.756*	−0.093*
	(0.011)	(0.011)
Private insurance (reference)		
Medicaid	2.200*	0.379*
	(0.033)	(0.010)
Medicare	1.695*	0.378*
	(0.025)	(0.010)
Other insurance	1.927*	0.129*
	(0.033)	(0.013)
No insurance/self-pay	1.465*	0.217*
	(0.017)	(0.009)
Rurality scale	0.731*	−0.080*
	(0.002)	(0.003)
Constant	0.180*	0.631*
	(0.003)	(0.014)
Observations	395 518	78 990

Note: Standard error values are listed in parentheses.
ED: emergency department; SES: socioeconomic status.
*$P < .01$.

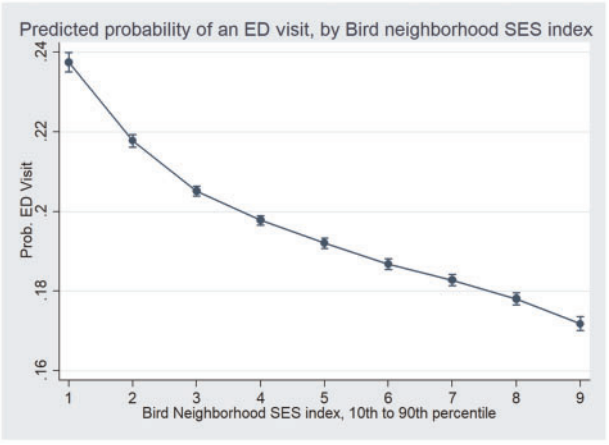

Figure 4. Logistic regression model demonstrating the predicted probability of visiting the emergency department (ED). The probability of visiting the ED increases as the socioeconomic status (SES) decreases, holding race, age, gender, insurance type, and rurality constant.

Figure 5 shows how the predicted number of visits (again holding all other factors in the model constant) declines as the SES index rises from the 10th to the 90th percentile.

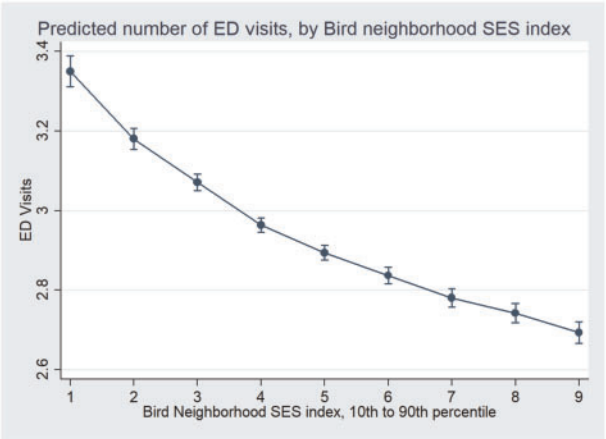

Figure 5. Negative binomial analysis of number of emergency department (ED) visits. For all patients who visited the ED, the frequency of visits increases as socioeconomic status (SES) decreases while holding age, gender, race, insurance type, and rurality constant.

DISCUSSION

Limitations

This study is limited by only including patients from Nebraska and Iowa. Additionally, the clinical data utilized do not encompass all hospitals and clinics patients may visit. Integrating health information exchange data would enhance the clinical picture and facilitate studies investigating readmission.

This study is also limited by the quality of data available in the EHR. For instance, 28 893 of 508 315 (5.68%) of all addresses were unable to be geocoded as they were recorded as some variant of unknown, were null, or were post office boxes rather than a physical address of a residence. In addition, there is varying quality of confidence in the results returned by the geocoding software, with 81 179 of 508 315 (15.97%) having a geo-rating with low confidence (PostGIS geo-rating >10). As addresses are nonuniform and may contain errors, some may not geocode accurately. By excluding patients and potentially mismapping a small portion of the patient population, the potential for bias is introduced. While race and gender showed no significant difference between the included and excluded populations, as is evidenced in other studies, rural locations had a lower percentage of successful geocoding.[36] Differences in these populations were also seen for age (included population slightly younger) and financial class (Medicare patients more likely to be excluded). Recognizing these population differences is essential, as they may impact analyses when future studies rely on these data.[37] While the geocoding for this study did not reach 100% completeness or 100% accuracy, results were comparable with other first-pass geocoding efforts.[36,38–43]

