Implementation and Validation of the Use of N-Gram Models for Classifying Social Determinants of Health Status from Real World Unstructured Clinical Notes

Vikas Kumar | Amanda Mummert | Tracy d'Arbeloff | Lawrence Rasouliyan | Amanda G. Althoff | Danae A. Black | Stella Chang | Stacey Long
OMNY Health | Atlanta, GA, United States

MSR82

BACKGROUND

- Social determinants of health (SDoH) account for up to 50 percent of variation in health outcomes in the United States.[1]
- Unstructured electronic health records can supplement underutilized clinical codes to provide a more accurate portrayal of economic and social vulnerabilities.
- SDoH factors include housing, food and nutrition, transportation, and education. Natural language processing (NLP) algorithms categorize key SDoH domains to leverage unstructured EHR.
- Use of n-gram models enables the extraction and classification of a contiguous sequence of words from clinical notes to applicable SDoH domains.[2]

OBJECTIVE

- The objective of this research was to implement and validate the use of n-gram models using unstructured clinical notes to classify patient risk status across five key SDoH domains.

METHODS

- Deidentified clinical notes from three hospital systems (2017-2022) in the OMNY Health real-world data platform were examined for the presence of indicative phrases previously published [2] for four SDoH insecurity domains (economics [EI]; housing [HI]; social environment [SI]; and transportation [TI]). For a fifth domain (undereducation [ED]), n-grams were composed.
- To measure precision, 50 random positive occurrences ("hits") from each domain-hospital combination were manually annotated.
- Models and included notes were iteratively refined until model precision met a pre-defined threshold (80%).
- Final patient, encounter, and mention counts were calculated tabulated.
- Recall was estimated using the overlap between hits and presence of SDoH domain-specific ICD-10 diagnosis codes.

REFERENCES

1. Whitman, A., De Lew, N., Chappel, A., Aysola, V., Zuckerman, R., Sommers, B. D. (2022) Addressing Social Determinants of Health: Examples of Successful Evidence-Based Strategies and Current Federal Efforts. Department of Health & Human Services, USA: Office of Health Policy.(HP-2022-12) https://aspe.hhs.gov/sites/default/files/documents/e2b650cd64cf84aae8ff0fae7474af82/SDOH-Evidence-Review.pdf
2. Wu W, Holkeboer KJ, Kolawole TO, Carbone L, Mahmoudi E. Natural language processing to identify social determinants of health in Alzheimer's disease and related dementia from electronic health records. *Health Serv Res*. 2023 Dec;58(6):1292-1302. doi: 10.1111/1475-6773.14210

RESULTS

- Clinical notes from 9.34 million patients were available for assessment.
- 703,235 unique patients and 1,644,376 unique encounters had a mention of an insecurity phrase.
- The volume of mentions of insecurity phrases varied by domain. Economic insecurity phrases were documented most frequently (2,806,465), followed by housing insecurity (677,349), and social isolation (171,440) [Table 1].

Table 1: Overall Positive Phrases ("Hits") by SDoH Domain and Associated Encounters and Patients

SDoH Domains	Positive Mentions Overall ("Hits")	Unique Encounters	Unique Patients
Economics (EI)	2,806,465	1,472,833	628,187
Housing (HI)	677,349	198,963	100,437
Social Environment (SI)	768,474	282,014	154,373
Transportation (TI)	171,440	125,772	91,341
Undereducation (ED)	18,370	12,142	8,984

Figure 1: OMNY Health Social Determinants of Health (SDoH) Domains and Corresponding ICD-10-CM Codes Leveraged

Economic
- Economic instability indicators (i.e., low income, poverty)
- ICD-10: Z59.4, Z59.5, Z59.6, Z59.7, Z59.86, Z59.87, Z59.89

Housing
- Housing insecurity (i.e., homelessness, inadequate housing)
- Z59.0, Z59.1, Z59.2, Z59.3, Z59.81, Z59.89

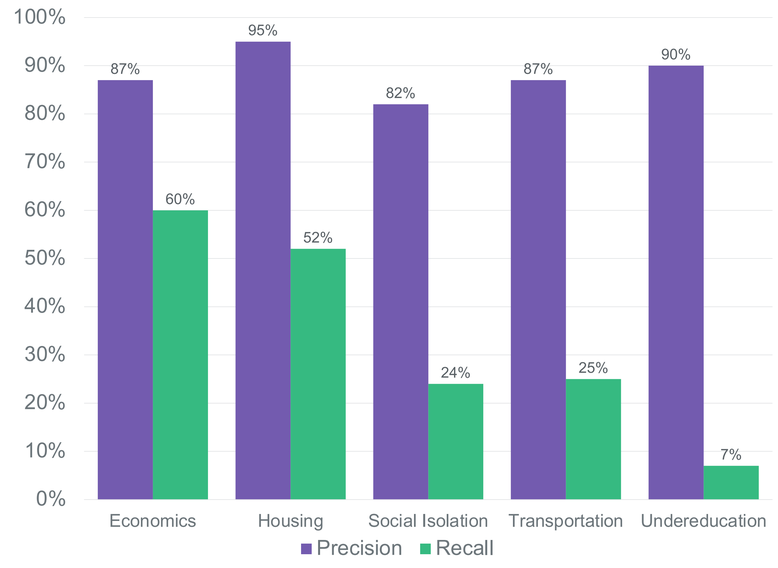
Social Environment
- Problems related to social environment (i.e., isolation, discrimination)
- Z60.0 to Z60.9

Transportation
- Adequate transportation availability
- Z59.82

Undereducation
- Problems related to education and literacy (i.e., low literacy, less than high school education)
- Z55.0 to Z55.9

- All models were successfully trained to meet and exceed the pre-defined 80% threshold.
- Precision and recall were highest for the Economic domain (87% precision, 60% recall) and the Housing domain (95% precision, 52% recall) [Figure 2].

Figure 2: Precision and Recall across the Five OMNY Health SDoH Domains



DISCUSSIONS AND CONCLUSIONS

- 7.5% of patients had positive hits in an SDoH domain when using unstructured clinical notes and n-grams.
- Drawbacks relative to transformer-based techniques may include lower precision/recall ceilings, due to factors including false-positive hits from negated n-grams and exact term mismatch.
- Recall may be artificially reduced due to provide under-utilization of available ICD-10 Z codes.
- This method generated a patient-level indicator of SDoH risk status based on information collected during a clinical encounter that can be leveraged in health economics and outcomes research.

CONTACT INFORMATION

Vikas Kumar, MD MS | Principal Data Scientist | **OMNY Health** |
Email: vikas@omnyhealth.com | Website: www.omnyhealth.com

Presented at ISPOR 2024: May 5-8, 2024; Atlanta, GA, USA