Journal of the American Medical Informatics Association, 25(4), 2018, 419–422
doi: 10.1093/jamia/ocx117
Advance Access Publication Date: 27 October 2017
Brief Communication

Brief Communication

FACETS: using open data to measure community social determinants of health

Michael N Cantor,[1,2,3] Rajan Chandras,[3] and Claudia Pulgarin[2]

[1]Department of Medicine, [2]Department of Population Health, New York University School of Medicine, [3]Medical Center Information Technology, NYU Langone Health, New York, NY, USA

Corresponding Author: Michael N Cantor, 227 E 30th St, #725, New York, NY 10016, USA. E-mail: michael.cantor@nyumc.org

Received 14 June 2017; Revised 9 August 2017; Editorial Decision 14 September 2017; Accepted 23 September 2017

ABSTRACT

Objective: To develop a dataset based on open data sources reflective of community-level social determinants of health (SDH).

Materials and Methods: We created FACETS (Factors Affecting Communities and Enabling Targeted Services), an architecture that incorporates open data related to SDH into a single dataset mapped at the census-tract level for New York City.

Results: FACETS (https://github.com/mcantor2/FACETS) can be easily used to map individual addresses to their census-tract-level SDH. This dataset facilitates analysis across different determinants that are often not easily accessible.

Discussion: Wider access to open data from government agencies at the local, state, and national level would facilitate the aggregation and analysis of community-level determinants. Timeliness of updates to federal non-census data sources may limit their usefulness.

Conclusion: FACETS is an important first step in standardizing and compiling SDH-related data in an open architecture that can give context to a patient's condition and enable better decision-making when developing a plan of care.

Key words: social determinants of health, open data, vulnerable populations

BACKGROUND AND SIGNIFICANCE

Much of health is determined not by people's interactions with the health care system, but by social factors that include their daily environment, neighborhood, and community and family cultural influences. These personal-, family-, and community-level social determinants of health (SDH) have been shown across clinical conditions, age groups, and populations to have effects on multiple health-related outcomes.[1–5] Understanding the environment in which patients live and other nonmedical factors that may be affecting their health also gives a more realistic model for risk stratification[1] and the true complexity of patient populations, which may lead to more effective patient, family, and community engagement.[6,7] Similarly, SDH are gaining importance as factors that can be incorporated into new schemas for managing patients in the framework of value-based care, including the Centers for Medicare and Medicaid Services' accountable care communities and accountable care organizations.[8] Data related to SDH are also of great interest to large-scale cooperative research projects like the PCORnet[9]

Data related to SDH abound in different open datasets (such as the US Census Bureau), but often require extensive processing or multiple queries to create a usable dataset. Vendor-supported datasets (eg, HealthLandscape)[10] are often of higher quality, but may not be granular enough to meet the needs of a specific project, may be too costly for routine use, or may not provide access to data programmatically at a scale to be effectively incorporated into predictive models or a population-level dataset. With this in mind, we developed the New York City instance of FACETS (Factors Affecting Communities and Enabling Targeted Services), a database of community-level (as opposed to self-reported) SDH data, with the expectation that access to this data would be of interest to both health systems and researchers who are evaluating the importance of SDH in managing patient populations.

We focused on data in New York City for the initial instance because it was most relevant to our local patient population, and also because of the wide range of data related to SDH that is available through city and state databases. Data in NYC-FACETS is represented at the census-tract level. We chose census tracts rather than smaller block groups, because several important data points are only reported at the tract level. Additionally, linking patients to block group data may lead to privacy issues, as it could potentially provide more unique characteristics that could be used for reidentification attempts. The initial use case of NYC-FACETS is as a data source that can be integrated with data from electronic health records (EHRs) for use in research and operational projects. Specifically, one can map a patient's address to its respective census tract and then use FACETS to obtain data on community-level determinants. Because FACETS is a relatively small dataset, it is currently presented as a lookup table in a spreadsheet.

