Discovering Associations between Social Determinants and Health Outcomes: Merging Knowledge Graphs from Literature and Electronic Health Data

Yoonyoung Park[1], Natasha Mulligan[2], Martin Gleize[2], Morten Kristiansen[3], Joao H Bettencourt-Silva[2]
[1]IBM Research, Cambridge, MA, USA; [2]IBM Research Europe, Dublin, Ireland; [3]IBM Watson Health, Dublin, Ireland

Abstract

Social Determinants of Health (SDoH) are an increasingly important part of the broader research and public health efforts in understanding individuals' physical and mental well-being. Despite this, non-clinical factors affecting health are poorly recorded in electronic health databases and techniques to study how SDoH might relate to population outcomes are lacking. This paper proposes an approach to systematically identify and quantify associations between SDoH and health-related outcomes in a specific cohort of people by (1) leveraging published evidence from literature to build a knowledge graph of health and social factor associations and (2) analysing a large dataset of claims and medical records where those associations may be found. This work demonstrates how the proposed approach could be used to generate hypotheses and inform further research on SDoH in a data-driven manner.

Introduction

With the increasing emphasis on delivering high value and accessible care, there is a growing awareness of the importance of non-clinical factors contributing to people's physical and mental well-being collectively called social determinants of health (SDoH). A large body of prior work has shown significant effects of SDoH on health related outcomes. For example, measures of SDoH were shown to be associated with increased risk of preterm birth,[1] readmission risk,[2] hospitalization rate,[3] or healthcare utilization.[4] The global SARS-COV-2 pandemic has further revealed the stark inequity and inequality in healthcare resource across the US, often associated with different dimensions of SDoH such as race, income, education level, or job security.[5] Timely and effective intervention regarding SDoH seems more important than ever.

From a healthcare provider or policy maker point of view, having insights on social factors that affect his/her population of interest can help design monitoring and public health interventions, prioritize resource allocation, and achieve health equity.[6–8] Limited insights on the prevalent social issues of the population may exist, but it's less likely that a thorough understanding of how much a particular issue affects which subset of the population exists. For example, a hospital knows that a large proportion of the population in its catchment area suffers from poor transportation, but is unclear about how the transportation problem adversely impacts different aspects of population health.

The interest in SDoH has led to a need for analytic and data-driven solutions, with a newly coined term *social informatics*.[9] Probably the biggest challenge pointed out by both researchers and practitioners is the lack of data on SDoH.[10] Existing health databases such as electronic health records or administrative claims data often do not collect information on SDoH in a reliable manner. A new set of codes was introduced during the conversion from ICD-9 to ICD-10 to better capture SDoH in claims data, but the utility of these Z-codes known to be under utilized are not fully understood.[11, 12] Area-based composite indices of SDoH based on US Census data have become more popular as a public resource.[13–15] However, limited applications exist to date among healthcare organizations possibly due to both technical and privacy related concerns. Research on the impact of social factors is active in domains other than healthcare as well and vast amount of literature exist, but incrementally digesting that information and producing actionable insights would not be a feasible or efficient task. Therefore, developing tools to utilize previously untapped sources of data to guide clinical and policy decision making would be highly beneficial from both research and practice point of views.

In this work, we introduce a systematic approach to identify and quantify associations between SDoH and health related outcomes in a specific cohort of people. Our approach has two parallel pipelines, one that mines associations from published evidence in PubMed, and the second which identifies related cohorts in electronic health data. The combined results then inform the subsequent analysis with stronger hypotheses on the association between SDoH of interest and specific health-related outcomes for further investigation. The main contributions of this work are the construction of a PubMed knowledge graph for identifying SDoH associations based on our prior work and the demonstration of methods to augment it with a real-world health data set.

Background and Related Work

Social factors are reported to account for more than half of the deaths in the US in any given year.[16] As the focus on population health grows, interest is also rising in data and analytic solutions to better understand SDoH and how we might integrate them with clinical outcome information. One way to achieve this is to analyze electronic health data based on *a priori* hypotheses, such as administrative claims data generated from insurance processes or electronic medical records (EMRs) generated by providers. However, these data are built for purposes other than research and do not suitably capture information on patients' social factors.[4,17] Despite the efforts to address this limitation by, for example, introducing Z-codes in ICD-10, the utilization level of Z-codes has been very low in Medicare beneficiaries[12]. Unstructured data, such as case notes, may be an additional source of SDoH information when coding is unavailable or sparingly documented.[17] Recent efforts are also underway to create or improve existing standards and represent SDoH information, such as the SIREN and Gravity Projects[18,19], LOINC's models for the representation of screening assessments and measures of SDoH, or HL7 FHIR profiles and extensions, among others.[20]

Knowledge extraction and text mining techniques may be used to discover associations and patterns in large datasets. The application of data and text mining in health informatics is increasing especially with the availability of large electronic health data and improved computing power.[21] Recently, researchers are realizing the potential of text mining through Natural Language Processing (NLP) in medicine to gain additional insights from traditionally underused unstructured data.[22] For example, text mining has been used to extract medical concepts[23] or SDoH related information[17,24] from patients' medical records. Previous work has also focused on extracting homelessness and adverse childhood experiences large corpora of clinical notes[25] while other work has used ontology-driven tools embedded in health records to identify individuals at an increased psychosocial risk.[26]

Our prior work in this domain[27] is probably one of the first attempts to utilize a novel source of textual data, PubMed, for SDoH research. Published peer-reviewed articles available in PubMed provide a great opportunity to gather new insights from a very large corpus of data. In this paper, we build on the prior work and extend the approach to identify and quantify the relations between SDoH and health outcomes in a real-world cohort through electronic health data analysis.

