Ontology-based Integration of Consumer Data and EHR Systems to Fill Gaps in Social Determinants of Health Data

S. Clint Dowland [1], Melody L. Greer [1], Sudeepa Bhattacharyya [1,2], and Mathias Brochhausen [1]

[1] *University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA*
[2] *Arkansas State University, Jonesboro, Arkansas, USA*

Abstract

Social risk factors impact health outcomes. Understanding these risk factors requires increased collection and better maintenance of data on social determinants of health. Despite recent efforts to improve collection and organization of such data, there still are considerable hurdles to collecting these data in the clinical setting. To fill this gap, we propose extracting social determinants of health-relevant data from commercial consumer data as a source of additional individual-level, social risk factor-related data. We present early results of our efforts toward developing a social risk factor ontology and using an ontology-based approach to integrating commercial consumer data items with electronic health record data.

Keywords

social risk factor, social determinants of health, consumer data, ontologies

1. Introduction

It is widely recognized that social conditions influence health outcomes in many ways. These factors are called *social determinants of health* (SDOH). Due to the growing awareness of the importance of SDOH, efforts have been made to gather and organize data about them [1-6]. For example, there are SDOH-focused screening instruments such as the Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE); Kaiser Permanente's Structural Vulnerability Assessment Tool; and Epic's Healthy Planet module [7-9]. Additionally, there are codes for recording SDOH in coding systems such as SNOMED-CT and ICD-10 CM, and the Fast Healthcare Interoperability Resources (FHIR) data exchange standard includes SDOH condition categories [10].

However, studies have found that SDOH-relevant information is documented in clinical notes more often than through medical codes, that clinicians' lack training in gathering information about social risk factors as well as difficulties discussing sensitive subjects are obstacles to gathering SDOH data, and that some clinicians have concerns regarding the increase in administrative burden that comes with gathering social risk factor information [11, 12]. While there are area-level data that are relevant to SDOH, there are limits on the extent to which we can infer facts about an individual from these [13]. Greer, Zayas, and Bhattacharyya (2022) propose the use of commercial consumer data as an additional source of SDOH data as a solution to these problems [14].

In this paper we report early results on integrating commercial consumer data with electronic health record (EHR) data to address gaps in the availability of SDOH data for clinical and clinical research purposes. Our immediate goal is to create a pipeline from SDOH-relevant consumer data elements to EHR systems, thereby allowing health care providers to consider a more robust picture of a patient's social situation in a way that does not require gathering the data via additions to the workflow such as

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: clintdowland@gmail.com (S.C. Dowland); mlgreer@uams.edu (M.L. Greer); sbhattacharyya@astate.edu (S. Bhattacharyya); mbrochhausen@uams.edu (M. Brochhausen)
ORCID: 0000-0003-1909-9269 (S.C. Dowland); 0000-0002-1070-5907 (M.L. Greer); 0000-0003-3493-6029 (S. Bhattacharyya); 0000-0003-1834-3856 (M. Brochhausen)

using questionnaires. More broadly, we aim to identify SDOH-relevant consumer data and to enable the integration of SDOH data from disparate sources and of different types.

2. Background

To enable the pipeline from consumer data to EHR systems we aim to use an ontology-driven approach. By transforming relevant data items about a person into an ontology-enhanced form to represent phenomena the data items are about, as well as modeling FHIR-compatible medical codes in our ontology, we can make inferences about which of these codes characterize the person.

As the ontological basis for our project we are developing the Social Risk Factor Ontology (SRFON). SRFON is intended not only to represent types of entities and relations that are relevant to SDOH, but also to represent SDOH-related medical codes and link them to representations of the types of situations that they characterize. In these ways SRFON can enable the integration of SDOH-related data from disparate sources and inferences from data about an individual to medical codes that characterize that individual.

The above application of consumer data and SRFON requires ontologically modeling SDOH-relevant consumer data elements and medical codes, but does not require modelling EHR data because it is only intended to input information into EHR systems in the form of medical codes. This is advantageous for the purpose of providing information to health care providers in familiar formats, and saves time since fewer data elements need to be modeled and transformed. But we are also interested in the prospect of integrating SDOH-related data from consumer data sources with EHR data by ontologically modeling and transforming data of both types.