MATERIALS AND METHODS

Data sources

We used only open, freely available data sources to populate the FACETS database. Not surprisingly, many of the data sources were government agencies. Because many of our data points involve demographic characteristics, the source of the vast majority of data is the Census Bureau's annual American Community Survey (ACS), accessible through American FactFinder, among other sites.[11]

For the non-ACS results, many of the data points were available in an analysis-ready format (eg, the US Environmental Protection Agency's Respiratory Hazard Index (RHI)[12] and the US Department of Agriculture's calculation of low access to healthy food[13]). We obtained data on the distance to parks from researchers at the Centers for Disease Control and Prevention (CDC) based on their previous publication,[14] and data on built environment and walkability from the Built Environment and Health Research Group at Columbia University.[15] We used New York State Open Health data (which includes geocoding) to obtain the number of tobacco retailers[16] per census tract, and used NYC Open Data crime reports,[17] also geocoded, to calculate the number of serious crimes per census tract. The housing quality measure is based on the number of housing violations reported per units of housing, and was reported by the NYU Furman Center.[18] We obtained voter turnout and number of registered voters by state assembly district in 2014 from the State Board of Elections,[19] and used data from the Census Bureau to map those districts to census tracts. We also obtained composite indices from the Census Bureau (Gini[20]) and the CDC (Social Vulnerability Index[21]) that give an overall score or ranking for a tract for social determinants.

One issue with older data sources (eg, pre-2015 housing violations, distance to parks) was that data was mapped to 2000, rather than 2010, census tracts. In those cases, we used mapping tables from the Census Bureau to map between iterations,[22] and only used data for 2000 tracts that were fully (rather than partially) mapped to 2010 tracts.

The level of engineering required to integrate the various datasets into a single location varied by source. As noted above, Census Bureau and CDC data, which make up the majority of FACETS, are analysis-ready as provided and require only minor modifications to be compiled into a single database. State and city data were generally geocoded, so they only required minimal work to reverse geocode their latitude and longitude values to census tracts. Voting districts do not map directly to census tracts, so turnout data required additional mapping through Census Bureau tools.

Table 1. Data elements in FACETS

Community-level determinants in FACETS	
Measure	Source
Total population	ACS
Urban/rural classification	USDA
Total population	ACS
Racial diversity	ACS
Ethnic diversity (Hispanic/non-Hispanic)	ACS
US citizenship	ACS
Foreign vs native-born	ACS
Educational attainment	ACS
English proficiency	ACS
Poverty rate	ACS
Median household income	ACS
Unemployment rate	ACS
Health insurance status	ACS
Respiratory Hazard Index	EPA
Access to healthy food	USDA
Distance to parks	CDC
Walkability score	BEH
Tobacco retailers/1000 population	NYS
Felony crime/1000 population	NYC
Gini index of inequality	ACS
Social Vulnerability Index	CDC
Housing violations/1000 units	FC
Voter turnout	BOE

Abbreviations: ACS: American Community Survey; USDA: US Department of Agriculture Food Access Research Atlas; EPA: Environmental Protection Agency National Air Toxics Assessment; CDC: Centers for Disease Control and Prevention; BEH: Columbia Built Environment and Health Research Group; NYS: New York State Open Health Data; NYC: New York City Open Data; FC: Furman Center; BOE: New York State Board of Elections.

RESULTS

The content of the database and the source of the respective data elements in FACETS can be seen in Table 1. The NYC-FACETS dataset is available at https://github.com/mcantor2/FACETS.

Census data, presented through American FactFinder, was the easiest to obtain and map into the FACETS data table. Other data provided by federal agencies (RHI, Social Vulnerability Index [SVI], etc.) required slightly more involved searching and navigation, but were also easily obtainable. Non-census federal data were not as timely, with data provided as a onetime initiative (eg, park distance from the CDC) or updated on a standard timeline (eg, RHI). The usefulness of a dataset like FACETS can be seen in a hypothetical use case: looking for a correlation between changes in body mass index and neighborhood characteristics,[23] specifically walkability, distance to parks, and an overall measure like the SVI. Without FACETS, one would have to obtain the original data from 3 different sources, normalize the SVI data, and convert the park distance data from 2000 to 2010 census tracts. Preliminary work from our group in this area (currently under review for publication) has shown statistically a significant, directionally correct (ie, more reductions in body mass index with greater walkability) correlation among all these factors using patient data from our institution and the data in FACETS.