Methods

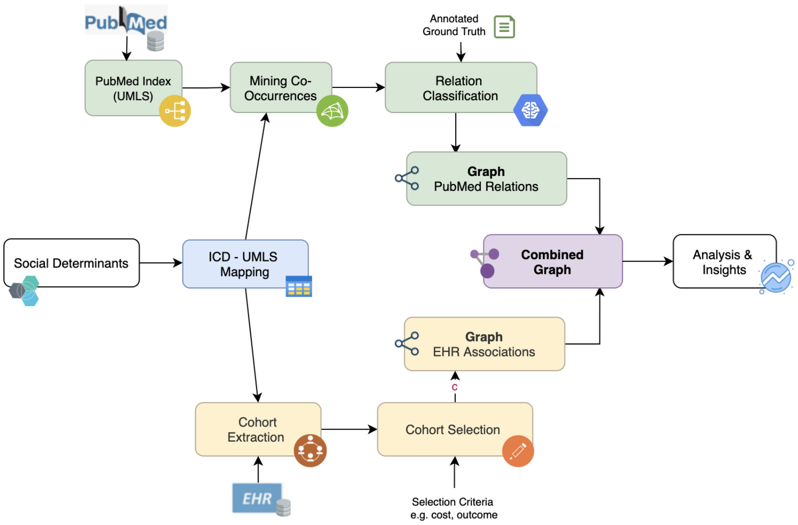


Figure 1: Illustration of the approach used in this paper

Figure 1 illustrates our approach to assessing the associations with SDoH using two sources of data. We begin by imagining the hospital decision maker mentioned earlier, who has some knowledge about SDoH challenges in her population and wishes to know more about their impact on an undetermined set of health related outcomes. Starting from the specific social factor of interest, we identify possible associations by creating a knowledge graph (KG) based on published abstracts in PubMed. By doing this, we are collecting all available insights on the chosen SDoH topics from all research publications. However, information extraction based on co-occurrence gives at best measures of correlation which may or may not be applicable for the specific target population. While they are useful for hypothesis generating purposes, this is a critical limitation for the decision maker because information extraction from such sizable corpus will return numerous pairs of correlated terms with weak evidence. We augment the findings from PubMed by adding electronic health data, which represents data collected from the population of interest. Cohort-specific data analysis may seem self-sufficient to understand the SDoH effect, but in reality, electronic health data poorly captures SDoH and even when these are recorded, some or all of the results may be spurious correlations due to nonrandom data collection and other unobserved confounders. In addition, electronic health data are highly fragmented, often unable to produce long-term effect estimations. We therefore focused on the intersection of the two KGs that generate stronger hypotheses informing the subsequent cohort analysis, which involves quantification of the strength of association for specific SDoH-outcome pairs to generate actionable insights.

SDoH for Information Extraction and Data Analysis

We focused on two specific SDoH factors, housing problems and unemployment. The motivation for choosing housing problems is largely based on the prior evidence from Centers for Medicare Medicaid Services (CMS) showing that the most utilized Z-code among Medicare beneficiaries was Homelessness,[12] in addition to other studies reporting adverse effects of unstable housing issue on health outcomes.[28] Ensuring we have a reasonable prevalence of the ICD social codes in the Claims-EMR Data was important to reliably quantify the association level for these codes. The motivation for choosing unemployment came from our prior work monitoring Google Trends across several SDoH dimensions in the year 2020.[27] The term 'Unemployment' had the most significant increase in terms of the average interest compared with previous years, implying that employment was a severely affected SDoH dimension at the outset of the global pandemic. Although it is not as frequently recorded in data as housing problems, we believe examining unemployment would be a timely and informative analysis in light of the SARS-COV-2 pandemic, as shown in recent studies.[5,29]

Information Extraction from PubMed

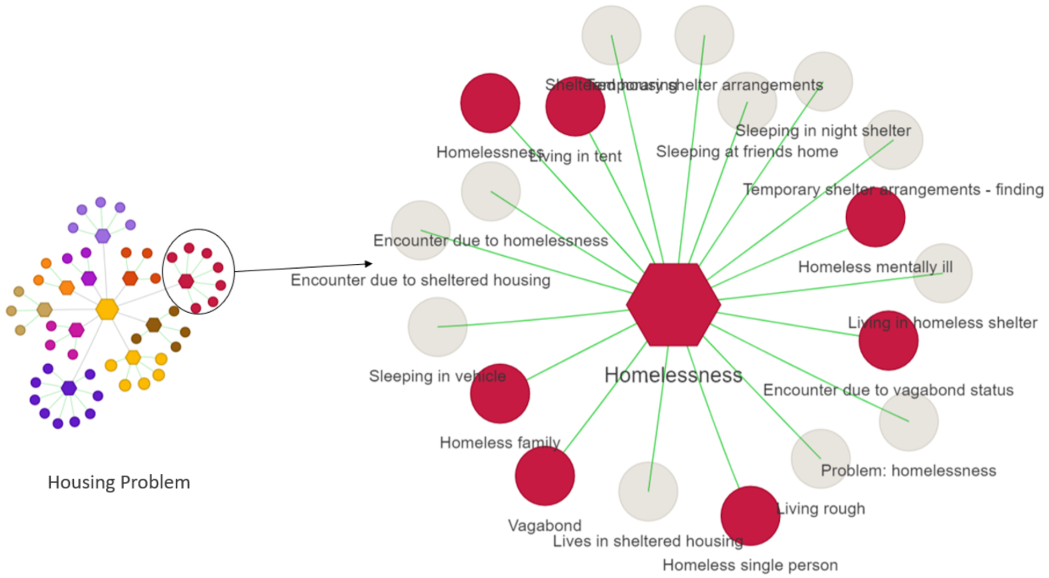

Housing Problem
Sheltered housing
Temporary shelter arrangements
Homelessness
Living in tent
Sleeping in night shelter
Sleeping at friends home
Temporary shelter arrangements - finding
Encounter due to homelessness
Encounter due to sheltered housing
Homeless mentally ill
Living in homeless shelter
Sleeping in vehicle
Homelessness
Encounter due to vagabond status
Homeless family
Problem: homelessness
Living rough
Vagabond
Lives in sheltered housing
Homeless single person

Figure 2: Concept network representing housing problems. Each hexagon is a category and circles are UMLS entities that represent each category. Circles that are grayed out were not found by PubMed MetaMap.