3. Materials and methods

Our methods are applied to two major components: a) developing an ontology covering SDOH, and b) identifying and selecting SDOH-relevant consumer data elements.

3.1. Developing SRFON

SDOH are not limited to direct influences on health, but instead include phenomena that influence health indirectly. In some cases the same phenomena have the potential to affect health through more than one pathway of influence, so that instead of only forming discrete causal pathways, SDOH can form interconnected webs of causes and effects. This web can include self-perpetuating cycles, such as when a person’s inadequate income is a barrier to transportation accessibility while the person’s lack of access to transportation is a barrier to employment opportunities that could lead to higher income. In order to develop an ontology to represent the relevant entities in the interconnected web of SDOH-related phenomena, we sought out academic literature that reports findings on how two or more SDOH are either correlated or causally related with one another or with effects on health. The starting point was a set of nineteen literature summaries from health.gov, each of which addresses a different SDOH area [15].

From these we extracted each assertion about pairs of social conditions and health effects that stand in a causal relation or are in some way correlated with each other. Many of the members of these pairs appeared in multiple assertions but denoted with different phrases, and so we identified such cases and made the phrasing uniform. In this initial list of terms for a number of interconnected phenomena, many terms did not name a single type of entity, but instead denoted a state of affairs involving multiple entities of certain types that are related in certain ways, and so each was analyzed in order to populate a list of terms for the relevant sorts of entities and relations. For example, household overcrowding is a matter of how several entities relate to one another, such as the members of some household and the rooms within their shared home.

Next, these terms were searched in three ontology repositories—Ontobee, BioPortal, and the Ontology Lookup Service—in order to find, when possible, preexisting ontology terms that represent the same types of entities [16-18]. Preference was given to terms from ontologies for which the Basic

Formal Ontology (BFO) is the upper ontology, as it is for SRFON [19]. Selected terms were imported, and new SRFON terms were created for types of entities or relations without matches. Terms in SRFON were arranged into a BFO-based hierarchy. Several SDOH-related clinical codes were included as well. These are represented as individual instances of *clinical code* and as members of their code sets.

3.2. Commercial consumer data and SDOH

Commercial consumer data include a wide range of types of information about an individual, the individual's household, and the area in which the individual resides. Commercial consumer data is gathered for the purpose of predicting a person's spending habits and it includes a vast amount of information about various aspects of the person's life.

In our project we use consumer data from a commercial database marketing company. Their database contains 6,260 distinct data elements that might each be populated with values for a given individual. We were provided with data dictionaries that include the value sets and written descriptions of each data element. We are manually reviewing these in order to find SDOH-related data elements. Additionally, to aid with finding and organizing relevant data elements, we have used keyword searches for SDOH-relevant terms, including but not limited to SRFON term labels and synonyms for them.

4. Results

4.1. SRFON

From the aforementioned literature summaries, we extracted 809 assertions about causal relations or correlations between pairs of phenomena including various social risk factors and health outcomes. Following the process of making the phrasing uniform and of replacing several terms that describe complex situations with corresponding collections of terms for the salient types of entities in those situations, there were 718 distinct class terms. After importing suitable terms from other ontologies—and in many cases importing superclasses of them as well—SRFON currently contains 677 new class terms and imports 255.

