The ENACT network is acting on housing instability and the unhoused using the open health natural language processing toolkit

Daniel R. Harris[1,2] , Sunyang Fu[3], Andrew Wen[3], Alexandria Corbeau[1,2], Darren Henderson[1,2], Jordan Hilsman[4], David Oniani[4] and Yanshan Wang[4,5,6]

[1]Center for Clinical and Translational Sciences, University of Kentucky, Lexington, KY, USA; [2]Institute for Biomedical Informatics, University of Kentucky, Lexington, KY, USA; [3]Center for Translational AI Excellence and Applications in Medicine, University of Texas Health Science Center at Houston, Houston, TX, USA; [4]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, USA; [5]Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, USA and [6]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Housing is an environmental social determinant of health that is linked to mortality and clinical outcomes. We developed a lexicon of housing-related concepts and rule-based natural language processing methods for identifying these housing-related concepts within clinical text. We piloted our methods on several test cohorts: a synthetic cohort generated by ChatGPT for initial infrastructure testing, a cohort with substance use disorders (SUD), and a cohort diagnosed with problems related to housing and economic circumstances (HEC). Our methods successfully identified housing concepts in our ChatGPT notes (recall = 1.0, precision = 1.0), our SUD population (recall = 0.9798, precision = 0.9898), and our HEC population (recall = N/A, precision = 0.9160).

Introduction

Environmental social determinants of health (SDOH), such as one's living circumstances and housing stability, significantly impact a person's overall health, and the lack of stable housing can lead to serious adverse effects [1]. Housing directly impacts a person's access and means to medical care; lack of housing is related to a disproportionately higher reliance on emergency medical services and ambulance transports [2]. The most severe manifestation of housing instability, known as housing deprivation or homelessness, can reduce life expectancy by as much as 12 years and increase rates of illness or disability [3,4]. Experiencing homelessness is linked to notably increased rates of hospital readmissions and extended hospital stays [5].

Despite the importance of housing and its relevance to health, housing issues are underreported in electronic health records (EHRs) due to a lack of national standards, social stigma, and reliance on self-reporting [6]. A previous study on housing found that diagnosis codes used for billing only identified 58.5% of the population experiencing housing instability or homelessness [7]; the remaining population was only identifiable through clinical notes or address data [7]. Clinical text combined with natural language processing (NLP) techniques may assist in identifying housing issues from unstructured data in EHRs [8–12]. We extend these housing-related techniques and findings as part of a national effort to capture housing-related concepts.

The Evolve to Next-Gen Accrual to Clinical Trials (ENACT) Network spans the Clinical and Translational Science Award (CTSA) consortium and connects CTSA sites with a single interface capable of querying over 142 million patients using the ENACT web-based query tool [13]. One of the goals of ENACT is to allow informatics researchers to develop and validate new EHR research tools; a working group (WG) for developing NLP tools was established across participating ENACT sites. This paper outlines the WG's progress on using clinical text to help identify housing issues and to supplement the known gap of underreported housing instability in structured clinical data by using NLP with unstructured EHR data. We present our custom lexicon of housing-related terms constructed after a literature review and discuss the performance of our initial implementation using three unique data sets.

Materials and methods

Lexicon development

We conducted a literature review of studies involving housing instability and homelessness to identify relevant works and to help construct a lexicon of housing-related terms [7,12,14,15].

Table 1. Housing-related concepts and phrases

Concept	Keywords or phrases
Homeless	Houseless, homeless, houselessness, homelessness, unhoused, unsheltered, unsheltered homelessness, lack of housing, lack of shelter, lacking housing, lacking shelter, housing lack, homelesss
Unstable housing	Vulnerably housed, unstably housed, unstable housed, unstable housing
Recovery housing	Recovery housing, recovery residence
Emergency housing	Emergency housing, emergency shelter
Temporary housing	Transitional housing, temporarily staying, itinerant, shelter, couch surfing, lodgers
Exposure	Evicted, living on the streets, residential institution problems, sleeping outside, sleeping outdoors

An existing Open Health Natural Language Processing (OHNLP) project on food and housing insecurity was reviewed to compare important words, phrases, and patterns [16]. We organized our relevant findings into six concepts: homeless, unstable housing, recovery housing, emergency housing, temporary housing, and exposure. These concepts and associated phrases are summarized in Table 1 and were selected to support fine-grain querying of housing in the ENACT query tool and clinical trial recruitment.