A natural way to mine relationships between socio-medical concepts is to look for their co-occurrence in published literature[30]. Although a high frequency of co-occurrence does not directly indicate association between concepts in the real world, knowing that they are often talked about together is a good indication that the link has been explored by the scientific community at the very least.

We indexed the full 2019 MEDLINE/PubMed Baseline[1], which notably includes the abstracts of research articles. We used MetaMap[31] to tokenize and identify UMLS concepts in the sentences of the abstracts, and indexed each single sentence with Lucene[2] so that it could be retrieved using multiple annotation layers, like words and phrases, UMLS semantic types, or UMLS concepts. It is possible to apply statistical methods on medical text alone but it is generally preferred to overlay a text corpus with concept annotations from an ontology like UMLS to address language variation: there is more than one way to designate the same entity. UMLS is a very extensive ontology[32] and the concepts identified by MetaMap vary widely in nature, so we restricted the medical concepts to only semantic types[3] representing health issues that individual people might have, like: Disease or Syndrome, Sign or Symptom. UMLS semantic types are an upper-level ontology on top of UMLS, providing broader categories that each UMLS concept can fit into[33]. In addition to this subset of UMLS concepts, we also highlighted Social Determinants of Health (SDoH) annotations in our indexed version of PubMed. As a starting point, we used a SDoH lexicon of concepts defined in-house and strongly inspired by the World Health Organization's definition[4]. Different coding terminologies represent SDoH in different ways with varying degrees of detail and specificity. Our team then semi-automatically mapped the SDoH concepts to UMLS concepts: looking manually for the closest denomination and checking that queries using this UMLS concept yielded a non-negligible amount of results in PubMed. The mapping process itself is a challenging task, with some identified SDoH not mapping to any single UMLS concept: for example *"Teenage Pregnancy"* could require a combination of concepts to properly describe both the *teenage* and *pregnancy* aspects of the concept. The SDoH looked at in this contribution, identified in the previous section – *Housing* and *Unemployment* – could be straightforwardly mapped to the corresponding UMLS concept. Figure 2 shows the SDoH Housing and related subterms, mapped to UMLS concepts.

In our experiments, we queried the index for any pair of a SDoH *Housing* or *Unemployment*, and another SDoH or another UMLS concept – as annotated by MetaMap and restricted as described above. Since our index is a sentence-level index, this defined the set of sentences where the SDoH appears together with another socio-medical concept of interest. It is important at this point to note that MetaMap's coverage is not perfect and some concepts in UMLS are never detected by it – even on such a large collection of documents as PubMed. In Figure 2, MetaMap detected no instance of the grayed out concepts in PubMed despite them being defined as children of *Homelessness* in UMLS. Grayed out concepts could be eligible for further revisions of MetaMap's annotation of PubMed as it is important to continuously improve coverage through time.

To build a knowledge graph, we created edges between concepts found together in at least one sentence, and weighted these edges using relative frequency. Formally, the edge (x, y) defined between concept x and concept y is weighted with $W(x, y) = P(y \mid x)$, a conditional probability estimated on our corpus by dividing the size of the set of sentences containing both x and y (called *cofrequency*) by the overall frequency of x in the corpus. An optional step in the graph building process is to prune edges with too low of a weight, to get rid of noisy relations. In this contribution however, this is not necessary as our conclusions will focus on the top ranked relations.

Analysis of Claims-EMR Data

We used IBM® MarketScan® Explorys® Claims-EMR Data (CED), created by linking administrative claims (IBM® MarketScan® Research Databases) and electronic medical records (IBM® Explorys® EMRs).[34] The claims data comes from both privately insured individuals through a variety of fee-for-service, fully and partially capitated plans and individuals with employer-based Medicare supplemental insurance. The EMR data provide additional details. The linked CED is a statistically de-identified, standardized, and normalized data set that contains 5 million patient-level records on demographics, diagnostic and procedure codes, lab tests and vital signs, admission records, payments, and

[1] https://www.nlm.nih.gov/databases/download/PubMed_medline.html

[2] https://lucene.apache.org/

[3] https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

[4] https://www.who.int/health-topics/social-determinants-of-health

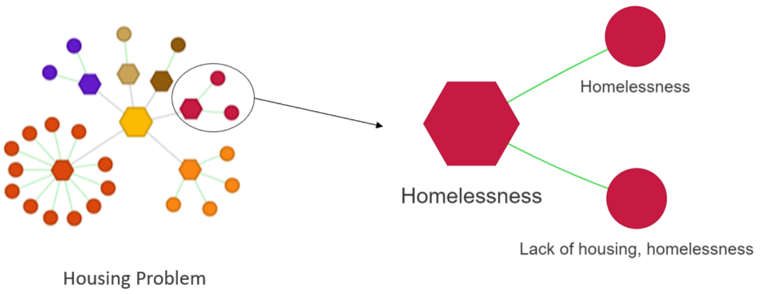
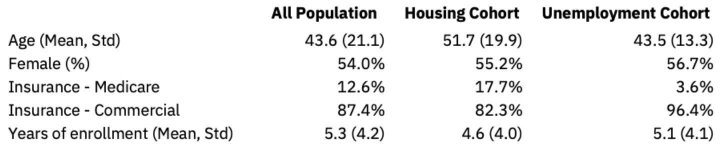