4.2. SDOH-relevant consumer data

While our review process is ongoing, we have thus far identified over 80 consumer data elements that are relevant to SDOH, either on their own or in combination. The consumer data include information about the employment status and education level of the person, each of which are important in relation to SDOH. In addition to the education level of the person the data is primarily about, there are also data elements about the education levels of up to four other individuals in the person's household. Additionally there are data elements about the occupation of the person and up to four other members of their household. Other information about the person's household can be derived from data elements about the total number of people in the household, the number of adults and the number of children in the household, whether there is a smoker in the household, and whether there is a single parent in the household. Relevant data about the person's home include the type of dwelling, whether the home has a source of heating or cooling, how many bedrooms and how many total rooms are in the home, and whether the home is owned or instead rented by the person. There are also data elements about the person's primary language and English proficiency, about the person's ethnicity at two levels of granularity, and how many vehicles are owned by members of the person's household. The consumer data also include area-level elements that are relevant to SDOH and specific to the area in which the person resides. These include for example seven cost of living indices at the county level, six of which concern the cost of specific types of products or services such as groceries, housing, and transportation.

The data elements described above are not an exhaustive list of SDOH-relevant consumer data elements, but suffice to reflect that information related to social risk factors can be derived at the individual level from commercial consumer data. Next, we take a closer look at some of these examples and how we ontologically represent what they are about.

4.3. Overcrowding

Household overcrowding occurs when too many people live together in the same residence, and it is a social risk factor [20-21]. The consumer data set we are utilizing does not contain any data elements that are explicitly about overcrowding nor any single value that indicates overcrowding on its own. However it includes data items about the person's household and residence that are relevant to measuring overcrowding.

Ways of measuring overcrowding tend to take as inputs both some measure of the household size and some measure of the home's capacity to house them [22]. For example, one standard that has been used is whether there are more than two persons per bedroom in the residence. In Figure 1, we represent a scenario in which some person P1's household consists of six members living in a residence with two bedrooms. In this figure, white nodes represent values of data items; blue nodes represent individual entities, including data items whose values are derived from consumer data; and green nodes represent individual entities whose values or relations to the other entities are inferred.

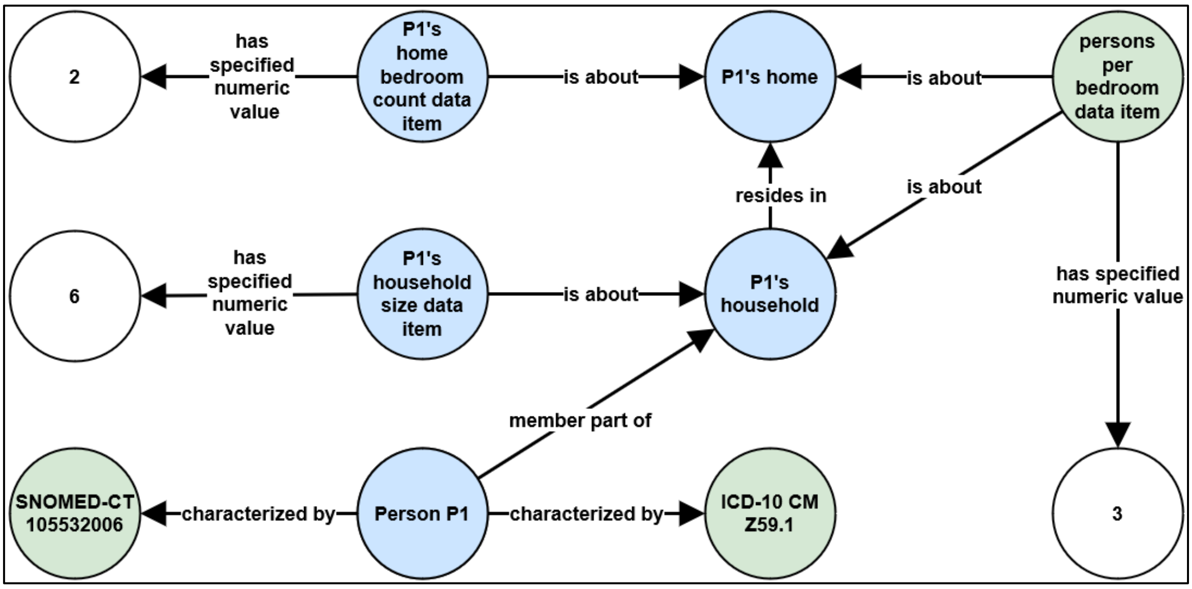
2
has specified numeric value
P1's home bedroom count data item
is about
P1's home
is about
persons per bedroom data item
resides in
is about
6
has specified numeric value
P1's household size data item
is about
P1's household
has specified numeric value
member part of
SNOMED-CT 105532006
characterized by
Person P1
characterized by
ICD-10 CM Z59.1
3