Algorithm development

We developed patterns to identify housing-related issues that were compatible with the OHNLP Toolkit, developed by the OHNLP consortium, for automated concept extraction from clinical notes [17]. The OHNLP Toolkit was selected due to its customizable interfaces which could support NLP efforts in multiple domains, including those beyond housing, and its easy integration with the ENACT query tool. This toolkit utilizes MedTagger, a lightweight tool for indexing based on dictionaries and patterns as the core component for information extraction [17,18]. The phrases in Table 1 were converted to regular expressions, which compressed the list. For example, "lack of housing" and "lack of shelter" are reduced to one pattern: "lack of (shelter|housing)." Furthermore, patterns were developed to allow flexible matching. For example, "living on the streets" became "living on the (the)? street(s)," where the article "the" is optional and streets may be plural or not. A common misspelling of "homeless" as "homelesss" was added based on the observational experience of the team in the housing domain.

The OHNLP Toolkit uses an expanded version of the ConText algorithm to classify whether identified entities are negated or part of a patient's medical history [19]. Irrelevant ConText rules, such as "did not demonstrate," were removed from the rule list to avoid wrongly negating detected entities; housing issues are not items for which patients test positive or negative. Each clinical document was divided into sentences as a preprocessing step; this was necessary after observing hits with negations generated from the wrong contextual window. These sentences were input into the OHNLP Toolkit, and annotated text files of the results were produced as output.

Testing

Each participating implementation site developed its own test data. For piloting the implementation, we developed a collection of emergency department notes using ChatGPT 3.5 that could be shared across sites for testing purposes. The initial question was "Can you write a sample discharge note from an Emergency Department for a homeless person?" Several additional prompts were used to generate positive cases in which the hypothetical patient has housing problems and negative cases in which there is no housing concern ("Can you generate a report for someone who is not homeless and not experiencing housing instability?").

We repurposed a selection of 250 documents from a related study on housing within a cohort with substance use disorders (SUD) specific to stimulant and opioid use disorders (randomly sampled from individuals having ICD-10-CM diagnosis codes of F11.*, F14.*, F15.*, T40.[1-6].*, and T43.6*) [7]. These documents were manually annotated as positive or negative for housing issues; patients experiencing housing issues have higher rates of SUD and are at higher risk of overdose, which highlights the importance of housing as an SDOH [20]. We also created a collection of 24,917 documents from a cohort ($n = 225$) diagnosed (ICD-10-CM Z59.6) with problems related to housing and economic circumstances (HEC) from UT Physicians, a multispecialty medical group associated with the University of Texas Health Sciences Center at Houston (UTHealth) and the UTHealth Harris County Psychiatric Center.

Results

The results of running the OHNLP Toolkit with our custom ENACT rule set and custom patterns on our three test data sets are summarized in Table 2. The results of MedTagger contain a flag for negation per each hit; a note was considered a positive case for housing issues if any of its hits were positive. True negatives were cases that either had no documented housing issues or all mentions of housing were negated. For the HEC cohort, the extracted hits were reviewed for correctness, so only precision is reported.

Table 3 lists common errors observed. For the SUD collection, false positives mostly stemmed from "Patient Education" notes that list dozens of community resources available for any patient; the "Homeless Veterans Center" caused false positives as it was only listed as a generic resource and did not imply the patient was a homeless veteran. Another false positive stemmed from a note describing someone who visited the emergency department after finding "a homeless person sleeping in her bathroom."

For the SUD cohort, there were only 2 false negatives at the note level, where all individual hits were negated, and at the individual hit level, there were 10 false negatives; these hits are described in Table 3. These examples are all failures to understand what concept is being negated in the sentence. For example, in "*Patient is not safe candidate for home IV abx therapy given active IVDA and homelessness,*" the concept of a safe candidate is intended to be negated instead of homelessness.