Figure 3: Concept network with ICD-9 and ICD-10 codes representing housing problems. Each hexagon represents a category and circles represent individual ICD codes. In the case of Homelessness they are ICD-10 Z59.0 and ICD-9 V60.0.

prescription drug information. Social aspects of patients' lives are captured by the V-codes in ICD9 and Z-codes in ICD10 in electronic health data, and both were used to identify recorded SDoH issues in patients' claims. Based on co-occurrence of codes, we extracted correlated diagnoses in the cohort of patients with that SDoH. For housing problems, we combined all housing related codes including Homelessness, Inadequate housing, Other specified housing or economic circumstances, etc. The ICD codes were grouped to represent more general diagnosis groups (Table 3) when summarizing and interpreting the results. By using the cohort created from all available CED data, we then create a very general cohort of people residing and receiving healthcare in the US (Table 1) - however, this can be replaced by any specific cohort data based on study settings.



	All Population	Housing Cohort	Unemployment Cohort
Age (Mean, Std)	43.6 (21.1)	51.7 (19.9)	43.5 (13.3)
Female (%)	54.0%	55.2%	56.7%
Insurance - Medicare	12.6%	17.7%	3.6%
Insurance - Commercial	87.4%	82.3%	96.4%
Years of enrollment (Mean, Std)	5.3 (4.2)	4.6 (4.0)	5.1 (4.1)

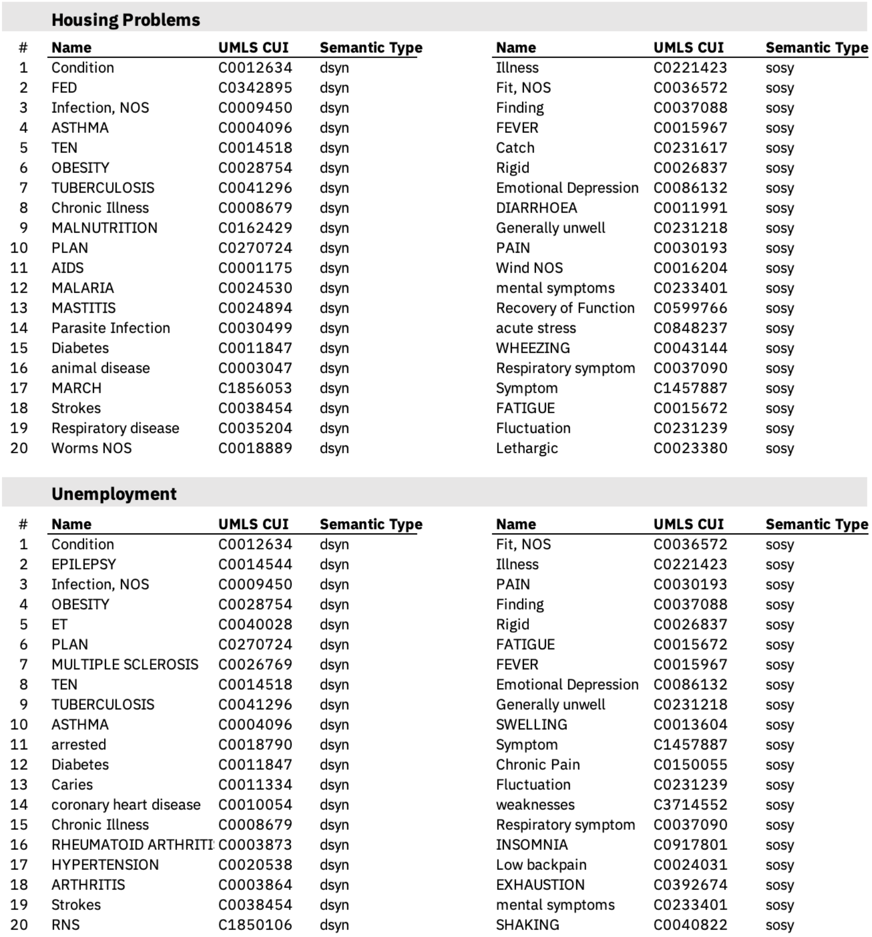
Table 1: Demographic characteristics of cohorts created from CED data

Quantification of Association

After creating two KGs from PubMed and CED data, we identified the overlapping region of the two graphs as Disease or Syndrome and Sign or Symptom types in PubMed graph, also captured in CED by ICD codes (Table 3). From PubMed, we identified the top 20 associations in terms of support for the two types of terminology. In CED data analysis, we identified top 10 most prevalent diagnoses in people with specific SDoH code. Knowledge extracted from PubMed covers a wider area of social and clinical relations due to the nature of unstructured data compared to the structured data of CED. Based on the result of knowledge extractions, we proceeded to quantify the association between each of the two SDoH (housing problems, unemployment) and health outcomes by comparing the group of people with and without the respective SDoH code in their claims. We focused on two disease conditions that appear in both PubMed top 20 and CED top 10 results, that also show significant difference in prevalence between SDoH-specific and general cohort. We examined whether having a SDoH code modifies the association between disease and health related outcomes by comparing the number of claims per patient per year and pay amount per claim per patient as proxies of the level of needs for healthcare. This analysis is not to identify causal relationship, but to generate hypothesis for more in-depth analysis tailored to each use case.