Figure 1: Person P1's household overcrowding

If we implement the aforementioned standard for assessing overcrowding, then from the combination of that standard and the inferred three persons per bedroom we can further infer that P1's household is experiencing overcrowding and thus that P1 is characterized by appropriate medical codes. For example, within the value set for the FHIR SDOH condition category 'inadequate-housing' are codes from both SNOMED-CT and ICD-10 CM. The SNOMED-CT codes include one that is specifically about overcrowding: 105532006, "Overcrowded in house." ICD-10 CM includes another that is applicable: Z59.1, "Inadequate housing." In Figure 1 we represent P1 as standing in the *characterized by* relation to each of these codes.

In addition to bedroom count and number of household members, the commercial consumer data also include the number of adults in the household and the number of children in the household, as well as the residence's number of rooms in general and its square footage. These are relevant for measuring overcrowding because, in addition to overcrowding standards that use the number of persons per bedroom, there are others that make use of the number of persons per room or the number of square feet per person, and some require a distinction between adult and child members of the household [22-23]. The consumer data is thus a potential source of information relevant to a number of ways that overcrowding has been measured. One advantage of this is that when the data required for one measure is not available for an individual, it might be possible to use a different measure. Another is the ability to compare and contrast different measures of overcrowding, for example by examining how often they evaluate the same households as overcrowded, or analyzing cases in which they evaluate the same households differently in order to investigate how other variables correlate with overcrowding as measured in different ways.

4.4. Language use and health care

Language barriers and limited English proficiency (LEP) can be detrimental to health in a number of ways. For example they can be obstacles to understanding health-related information from public sources [24]. Furthermore, language barriers between patients and providers are associated with lower quality of health care [25-26]. They can do so by inhibiting the patient's abilities to understand the provider's questions and to clearly convey problems and concerns to the provider, thus inhibiting the provider's ability to reach an accurate diagnosis. Additionally, language barriers can make it difficult for the patient to properly understand the provider's instructions once a diagnosis is made.

A terminological clarification will allow us to clearly distinguish two important concepts related to language use. By "primary language" we mean the language that a person is most adept at, or is most comfortable with, using. This is often the person's first language. In contrast, we use "preferred language" for the language a person selects to use in a given situation, for example during a health care encounter. These are often but not always the same, for preferred language can vary from situation to situation even while primary language stays the same. For example, a bilingual Spanish and English speaker might select English as their preferred language at a hospital in the U.S., while preferring Spanish when visiting a hospital in Mexico, so as to increase their chance of successful communication in each setting.

A preferred language field is found in EHR systems that meet Federal guidelines for stage 1 Meaningful Use Requirements [27]. But to know whether the preferred language is the patient's primary language and whether the choice of language might be cause for concern about a language barrier, we need more information. The commercial consumer data we are using can help with this because they include data about the person's primary language and about the person's ability to speak English. In Figure 2 we depict a scenario in which consumer data reflect that some person P2's primary language is Spanish and that P2 has LEP, while EHR data indicates P2 selected English as the preferred language for some particular health care encounter. The representation of this scenario in Figure 2 is based in part upon the way that languages, linguistic competences, and primary and preferred language data are represented in the Ontology of Medically Related Social Entities (OMRSE) [28].