We report the distribution of concepts identified for each cohort in Table 4. The most frequently identified concept across all cohorts was homelessness. The second most frequent concept

Table 2. Performance using ENACT NLP rules

Cohort		Present (actual)	Absent (actual)	Recall	Precision
ChatGPT	**Present (rules)**	25 (TP)	0 (FP)	1.0	1.0
	Absent (rules)	0 (FN)	10 (TN)		
SUD	**Present (rules)**	194 (TP)	4 (FP)	0.9798	0.9898
	Absent (rules)	2 (FN)	50 (TN)		
HEC	**Present (rules)**	403 (TP)	37 (FP)	N/A	0.9159
	Absent (rules)	N/A	N/A		

ENACT = Evolve to Next-Gen Accrual to Clinical Trials; FN = false negative; FP = false positive; HEC = housing and economic circumstances; NLP = natural language processing; SUD = substance use disorders; TN = true negative; TP = true positive.

Table 3. Examples of errors

Cohort	Error	Sentence
SUD	False positive	"... found a **homeless** person sleeping in her bathroom."
	False positive	"**Homeless** Veterans Center" (listed in a directory of community resources, but this patient had no documentation of experiencing housing issues.)
SUD	False negative	"He does **not** want to return to the community homeless shelter where he was living, and has no money, transportation or job."
	False negative	"Patient is **not** safe candidate for home IV abx therapy given active IVDA and homelessness."
	False negative	"He does **not** want to return to his previous living situation due to the drug use in the home but also refuses to go to a homeless shelter and has nowhere else to go."
	False negative	"Patient **not** reveal to staff that she was homeless, but per (transferred hospital) patient stated she was homeless."
	False negative	"Pt **not** agreeable to homeless shelter and requested Ronald McDonald House."
	False negative	"Patient reports he presents because he believes that he needs mental health services related to frequent "outbursts" which have gotten him in trouble in the past, and resulted in him **not** being able to stay with his father or mother, and thus being homeless."
	False negative	"Patient is **not** safe candidate for home IV abx therapy given active IVDA and homelessness – will need to be kept in monitored environment for duration of IV antibiotics."
	False negative	"Regarding TB risk factors – the patient was homeless for 4 months recently, but did **not** stay in a homeless shelter, he has never worked in healthcare, he was in jail in the 1980s."
HEC	False positive	"Son was homeless and patient was sending food."
HEC	False positive	"Unstable Housing in the Last Year: No"
HEC	False positive	"They do not remove people from the home in effort to avoid adding to the homeless population."

HEC = housing and economic circumstances; SUD = substance use disorders.

Table 4. Distribution of concepts per data set

Concept	Identified concept count		
	ChatGPT	SUD	HEC
Exposure	4 (4.4%)	2 (0.5%)	8 (1.8%)
Homeless	50 (54.9%)	333 (84.3%)	373 (84.8%)
Housing instability	26 (28.6%)	0 (0.0%)	3 (0.7%)
Temporary housing	5 (5.5%)	57 (14.4%)	2 (0.5%)
Unstable housing	6 (6.6%)	3 (0.75%)	54 (12.2%)
Grand total	91	395	440

HEC = housing and economic circumstances; SUD = substance use disorders.

varied across cohorts. Temporary housing was likely popular in the SUD cohort due to a large number of patients staying in shelters; the HEC cohort was a general population where unstable housing may be more common than staying in a shelter.

Discussion

Our pilot suggests that developing a lexicon for housing-related issues and rule-based NLP methods for identifying housing concepts in unstructured EHR data is a realistic goal for the ENACT Network. The OHNLP platform is easily deployable and customizable by any ENACT site. The OHNLP toolkit can be customized to read and write to any database; the input can be

clinical data warehouses containing the clinical notes and the output can be the ENACT database that stores the searchable patient observations.

The ENACT web-based query tool is based on a browsable ontology that organizes concepts and codes that can be used in a "drag and drop" fashion. Our housing results are searchable on two tiers: the overall housing concept and the embedded individual concepts described in Table 1.