Results

Several relations were found from PubMed under Disease or Syndrome and Sign or Symptom types (Table 2). For housing problems, the majority of the disease terms were associated with infectious diseases (Infection, Tuberculosis, Malaria, Mastitis, Parasite Infection, or Worms) or chronic conditions (Asthma, Obesity, AIDS, Diabetes, Stroke, and Respiratory disease). A number of mental health related symptoms were identified such as Emotional Depression, Mental symptoms, or Acute stress. There were also a number of irrelevant or erroneous terms found as expected. Notably, common English words like *fed*, *ten* or *march* are annotated incorrectly as acronyms by MetaMap (e.g. FED for Fish Eye Disease) and their frequencies are skewed as a result. Similarly for unemployment, a diverse range of diagnosis and symptoms was extracted. Some degree of overlap with housing problems is observed but with less emphasis on infection-related diagnoses. These associated terms from a large pool of published research can generate a number of hypotheses regarding the relationship between SDoH and clinical conditions. For example, one can hypothesize that housing problems, such as homelessness, may expose a person to environments with higher probabilities of microbial infection; or that unemployment can put a person under severe mental stress that results in insomnia. One can also think about less obvious but inferred hypotheses - living conditions with housing insecurity may lead women to suffer from mastitis with little postpartum support; being unemployed can delay dental visits and lead to more people having caries. But these hypotheses are not yet specific to the population of interest and may not be applicable.



Housing Problems

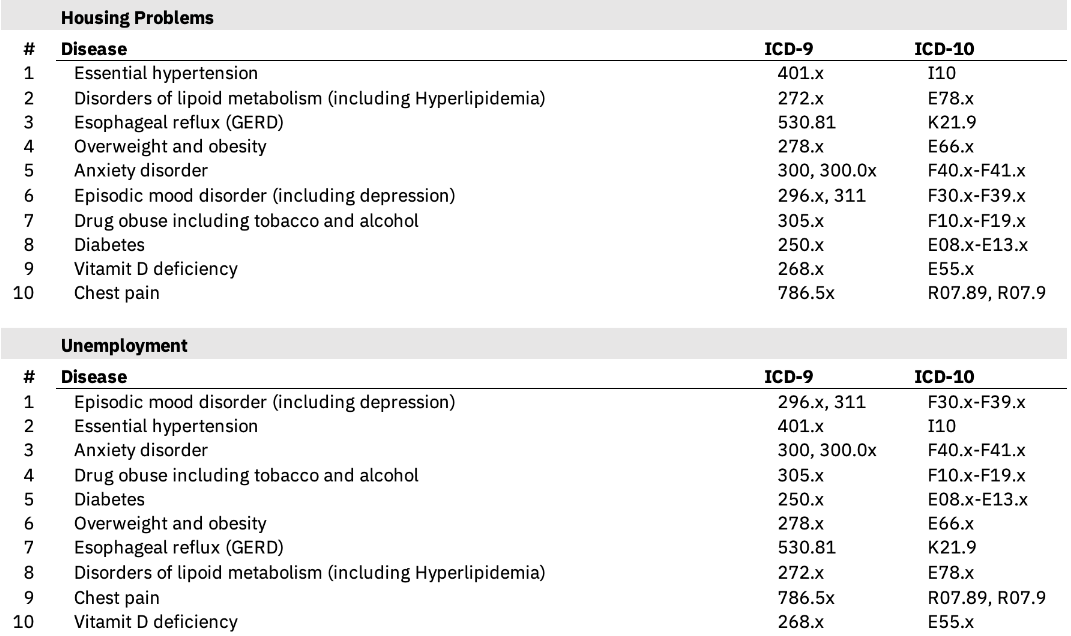
#	Name	UMLS CUI	Semantic Type	Name	UMLS CUI	Semantic Type
1	Condition	C0012634	dsyn	Illness	C0221423	sosy
2	FED	C0342895	dsyn	Fit, NOS	C0036572	sosy
3	Infection, NOS	C0009450	dsyn	Finding	C0037088	sosy
4	ASTHMA	C0004096	dsyn	FEVER	C0015967	sosy
5	TEN	C0014518	dsyn	Catch	C0231617	sosy
6	OBESITY	C0028754	dsyn	Rigid	C0026837	sosy
7	TUBERCULOSIS	C0041296	dsyn	Emotional Depression	C0086132	sosy
8	Chronic Illness	C0008679	dsyn	DIARRHOEA	C0011991	sosy
9	MALNUTRITION	C0162429	dsyn	Generally unwell	C0231218	sosy
10	PLAN	C0270724	dsyn	PAIN	C0030193	sosy
11	AIDS	C0001175	dsyn	Wind NOS	C0016204	sosy
12	MALARIA	C0024530	dsyn	mental symptoms	C0233401	sosy
13	MASTITIS	C0024894	dsyn	Recovery of Function	C0599766	sosy
14	Parasite Infection	C0030499	dsyn	acute stress	C0848237	sosy
15	Diabetes	C0011847	dsyn	WHEEZING	C0043144	sosy
16	animal disease	C0003047	dsyn	Respiratory symptom	C0037090	sosy
17	MARCH	C1856053	dsyn	Symptom	C1457887	sosy
18	Strokes	C0038454	dsyn	FATIGUE	C0015672	sosy
19	Respiratory disease	C0035204	dsyn	Fluctuation	C0231239	sosy
20	Worms NOS	C0018889	dsyn	Lethargic	C0023380	sosy