Future research

To demonstrate reproducibility and extend the results demonstrated in this paper, collaborations to implement this data integration approach will occur in an existing distributed research network (Greater Plains Collaborative [GPC]). Within the GPC, sites have implemented de-identified data warehouses on an i2b2 platform. Elements of the methodology described in this paper may be implemented at novel sites with varying levels of resource commitment. At sites with i2b2 deployed, the metadata we displayed to make data queryable in the i2b2 hierarchy may be readily incorporated with minimal investment. Creating the data facts at a novel site may be accomplished by following the pattern we outlined; however, personnel would need to be in place who are competent in geocoding. Of note, any geocoding process the institution has confidence in may be employed and readily incorporated into the pipeline we describe. Thus, creating i2b2 metadata and obtaining ACS data may be achieved by directly following the pattern we outline and geocoding may be accomplished in an institution-specific manner and incorporated into the pattern illustrated. While resource investment is requisite, the pattern we describe illustrates a functional workflow and sufficient detail for metadata creation.

Future efforts will also address incorporating additional deprivation indices and ACS variables into the de-identified clinical data warehouse.[34,44,45] As demonstrated in this study, a standard approach to identifying these indices and component variables within an i2b2 ontology will be maintained for interoperability. Integration of additional indices will allow both the comparison of the efficacy of indices in a variety of contexts as well as the application of validated indices within many disease phenotypes.

Future research may also address some limitations imposed by the geocoding process. A refinement of this process may reduce this bias and increase confidence of results relying on the generated data. While perfecting geocoding is beyond the scope of the current demonstration, an enhanced geocoding process could readily be integrated into this model for incorporating census data into a de-identified data warehouse. Future research may focus both on alternative geocoding software which may be used locally as well as appropriate methods to improve the quality and accuracy of address information extracted from the EHR.

Finally, a limitation of this work is using a single, static address for each patient. We used only a patient's most current recorded address we were able to confidently geocode. We recognize that patients move over time, and this may impact their neighborhood SES. While research has shown the majority of moves from mobile patients have little impact on SES, future work will focus on associating a valid time period for each address recorded.[46] In this way, researchers could use a patient's current SES or the SES for the time period of a historic event.

CONCLUSIONS

With evidence of the impact of environmental factors on health, facilitating comparative effectiveness clinical research incorporating social determinants of health is paramount.[47] Advances in geoinformatics make it possible to link patient location to data provided by the U.S. Census Bureau. We demonstrated an approach to link social determinants of health data from the ACS to clinical data in a de-identified data warehouse. All elements of this approach may be completed at individual sites, avoiding the need to send PHI to a third party. When sites utilize the same geocoding and linkage process, collaborations are possible without introducing unnecessary variability between sites. Institutions who implement this approach using i2b2 may share federated queries across networks such as PCORnet and the GPC, to increase the patient sample size for analysis.[48] Facilitating this research will inform efforts to incorporate location-based census data directly into the EHR and future clinical decision support at the point of care.

This study is an example using only a single contextual socioeconomic index. While many similar indices have been published to estimate SES, we selected Bird et al's[5] model because it is well-validated and fully reproducible using data elements from

the ACS.[34,45,49–51] Future work includes incorporating other well-validated models based on extant ACS data using the process described in this manuscript. These may readily be incorporated into the database within the ontology cell of i2b2. The process for creating each metadata row and developing appropriate XML for this metadata is described in this manuscript and readily applicable to novel indices and novel ACS variables. We demonstrated the efficacy of querying these data with a use case based on evidence from prior studies. We noted both a higher proportion and more frequent per patient utilization of the ED for patients living within areas of lower SES relative to patients from areas with a greater SES.

FUNDING

J.C.M. is partially supported by a Patient-Centered Outcomes Research Institute Program Award grant no. CDRN-1306-04631 and in part by the National Institute of General Medical Sciences grant no. 1U54GM115458-01.

CONTRIBUTORS

BG and JP are responsible for the initial conception and design of the original work and acquisition of data. BG, JP, and MC are responsible for the data analysis. All 4 authors participated in the interpretation of the data analysis. All authors participated in drafting and revising the work and reviewed the manuscript before submission for publication. All 4 authors are accountable for all aspects of the work and its accuracy. The authors take full responsibility for the work presented.