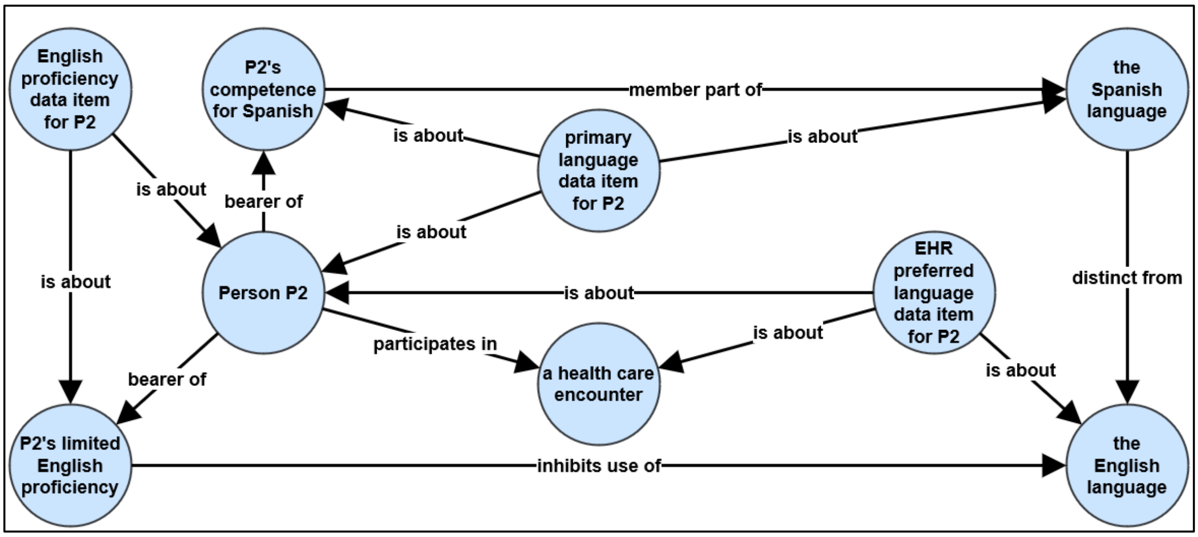

Figure 2: Data about P2's language use and capabilities

We can see that during the health care encounter, P2 was at risk for facing language-related barriers to the benefits of health care. Of course, not everyone whose primary language differs from their preferred language in a given encounter will face a language barrier during that encounter, since a person can be highly proficient in multiple languages. But having LEP and a primary language other than English indicates P2 faces a potentially detrimental language barrier when communicating with health care providers in English.

5. Conclusion

We have described a number of consumer data elements that are relevant to SDOH, and in more detail have discussed two sets of examples, how we represent what the data are about, and examples of what can be inferred from them. SRFON can aid in integrating such data with EHR or other SDOH-related data, as well as with enabling inferences from consumer data items to medical codes that characterize the person.

We will continue developing SRFON as well as identifying and ontologically modeling SDOH-related consumer data elements. Other future work includes using SRFON as the base for an additional ontological representation of SDOH-related correlations and causal relations that are reported as findings in academic literature—starting with those that informed the initial development of SRFON—thereby integrating findings from a number of sources. One possible use of this is to aid in identifying potential problem areas for individuals. For example, several variables in a person's life might each be of types that can cause or otherwise increase the risk of the same type of problem. Future work also includes looking into the potential utility of integrating individual-level consumer data with relevant area-level data from additional sources, such as from the US Census Bureau.

6. Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Award No. OIA-1946391. Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. References

[1] Blackman, P. H. (1994). Actual causes of death in the United States. *Jama*, *271*(9), 659-660.

[2] McGinnis, J. M., Williams-Russo, P., & Knickman, J. R. (2002). The case for more active policy attention to health promotion. *Health affairs*, *21*(2), 78-93.

[3] Wilensky, G. R., & Satcher, D. (2009). Don't forget about the social determinants of health. *Health Affairs*, *28*(2), w194-w198.

[4] Braveman, P., & Gottlieb, L. (2014). The social determinants of health: it's time to consider the causes of the causes. *Public health reports*, *129*(1_suppl2), 19-31.

[5] Gold, R., Cottrell, E., Bunce, A., Middendorf, M., Hollombe, C., Cowburn, S., ... & Melgar, G. (2017). Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *The Journal of the American Board of Family Medicine*, *30*(4), 428-447.