DISCUSSION

For the FACETS database, though we have incorporated many community-level factors, the information is not comprehensive. Many additional sources of data related to SDH can be incorporated into the database as necessary. Additionally, the many SDH-related indices, particularly the CDC's Social Vulnerability Index, may supersede the other individual measures and be sufficient to obtain an accurate estimation of the community-level factors affecting a patient's health. We chose to collect the individual and demographic data points both for completeness and because it is likely that the individual factors may also be important in more detailed analyses.

Much of the data within FACETS is based on estimates from the ACS, which are provided along with their respective margins of error. We chose to use only the estimates in the database, for usability and because the goal of the project is to characterize the patient's environment, rather than analyze the specific results. Another potential data-quality issue is the crosswalk between 2000 and 2010 census tracts; as noted above, we chose to ignore potentially useful data when mappings between iterations were only partial because of the difficulty of assigning populations and SDH-related factors to the 2010 tracts. The ACS is updated yearly, and we will update FACETS on the same schedule. However, other data sources (like the RHI) are updated less frequently, so, as with any retrospective observational dataset, current conditions may be different from when the data was collected. Expanding FACETS to the national level will be facilitated by similar open data efforts in other areas outside of New York City and New York State, but will require additional investment at all levels.

One potential challenge to the usefulness of FACETS and other similar databases is the quality of addresses obtained from the EHR. Manually entered addresses often have spelling or formatting errors, which makes mapping them to census tracts difficult. Additionally, street names may not be unique (eg, West 110th Street is also Cathedral Parkway) or may be present in several tracts. Fuzzy matching, relatively straightforward natural language processing techniques, and a database of alternative street names are all techniques one may use to improve the accuracy of the address mapping.

Perhaps more important than the technical issues described above, however, is the role of community-level SDH in clinical care. While FACETS focuses on these community-level determinants, many other groups have shown the importance of individual-level (patient-reported) determinants in clinical care.[24–26] The Institute of Medicine report emphasizes individual-level over community-level determinants, at least partly because one is addressing a patient's specific needs in the moment (eg, a social worker's referral for housing instability), and it is therefore easier to take action and, ideally, see impact. We believe that both community- and individual-level determinants can have a significant impact on care, with community-level determinants functioning more on the "back end" of a care system. For example, community-level determinants can be integrated into predictive models and other decision support to help tailor interventions. A recommendation of diet and exercise for a prediabetic asthmatic patient may need to be tailored differently if the patient lives in an area that has high levels of air pollution and lacks fresh food and access to parks. Ideally, community-level determinants will be another data point that, along with the patient's clinical profile and individual needs, will form the basis of an effective plan of care. The impact of this data can also extend beyond the health care system and influence areas such as city planning, which can also have a positive influence on patient outcomes.

CONCLUSIONS

Data related to SDH are easily obtainable from public sources, and it requires varying degrees of effort to compile the disparate sources into a single database. Mapping and compiling the data points into a unified database allows for both aggregate analysis and comparisons of the impact of individual factors on specific clinical outcomes. Though the data require processing and occasional cross-mapping, the process is tractable. The process of accessing SDH data in the clinic would benefit from a standardized representation of SDH in the EHR, which is part of our group's future work. Additionally, we plan to populate FACETS with additional data related to SDH from census tracts outside of NYC, and to work with local and national resources to obtain data (eg, crime rates, voter turnout, walk scores) at similar levels of granularity.

The key questions regarding SDH data, as noted above, remain around their actual application in clinical care, since knowing a patient's situation without having tools to address it may lead to frustration rather than better outcomes. Initiatives to create these tools are growing and being supported by diverse entities, from health care networks themselves to startup vendors. We believe that FACETS is an important first step in providing the high-quality data that will enable these tools and the interventions they permit to work most effectively.

FUNDING

MNC is supported in part by NYU Clinical and Translational Science Award grant UL1 TR001445 from the National Center for Advancing Translational Sciences, National Institutes of Health. MNC and CP are supported in part by New York City Clinical Data Research Network grant CDRN-1306-03961 from the Patient-Centered Outcomes Research Institute.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

MNC, RC, and CP conceived and developed the database structure. MNC and CP obtained the data. MNC wrote the final manuscript. All authors approved the final version.

ACKNOWLEDGMENTS

We would like to thank James Holt, Andrew Rundle, and David Lee for their contributions to the FACETS dataset.