Unemployment

#	Name	UMLS CUI	Semantic Type	Name	UMLS CUI	Semantic Type
1	Condition	C0012634	dsyn	Fit, NOS	C0036572	sosy
2	EPILEPSY	C0014544	dsyn	Illness	C0221423	sosy
3	Infection, NOS	C0009450	dsyn	PAIN	C0030193	sosy
4	OBESITY	C0028754	dsyn	Finding	C0037088	sosy
5	ET	C0040028	dsyn	Rigid	C0026837	sosy
6	PLAN	C0270724	dsyn	FATIGUE	C0015672	sosy
7	MULTIPLE SCLEROSIS	C0026769	dsyn	FEVER	C0015967	sosy
8	TEN	C0014518	dsyn	Emotional Depression	C0086132	sosy
9	TUBERCULOSIS	C0041296	dsyn	Generally unwell	C0231218	sosy
10	ASTHMA	C0004096	dsyn	SWELLING	C0013604	sosy
11	arrested	C0018790	dsyn	Symptom	C1457887	sosy
12	Diabetes	C0011847	dsyn	Chronic Pain	C0150055	sosy
13	Caries	C0011334	dsyn	Fluctuation	C0231239	sosy
14	coronary heart disease	C0010054	dsyn	weaknesses	C3714552	sosy
15	Chronic Illness	C0008679	dsyn	Respiratory symptom	C0037090	sosy
16	RHEUMATOID ARTHRITI	C0003873	dsyn	INSOMNIA	C0917801	sosy
17	HYPERTENSION	C0020538	dsyn	Low backpain	C0024031	sosy
18	ARTHRITIS	C0003864	dsyn	EXHAUSTION	C0392674	sosy
19	Strokes	C0038454	dsyn	mental symptoms	C0233401	sosy
20	RNS	C1850106	dsyn	SHAKING	C0040822	sosy

Table 2: Top 20 Diseases or Syndromes (*dsyn*) and Signs or Symptoms (*sosy*) associated with Housing Problems (C0014003) and Unemployment (C0014003) in PubMed.

We then obtained the top 10 most prevalent diagnoses associated with either housing problems or unemployment from an analysis of CED data (Table 3). Unsurprisingly, many were chronic conditions known to be highly prevalent among general population such as hypertension, diabetes, hyperlipidemia, or reflux disease. However, an interesting observation was that the two SDoH generated the same list of top 10 diagnosis with different order of prevalence. Compared to housing problems cohort in which chronic physical conditions were the top 3 most prevalent, psychiatric diagnosis such as mood disorders or anxiety took place in the top 3 diagnoses in the unemployment cohort. The nature of health related challenges associated with a social factor may differ across different SDoH.



	Housing Problems		
#	**Disease**	**ICD-9**	**ICD-10**
1	Essential hypertension	401.x	I10
2	Disorders of lipoid metabolism (including Hyperlipidemia)	272.x	E78.x
3	Esophageal reflux (GERD)	530.81	K21.9
4	Overweight and obesity	278.x	E66.x
5	Anxiety disorder	300, 300.0x	F40.x-F41.x
6	Episodic mood disorder (including depression)	296.x, 311	F30.x-F39.x
7	Drug obuse including tobacco and alcohol	305.x	F10.x-F19.x
8	Diabetes	250.x	E08.x-E13.x
9	Vitamit D deficiency	268.x	E55.x
10	Chest pain	786.5x	R07.89, R07.9

	Unemployment		
#	**Disease**	**ICD-9**	**ICD-10**
1	Episodic mood disorder (including depression)	296.x, 311	F30.x-F39.x
2	Essential hypertension	401.x	I10
3	Anxiety disorder	300, 300.0x	F40.x-F41.x
4	Drug obuse including tobacco and alcohol	305.x	F10.x-F19.x
5	Diabetes	250.x	E08.x-E13.x
6	Overweight and obesity	278.x	E66.x
7	Esophageal reflux (GERD)	530.81	K21.9
8	Disorders of lipoid metabolism (including Hyperlipidemia)	272.x	E78.x
9	Chest pain	786.5x	R07.89, R07.9
10	Vitamit D deficiency	268.x	E55.x

Table 3: Top 10 most prevalent diseases associated with Housing problems and Unemployment in the Claims-EMR dataset.

Combining the two KGs, represented with the extracted relations, led to the identification of several concepts that are directly or indirectly linked to the concepts in the other KG. Disease concepts such as diabetes, obesity, and hypertension appear in both PubMed and CED results. Concepts from PubMed like emotional depression or acute stress can be linked to mood disorder diagnosis in CED. We calculated the prevalence of the top 10 diseases in the general CED population and in the cohort specified by each of the two SDoH examined. The largest relative difference in prevalence was observed for diabetes and episodic mood disorder (Table 4) so we focused on these two in the subsequent analysis.

The last part of our experiment was quantifying the level of association in the CED cohort. Specifically, we examined whether the healthcare cost and utilization level among those with a disease diagnosis differ across the general cohort and the subgroups of people with housing problems or unemployment codes. As Table 4 shows, for patients with diabetes diagnosis, having a code for housing problems was associated with greater outpatient cost (per year per patient) and outpatient service use measured by number of claims (per year per patient). Similarly, for patients with mood disorders diagnosis, having a code for housing problems was associated with greater outpatient cost and outpatient service use. Notably, emergency department cost was significantly higher for patients with housing problems or unemployment code compared to the general population, for both diabetic and mood disorder patients. On the other hand, we observed that inpatient service utilization and related cost was higher for the general CED population compared to the SDoH-specific subgroups.