[6] LaForge, K., Gold, R., Cottrell, E., Bunce, A. E., Proser, M., Hollombe, C., ... & Clark, K. D. (2018). How 6 organizations developed tools and processes for social determinants of health screening in primary care: an overview. *The Journal of ambulatory care management*, *41*(1), 2.

[7] PRAPARE. Who We Are. https://prapare.org/who-we-are/.

[8] Bourgois, P., Holmes, S. M., Sue, K., & Quesada, J. (2017). Structural vulnerability: operationalizing the concept to address health disparities in clinical care. *Academic medicine: journal of the Association of American Medical Colleges*, *92*(3), 299.

[9] OCHIN. Building the Foundation for Population Health at OCHIN. https://ochin.org/blog/population-health-at-ochin.

[10] HL7 International – Patient Care WG. (2023, June 02). SDOH Clinical Care: 14.6.1 Resource Profile: SDOHCC Condition.

[11] Guo, Y., Chen, Z., Xu, K., George, T. J., Wu, Y., Hogan, W., ... & Bian, J. (2020). International Classification of Diseases, Tenth Revision, Clinical Modification social determinants of health codes are poorly used in electronic health records. *Medicine*, *99*(52).

[12] 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.

[13] Greer, M. L., Garza, M. Y., Sample, S., & Bhattacharyya, S. (2023). Social Determinants of Health Data Quality at Different Levels of Geographic Detail. *medRxiv*, 2023-02.

[14] Greer, M. L., Zayas, C. E., & Bhattacharyya, S. (2022). Repeatable enhancement of healthcare data with social determinants of health. *Frontiers in big Data*, *5*.

[15] US Department of Health and Human Services, & Office of Disease Prevention and Health Promotion. Social Determinants of Health Literature Summaries - Healthy People 2030. https://health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries.

[16] Xiang, Z., Mungall, C., Ruttenberg, A., & He, Y. (2011, July). Ontobee: A linked data server and browser for ontology terms. In *ICBO*.

[17] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., ... & Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, *37*(suppl_2), W170-W173.

[18] Côté, R. G., Jones, P., Apweiler, R., & Hermjakob, H. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC bioinformatics*, *7*(1), 1-7.

[19] Arp, R., Smith, B., & Spear, A. D. (2015). *Building ontologies with Basic Formal Ontology*. MIT Press.

[20] Lepore, S. J., Evans, G. W., & Palsane, M. N. (1991). Social hassles and psychological health in the context of chronic crowding. Journal of Health and Social Behavior, *32*(4), 357–367.

[21] Cardoso, M. R. A., Cousens, S. N., de Góes Siqueira, L. F., Alves, F. M., & D'Angelo, L. A. V. (2004). Crowding: Risk factor or protective factor for lower respiratory disease in young children?. BMC Public Health, *4*(1), 1–8.

[22] Blake, K. S., Kellerson, R. L., & Simic, A. (2007). Measuring overcrowding in housing.

[23] OECD. (2023). Housing overcrowding (indicator). doi: 10.1787/96953cb4-en.

[24] Greer, M. L., Sample, S., Jensen, H. K., McBain, S., Lipschitz, R., & Sexton, K. W. (2021). COVID-19 is connected with lower health literacy in rural areas. *Studies in health technology and informatics*, *281*, 804.

[25] Espinoza, J., & Derrington, S. (2021). How Should Clinicians Respond to Language Barriers That Exacerbate Health Inequity? *AMA Journal of Ethics*, *23*(2), 109-116.

[26] Flores, G. (2005). The impact of medical interpreter services on the quality of health care: A systematic review. *Medical Care Research and Review, 62*(3), 255–299.

[27] Centers for Medicare & Medicaid Services (CMS), "Eligible Hospital and Critical Access Hospital Meaningful Use Core Measures Measure 6 of 11, Stage 1, Record Demographics."

[28] Dowland, S. C., Diller, M. A., Landgrebe, J., Smith, B., & Hogan, W. R. (forthcoming). Ontology of Language, with Applications to Demographic Data. *Applied Ontology*.