REFERENCES

1. Bieler G, Paroz S, Faouzi M, *et al*. Social and medical vulnerability factors of emergency department frequent users in a universal health insurance system. *Acad Emerg Med*. 2012;19(1):63–68.
2. Chetty R, Stepner M, Abraham S, *et al*. The association between income and life expectancy in the United States, 2001–2014. *JAMA*. 2016;315(16):1750–66.

3. Kind AJ, Jencks S, Brock J, *et al*. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. *Ann Intern Med*. 2014;161(11):765–74.
4. Sills MR, Hall M, Colvin JD, *et al*. Association of social determinants with Children's Hospitals' preventable readmissions performance. *JAMA Pediatr*. 2016;170(4):350–58.
5. Walker RJ, Gebregziabher M, Martin-Harris B, Egede LE. Relationship between social determinants of health and processes and outcomes in adults with type 2 diabetes: validation of a conceptual framework. *BMC Endocr Disord*. 2014;14:82.
6. Kressin NR, Chapman SE, Magnani JW. A tale of two patients: patient-centered approaches to adherence as a gateway to reducing disparities. *Circulation*. 2016;133(24):2583–92.
7. Woolf SH, Zimmerman E, Haley A, Krist AH. Authentic engagement of patients and communities can transform research, practice, and policy. *Health Aff (Millwood)*. 2016;35(4):590–94.
8. Adler NE, Stead WW. Patients in context: EHR capture of social and behavioral determinants of health. *N Engl J Med*. 2015;372(8):698–701.
9. Patient-Centered Outcomes Research Institute. The impact of patient complexity on healthcare utilization. www.pcori.org/research-results/2016/impact-patient-complexity-healthcare-utilization. Accessed September 9, 2017.
10. 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.
11. US Census Bureau. American FactFinder. http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml. Accessed September 9, 2017.
12. US Environmental Protection Agency. *National Air Toxics Assessment: 2011 NATA: Assessment Results*. www.epa.gov/national-air-toxics-assessment/2011-nata-assessment-results.
13. US Department of Agriculture. *Food Access Research Atlas*. www.ers.usda.gov/data-products/food-access-research-atlas.aspx. Accessed September 9, 2017.
14. Wen M, Zhang X, Harris CD, Holt JB, Croft JB. Spatial disparities in the distribution of parks and green spaces in the USA. *Ann Behav Med*. 2013;45 (Suppl 1):S18–27.
15. Columbia University Built Environment and Health Research Group. https://beh.columbia.edu/. Accessed September 9, 2017.
16. New York State Department of Health. *Active retail tobacco vendors*. https://health.data.ny.gov/Health/Active-Retail-Tobacco-Vendors/9ma3-vsuk. Accessed September 9, 2017.
17. NYC Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-7-Major-Felony-Incidents/hyij-8hr7. Accessed September 23, 2016.
18. NYU Furman Center. *CoreData.nyc*. http://coredata.nyc/. Accessed September 9, 2017.
19. New York State Board of Elections. 2014 *Election Results*. www.elections.ny.gov/2014ElectionResults.html. Accessed September 9, 2017.
20. US Census Bureau. Income Inequality. www.census.gov/topics/income-poverty/income-inequality.html. Accessed September 9, 2017.
21. Agency for Toxic Substances and Disease Registry. *The Social Vulnerability Index*. http://svi.cdc.gov/. Accessed September 9, 2017.
22. US Census Bureau. Geography. www.census.gov/geo/maps-data/data/relationship.html. Accessed September 9, 2017.
23. Loo CK, Greiver M, Aliarzadeh B, Lewis D. Association between neighbourhood walkability and metabolic risk factors influenced by physical activity: a cross-sectional study of adults in Toronto, Canada. *BMJ Open*. 2017;7(4):e013889.
24. Gottlieb LM, Hessler D, Long D, *et al*. Effects of social needs screening and in-person service navigation on child health: a randomized clinical trial. *JAMA Pediatr*. 2016;170(11):e162521.
25. Patel MR, Piette JD, Resnicow K, Kowalski-Dobson T, Heisler M. Social determinants of health, cost-related nonadherence, and cost-reducing behaviors among adults with diabetes: findings from the national health interview survey. *Medical Care*. 2016;54(8):796–803.
26. Venn D, Strazdins L. Your money or your time? How both types of scarcity matter to physical activity and healthy eating. *Soc Sci Med*. 2017;172:98–106.