Discussion and Conclusion

We describe a systematic approach to identify and quantify meaningful associations between SDoH and health related outcomes using published knowledge as well as electronic health data. The novelty of this work is the augmentation of knowledge extraction from PubMed by adding a cohort-specific, independent source of clinical data. The final set of pruned associations with the chosen SDoH can be informative and actionable from a decision maker point of view – whether it be a population health specialist in a hospital or a policy maker at a public health department – because it provides augmented insights specific to the population group of interest that may not be observable in data alone. As

		All Population (n=5,662,591)	Housing Cohort (n=9,835)	Unemployment Cohort (n=1,601)
Diabetes	N patients (%)	522944 (9.2%)	2156 (21.9%)	111 (6.9%)
	Inpatient cost	$27,436.26	$10,692.88	$15,096.53
	Inpatient N claims	49.5	24.8	45.3
	Outpatient cost	$7,044.90	$8,586.06	$8,438.46
	Outpatient N claims	54.1	69.3	77.9
	ED cost	$2,561.84	$4,470.04	$3,555.54
	ED N claims	12.3	23.5	16.8
Mood disorders	N patients (%)	620927 (11.0%)	2451 (24.9%)	335 (20.9%)
	Inpatient cost	$24,555.65	$10,511.55	$11,484.27
	Inpatient N claims	42.5	24.0	27.2
	Outpatient cost	$5,918.04	$7,669.82	$6,248.16
	Outpatient N claims	44.8	60.0	52.0
	ED cost	$1,085.70	$1,438.36	$1,679.53
	ED N claims	6.3	8.6	8.2

Table 4: Impact quantification results for housing problems and unemployment in CED data. All cost and number of claims measures are per patient per year values.

our approach does not attempt to identify causal relationship, evaluating the impact of SDoH should be followed by additional data collection or more rigorous analysis. Ongoing efforts to address the need for a more comprehensive SDoH terminology, such as methodologies to detect new SDoH concepts[17] or the Gravity project[19] should, in the future, provide additional structure to the work presented in this paper.

The ability of our proposed approach to identify meaningful SDoH associations from KGs were supported by literature findings. The increased emergency service utilization among people with housing problems we observed in the data has also been reported in prior literature[28]. The observation that unemployment is associated with higher prevalence of mental health diagnosis can also be supported through prior evidence.[35] The validation of selected hypotheses through literature means that other associations we observe in data are potentially worth further investigating. For example, one relation we observe from PubMed knowledge extraction is between housing problems and respiratory diseases or symptoms, which is not captured by the 10 most prevalent ICD codes in the CED data. While it is possible that this is just a spurious association, it can also be a meaningful association not captured due to incomplete data collection or low prevalence. Considering the disease burden and clinical significance of managing chronic respiratory diseases, it may be worth investigating this association further to see whether this association is true, and if so what may be the factors that contribute to both housing problems and diseases like asthma or COPD, such as geographic characteristics. There are extended application use cases for this approach. For example, housing problems may be selected as an input for outcome risk modelling to adjust for confounding. Also, it can be used as an input for predictions to improve the model performance.

The known limitations of observational health data are applicable for this study, including non-randomly missing diagnosis, incomplete capture of data such as lack of information on the duration of illness or SDoH, and coding errors or variability in coding practice. A claims data point is generated at the time of a patient's interaction with health system, so those who lack access to care, who will likely have more severe SDoH problems, may be absent from the data. By combining claims with the EMR data, information from CED can provide more complete picture of individual patients and increases the likelihood of capturing all data compared to using only claims.

References

1. Renee Mehra, F. Shebl, S. Cunningham, U. Magriples, Eric Barrette, Carolina Herrera, K. Kozhimannil, and J. Ickovics. Area-level deprivation and preterm birth: results from a national, commercially-insured population. *BMC Public Health*, 19, 2019.

2. J. Meddings, Heidi Reichert, S. Smith, T. Iwashyna, K. Langa, T. Hofer, and Laurence F. McMahon. The impact of disability and social determinants of health on condition-specific readmissions beyond medicare risk adjustments: A cohort study. *Journal of General Internal Medicine*, 32:71–80, 2016.

3. E. Hatef, Hadi Kharrazi, Karin Nelson, Philip Sylling, X. Ma, E. Lasser, K. Searle, Zachary Predmore, Adam J Batten, Idamay Curtis, S. Fihn, and J. Weiner. The association between neighborhood socioeconomic and housing characteristics with hospitalization: Results of a national study of veterans. *The Journal of the American Board of Family Medicine*, 32:890 – 903, 2019.

4. E. Hatef, X. Ma, M. Rouhizadeh, Gurmehar Singh, J. Weiner, and Hadi Kharrazi. Assessing the impact of social needs and social determinants of health on health care utilization: Using patient- and community-level data. *Population health management*, 2020.

5. Lauren Paremoer, S. Nandi, H. Serag, and Fran Baum. Covid-19 pandemic and the social determinants of health. *The BMJ*, 372, 2021.

6. J. Phelan, B. Link, and P. Tehranifar. Social conditions as fundamental causes of health inequalities: Theory, evidence, and policy implications. *Journal of Health and Social Behavior*, 51:S28 – S40, 2010.

7. L. Horwitz, Carol Chang, Harmony N Arcilla, and J. Knickman. Quantifying health systems' investment in social determinants of health, by sector, 2017-19. *Health affairs*, 39 2:192–198, 2020.

8. T. Veinot, J. Ancker, and S. Bakken. Health informatics and health equity: improving our reach and impact. *Journal of the American Medical Informatics Association : JAMIA*, 2019.

9. Matthew S. Pantell, Julia Adler-Milstein, Michael D Wang, A. Prather, N. Adler, and L. Gottlieb. A call for social informatics. *Journal of the American Medical Informatics Association : JAMIA*, 27 11:1798–1801, 2020.

