Clinical Validation of an N-Gram Model for Detecting Economic Insecurity in Patient Populations Using Unstructured Clinical Notes

Vikas Kumar | Tracy d'Arbeloff | Amanda Mummert | Lawrence Rasouliyan
OMNY Health | Atlanta, GA, United States

MSR62

BACKGROUND

- Social determinants of health (SDoH) account for up to 50 percent of variation in health outcomes in the United States.[1]
- Natural language processing (NLP) applied to unstructured electronic health records (EHRs) can supplement underutilized or non-existent clinical codes to portray individual-level economic and social vulnerabilities.
- N-gram models enable the extraction and classification of a contiguous sequence of words from clinical notes to applicable SDoH domains.[2]
- While several n-gram models for identifying SDoH factors have been developed, their credibility and validity have not yet been verified.

OBJECTIVE

- Our objective was to apply and validate a previously developed n-gram model [2] for predicting economic instability to unstructured clinical notes from large health systems.

METHODS

- Unstructured clinical notes (2017-2022) from encounters across three hospital systems in the OMNY Health real-world data platform were examined for the presence of phrases indicative of economic insecurity as published previously.[2]
- Model performance and clinical validity were measured by assessing frequency of predictive terms and geographic distribution of those terms relative to United States Census data.
- Clinical characteristics such as age, gender, payer type, urban/rural location, and employment status were compared between patient encounters positively identified for economic insecurity compared to all encounters.

REFERENCES

1. Whitman A, De Lew N, Chappel A, Aysola V, Zuckerman R, Sommers BD. (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. 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.
3. Kumar V, Mummert A, d'Arbeloff T, et al. Implementation and Validation of the Use of N-gram Models for Classifying Social Determinants of Health Status from Real World Unstructured Clinical Notes. Poster presentation at ISPOR 2024: May 5-8, 2024; Atlanta, GA, USA.
4. Fienberg B. A dictionary of 3-digit zip code prefixes and associated metro areas. 2020. https://github.com/billfienberg/zip3/blob/master/zip3.csv
5. US Census Data. 2022 https://data.census.gov

RESULTS

- 9,460,313 patients with 126,306,593 patient-encounters with unstructured clinical notes were assessed. 628,187 unique patients (6.6%) had 1+ notes with an economic insecurity phrase. 1.47 million encounters (1.2%) were positive for economic insecurity phrases.
- Notes for encounters associated with economic insecurity had 7.7 times as many characters (mean, 47,392; SD, 244,941) compared to all notes (mean, 6,109; SD, 42,774).
- The five most frequent phrases predicting a positive indication of economic insecurity were: "Medicaid," "homeless," "shelter," "too expensive," and "not covered by insurance." [Table 1].

Table 1: Most Frequent N-grams Appearing in the Notes of Patients Positive for Economic Insecurity

Rank	N-gram	Frequency
1	Medicaid	1,170,043
2	Homeless	406,393
3	Shelter	176,339
4	Too expensive	142,495
5	Not covered by insurance	93,917
6	Unable to afford	84,716
7	Financial assistance	77,540
8	Meals on Wheels	76,053

Note: N-grams included all case variants of the terms

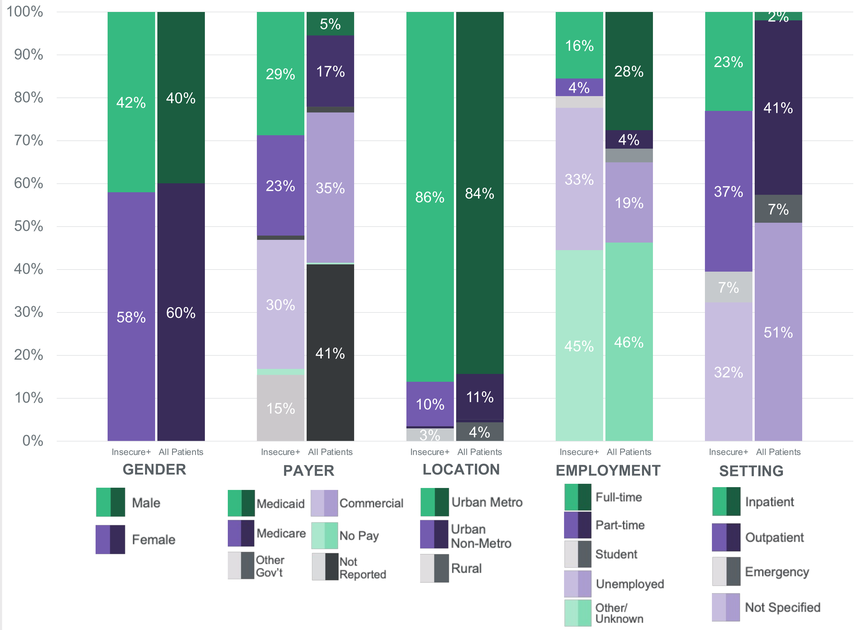
- In 2022, the median US household income was $74,580 and Federal poverty threshold was $22,556.[5] Among the ten ZIP-3 locations with the most positive encounters for economic instability:
  - All but one (180, Lehigh Valley, PA) were below the US median.
  - None were below the Federal poverty threshold [Table 2].

Table 2: Most Frequent 3-Digit ZIP Codes of Patients Positive for Economic Insecurity

Rank	3 Digit ZIP – City, State[4]	Frequency	Median Income (Census)
1	436 - Toledo, OH	117,290	$48,339
2	232 - Richmond, VA	116,723	$66,739
3	452 - Cincinnati, OH	96,417	$65,687
4	180 - Lehigh Valley, PA	91,853	$80,510
5	458 - Lima, OH	75,580	$62,417
6	296 - Greenville, SC	58,379	$60,321
7	237 - Portsmouth, VA	57,263	$53,775
8	451 - Cincinnati, OH	49,760	$71,105
9	238 - Petersburg, VA	49,541	$68,571
10	440 - Cleveland, OH	43,680	$71,625

- Encounters associated with economic insecurity were more likely to have Medicaid (28.7%), Medicare (23.4%), or no payer (1.4%), as payer type and unemployed as employment status (33.3%) as compared to all encounters (5.4%, 16.7%, 0.4% and 18.7%, respectively) [Figure 1].
- A higher frequency of encounters associated with economic insecurity occurred in inpatient settings (23.0% vs. 1.9%) [Figure 1].

Figure 1: Distribution of Encounter-Level Demographic Characteristics for Entire Sample ("All") vs. Economically Insecure



DISCUSSION AND CONCLUSIONS

- Results show that positive economic insecurity using an n-gram model is confounded by the note length.
- Qualitative examination of the most common n-grams and demographic characteristics align with intuition. Patients with encounters associated with mentions of economic insecurity tended to come from ZIP codes with lower median incomes, but this may reflect the underlying distribution of patients in the hospital systems.

CONTACT INFORMATION

Vikas Kumar, MD, MS | Principal Data Scientist | vikas@omnyhealth.com

Amanda Mummert, PhD | Senior Director Customer Success | Amanda.Mummert@omnyhealth.com

OMNY Health | www.omnyhealth.com

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