Conflict of interest statement. The authors have no competing interests to declare.

REFERENCES

1. Hsiao CJ, Hing E. Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001-2013. *NCHS Data Brief* 2014; 143: 1–8.
2. Hufnagel SP. National electronic health record interoperability chronology. *Mil Med* 2009; 174 (Suppl 5): 35–42.
3. Charles D, Gabriel M, Furukawa M. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2013. https://www.healthit.gov/sites/default/files/briefs/oncdatabrief16.pdf. Accessed December 8, 2015.
4. Adler-Milstein J, DesRoches CM, Kralovec P, *et al*. Electronic health record adoption in US hospitals: progress continues, but challenges persist. *Health Aff* 2015; 34 (12): 2174–80.
5. Bird CE, Seeman T, Escarce JJ, *et al*. Neighbourhood socioeconomic status and biological 'wear and tear' in a nationally representative sample of US adults. *J Epidemiol Community Health* 2010; 64 (10): 860–5.
6. Estiri H, Patel CJ, Murphy SN. Informatics can help providers incorporate context into care. *JAMIA Open* 2018; 1 (1): 3–6.
7. Casey JA, Pollak J, Glymour MM, Mayeda ER, Hirsch AG, Schwartz BS. Measures of SES for electronic health record-based research. *Am J Prev Med* 2018; 54 (3): 430–9.
8. Nkoy FL, Stone BL, Knighton AJ, *et al*. Neighborhood deprivation and childhood asthma outcomes, accounting for insurance coverage. *Hosp Pediatr* 2018 Jan 9 [E-pub ahead of print]. 8 (2): 59–67.
9. LaVeist T, Pollack K, Thorpe RJr, Fesahazion R, Gaskin D. Place, not race: disparities dissipate in southwest Baltimore when blacks and whites live under similar conditions. *Health Aff* 2011; 30 (10): 1880–7.
10. Gaskin DJ, Thorpe RJ Jr,McGinty EE, *et al*. Disparities in diabetes: the nexus of race, poverty, and place. *Am J Public Health* 2014; 104 (11): 2147–55.
11. Diez-Roux AV, Nieto FJ, Muntaner C, *et al*. Neighborhood environments and coronary heart disease: a multilevel analysis. *Am J Epidemiol* 1997; 146 (1): 48–63.
12. LeClere FB, Rogers RG, Peters K. Neighborhood social context and racial differences in women's heart disease mortality. *J Health Soc Behav* 1998; 39 (2): 91–107.
13. Kramer MR, Hogue CR. Is segregation bad for your health? *Epidemiol Rev* 2009; 31: 178–94.
14. Kandula NR, Wen M, Jacobs EA, Lauderdale DS. Association between neighborhood context and smoking prevalence among Asian Americans. *Am J Public Health* 2009; 99: 885–92.
15. Kimbro RT. Acculturation in context: gender, age at migration, neighborhood ethnicity, and health behaviors. *Soc Sci Q* 2009; 90 (5): 1145–66.
16. White K, Borrell LN. Racial/ethnic neighborhood concentration and self-reported health in New York City. *Ethn Dis* 2006; 16: 900–8.
17. Dominici F, Peng RD, Bell ML, *et al*. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* 2006; 295 (10): 1127–34.
18. Bazemore AW, Cottrell EK, Gold R, *et al*. "Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. *J Am Med Inform Assoc* 2016; 23 (2): 407–12.
19. 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* 2017 Nov 8 [E-pub ahead of print]. 25 (3): 309–314.
20. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The Greater Plains Collaborative: a PCORnet clinical research data network. *J Am Med Inform Assoc* 2014; 21 (4): 637–41.
21. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc* 2014; 21 (4): 576–7.
22. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014; 21 (4): 578–82.
23. Jacquez GM. A research agenda: does geocoding positional error matter in health GIS studies? *Spat Spatiotemporal Epidemiol* 2012; 3 (1): 7–16.
24. Zandbergen PA. A comparison of address point, parcel and street geocoding techniques. *Comput Environ Urban Syst* 2008; 32 (3): 214–32.
25. Lemke D, Mattauch V, Heidinger O, Hense HW. Who hits the mark? A comparative study of the free geocoding services of google and openstreetmap. *Gesundheitswesen* 2015; 77: e160–5.
26. Jacquemin B, Lepeule J, Boudier A, *et al*. Impact of geocoding methods on associations between long-term exposure to urban air pollution and lung function. *Environ Health Perspect* 2013; 121 (9): 1054–60.
27. Tang N, Stein J, Hsia RY, Maselli JH, Gonzales R. Trends and characteristics of US emergency department visits, 1997-2007. *JAMA* 2010; 304 (6): 664–70.
28. Weber GM, Murphy SN, McMurry AJ, *et al*. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc* 2009; 16 (5): 624–30.
29. Office of the National Coordinator for Health IT. 2016 Interoperability Standards Advisory. https://www.healthit.gov/sites/default/files/2016-interoperability-standards-advisory-final-508.pdf. Accessed September 21, 2016.
30. Murphy SN, Weber G, Mendis M, *et al*. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; 17 (2): 124–30.
31. Holl S, Plum H. PostGIS. *GeoInformatics* 2009; 3: 34–6.
32. Hart G. Temporary zip RUCA 3.10 file access page. https://ruralhealth.und.edu/ruca. Accessed September 26, 2018.
33. Cromartie J. Documentation: 2010 Rural-Urban Commuting Area (RUCA) codes. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/documentation/ Accessed September 26, 2018.
34. Messer LC, Laraia BA, Kaufman JS, *et al*. The development of a standardized neighborhood deprivation index. *J Urban Health* 2006; 83 (6): 1041–62.

35. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: the Public Health Disparities Geocoding Project (US). *J Epidemiol Community Health* 2003; 57 (3): 186–99.
36. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *Int J Health Geogr* 2003; 2 (1): 10.
37. Zimmerman DL, Rushton G, Armstrong MP, *et al*. *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice*. Boca Raton, FL: CRC Press; 2007.
38. Gregorio DI, Cromley E, Mrozinski R, Walsh SJ. Subject loss in spatial analysis of breast cancer. *Health Place* 1999; *5* (2): 173–7.
39. Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW. Geographic bias related to geocoding in epidemiologic studies. *Int J Health Geogr* 2005; 4 (1): 29.
40. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *Am J Public Health* 2001; 91 (7): 1114–6.
41. Dearwent SM, Jacobs RR, Halbert JB. Locational uncertainty in georeferencing public health datasets. *J Expo Sci Environ Epidemiol* 2001; 11 (4): 329–34.
42. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology* 2003; 14 (4): 408–12.
43. Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. *Health Place* 2007; 13 (1): 293–8.
44. Dubowitz T, Heron M, Bird CE, *et al*. Neighborhood socioeconomic status and fruit and vegetable intake among whites, blacks, and Mexican Americans in the United States. *Am J Clin Nutr* 2008; 87 (6): 1883–91.
45. Knighton AJ, Savitz L, Belnap T, Stephenson B, VanDerslice J. Introduction of an area deprivation index measuring patient socioeconomic status in an integrated health system: implications for population health. *EGEMS (Wash DC)* 2016; 4: 1238.
46. Knighton AJ. Is a patient's current address of record a reasonable measure of neighborhood deprivation exposure? A case for the use of point in time measures of residence in clinical care. *Health Equity* 2018; 2 (1): 62–9.
47. Braveman P, Egerter S, Williams DR. The social determinants of health: coming of age. *Annu Rev Public Health* 2011; 32: 381–98.
48. McMurry AJ, Murphy SN, MacFadden D, *et al*. SHRINE: enabling nationally scalable multi-site disease studies. *PLoS One* 2013; 8 (3): e55811.
49. UW Health Innovation Program. Health innovation program. Area deprivation index. http://www.hipxchange.org/AD. Accessed April 21, 2018.
50. Singh GK. Area deprivation and widening inequalities in US mortality, 1969-1998. *Am J Public Health* 2003; 93 (7): 1137–43.
51. Butler DC, Petterson S, Phillips RL, Bazemore AW. Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. *Health Serv Res* 2013; 48 (2 Pt 1): 539–59.