10. G. F. Murray, Hector P. Rodriguez, and V. Lewis. Upstream with a small paddle: How acos are working against the current to meet patients' social needs. *Health affairs*, 39 2:199–206, 2020.

11. W. Weeks, Stacey Y Cao, Chris M Lester, J. Weinstein, and N. Morden. Use of z-codes to record social determinants of health among fee-for-service medicare beneficiaries in 2017. *Journal of General Internal Medicine*, 35:952–955, 2019.

12. J. Mathew, C. Hodge, and M. Khau. Z codes utilization among medicare fee-for-service (ffs) beneficiaries in 2017. *CMS OMH Data Highlight No. 17. Baltimore, MD: CMS Office of Minority Health*, 2019.

13. Mirjam Allik, A. Leyland, Maria Yury Travassos Ichihara, and R. Dundas. Creating small-area deprivation indices: a guide for stages and options. *Journal of Epidemiology and Community Health*, 74:20 – 25, 2019.

14. G. Singh. Area deprivation and widening inequalities in us mortality, 1969-1998. *American journal of public health*, 93 7:1137–43, 2003.

15. Danielle C Butler, S. Petterson, R. Phillips, and A. Bazemore. Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. *Health services research*, 48 2 Pt 1:539–59, 2013.

16. M. Cantor and L. Thorpe. Integrating data on social determinants of health into electronic health records. *Health affairs*, 37 4:585–590, 2018.

17. J. H. Bettencourt-Silva, N. Mulligan, M. Sbodio, J. Segrave-Daly, R. Williams, V. Lopez, and C. Alzate. Discovering New Social Determinants of Health Concepts from Unstructured Data: Framework and Evaluation. *Stud Health Technol Inform*, 270:173–177, Jun 2020.

18. Arons A, Desilvey S, Fichtenberg C, and Gottlieb L. Improving the interoperability of social determinants data in electronic health records: working paper for november 2017 expert panel convening, 2017.

19. Health Level Seven International. Gravity project. url: https://www.hl7.org/gravity. (accessed: 07.05.2021).

20. M. Watkins, B. Viernes, V. Nguyen, L. Rojas Mezarina, J. Silva Valencia, and D. Borbolla. Translating Social Determinants of Health into Standardized Clinical Entities. *Stud Health Technol Inform*, 270:474–478, Jun 2020.

21. Matthew Herland, T. Khoshgoftaar, and Randall Wald. A review of data mining using big data in health informatics. *Journal Of Big Data*, 1:1–35, 2013.

22. J. Vest, S. Grannis, Dawn P. Haut, P. Halverson, and N. Menachemi. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *International journal of medical informatics*, 107:101–106, 2017.

23. C. Luque, J. Luna, M. Luque, and Sebastián Ventura. An advanced review on text mining in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9, 2019.

24. C. Bejan, J. Angiolillo, D. Conway, R. Nash, J. Shirey-Rice, Loren Lipworth-Elliot, R. Cronin, J. Pulley, S. Kripalani, S. Barkin, K. Johnson, and J. Denny. Large-scale text mining of social determinants from electronic health records: Case studies of homelessness and adverse childhood experiences. In *AMIA*, 2017.

25. C. A. Bejan, J. Angiolillo, D. Conway, R. Nash, J. K. Shirey-Rice, L. Lipworth, R. M. Cronin, J. Pulley, S. Kripalani, S. Barkin, K. B. Johnson, and J. C. Denny. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc*, 25(1):61–71, 01 2018.

26. N. M. Oreskovic, J. Maniates, J. Weilburg, and G. Choy. Optimizing the Use of Electronic Health Records to Identify High-Risk Psychosocial Determinants of Health. *JMIR Med Inform*, 5(3):e25, Aug 2017.

27. JH Bettencourt-Silva, N Mulligan, C Jochim, N Yadav, W Sedlazek, V Lopez, and M Gleize. Exploring the social drivers of health during a pandemic: Leveraging knowledge graphs and population trends in covid-19. *Studies in Health Technology and Informatics*, 275:6–11, 2020.

28. S. A. Berkowitz, S. Kalkhoran, S. T. Edwards, U. R. Essien, and T. P. Baggett. Unstable Housing and Diabetes-Related Emergency Department Visits and Hospitalization: A Nationally Representative Study of Safety-Net Clinic Patients. *Diabetes Care*, 41(5):933–939, 05 2018.

29. W. Kawohl and C. Nordt. COVID-19, unemployment, and suicide. *Lancet Psychiatry*, 7(5):389–390, 05 2020.

30. Martin Theobald, Nigam Shah, and Jeff Shrager. Extraction of conditional probabilities of the relationships between drugs, diseases, and genes from PubMed guided by relationships in PharmGKB. In *2009 AMIA Summit on Translational Bioinformatics*, pages 124–128. American Medical Informatics Association, AMIA, 2009.

31. Alan Aronson and François-Michel Lang. An overview of metamap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17:229–36, 05 2010.

32. Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

33. Alexa T McCray. An upper-level ontology for the biomedical domain. *Comparative and Functional Genomics*, 4(1):80–84, 2003.

34. Y. Park, H. Yang, A. K. Das, and G. Yuen-Reed. Prescription fill rates for acute and chronic medications in claims-EMR linked data. *Medicine (Baltimore)*, 97(44):e13110, Nov 2018.

35. A. Zuelke, T. Luck, M. Schroeter, A. V. Witte, A. Hinz, C. Engel, Cornelia Enzenbach, S. Zachariae, M. Loeffler, J. Thiery, A. Villringer, and S. Riedel-Heller. The association between unemployment and depression-results from the population-based life-adult-study. *Journal of affective disorders*, 235:399–406, 2018.