Our ChatGPT performance was without error; this performance is largely unrealistic and likely a reflection of how formulaic ChatGPT output appears. Additionally, ChatGPT occasionally documented negative cases of housing issues as "no visible signs of homelessness" which is highly unlikely to occur in a real note; if a patient does not appear homeless, the clinical documentation for homelessness would simply be absent. The phrase "no visible signs of homelessness" may be pejorative if included in clinical documentation. Despite these limitations, the ChatGPT notes are useful for prototyping and setting up the infrastructure needed to run MedTagger and to interface with the ENACT Network. We leave improving ChatGPT's formulaic responses as future work where prompt engineering could potentially produce a more realistic data set. We leave exploring the role of generative models in identifying housing issues as future work.

Our SUD results highlighted a false positive where the note references an unhoused individual who is not the patient; this example would be difficult to fix using rule-based methods as there are very little contextual clues or markers in these sentences to emphasize the unhoused person was not the patient. The HEC results highlighted an example where the patient's family member was experiencing homelessness, which may be addressable by fine-tuning the ConText algorithm to correctly identify family history.

Our study is limited by the breadth and depth of our housing lexicon. Although our intent was to be comprehensive, there may exist phrases or patterns that were not found during our literature review or during our tests. Furthermore, the language used to describe patients experiencing housing problems may change over time. We did not study recall in the HEC cohort due to the large number of notes; a smaller sampling strategy may be needed to manually review and validate recall. We also did not evaluate the temporality of the housing concepts or occurrences of where stable housing is explicitly mentioned.

Conclusion

The ENACT Network is based largely on querying structured, standardized codes; diagnostic billing codes are insufficient for identifying patients experiencing housing instability or homelessness. We designed our housing lexicon and rule-based NLP methods based on a literature review of other studies and how they reference housing issues. We piloted our methods across a small group of ENACT sites and will be moving to implement these findings as routine updates to the entire ENACT Network, where cohort size estimates can be calculated across sites and in support of innovation clinical trials involving those experiencing housing instability.

References

1. **Kushel MB, Gupta R, Gee L, Haas JS.** Housing instability and food insecurity as barriers to health care among low-income Americans. *J Gen Intern Med*. 2006;**21**(1):71–77. doi: 10.1111/j.1525-1497.2005.00278.x.

2. **Abramson TM, Sanko S, Eckstein M.** Emergency medical services utilization by homeless patients. *Prehosp Emerg Care*. 2021;**25**(3):333–340. doi: 10.1080/10903127.2020.1777234.

3. **National Health Care for the Homeless Council. homelessness-and-health.pdf**. Homelessness and Health: What's the connection? Published February 1, 2019. https://nhchc.org/wp-content/uploads/2019/08/homelessness-and-health.pdf. Accessed November 22, 2022

4. **Kushel MB, Vittinghoff E, Haas JS.** Factors associated with the health care utilization of homeless persons. *JAMA*. 2001;**285**(2):200–206. doi: 10.1001/jama.285.2.200.

5. **Khatana SAM, Wadhera RK, Choi E, et al.** Association of homelessness with hospital readmissions—an analysis of three large states. *J Gen Intern Med*. 2020;**35**(9):2576–2583. doi: 10.1007/s11606-020-05946-4.

6. **Brown RT, Steinman MA.** Characteristics of emergency department visits by older versus younger homeless adults in the United States. *Am J Public Health*. 2013;**103**(6):1046–1051. doi: 10.2105/AJPH.2012.301006.

7. **Harris DR, Anthony N, Quesinberry D, Delcher C.** Evidence of housing instability identified by addresses, clinical notes, and diagnostic codes in a real-world population with substance use disorders. *J Clin Transl Sci*. 2023;**7**(1):e196. doi: 10.1017/cts.2023.626.

8. **Bejan CA, Angiolillo J, Conway D, et al.** 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*. 2018;**25**(1):61–71. doi: 10.1093/jamia/ocx059.

9. **Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R.** Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA Open*. 2021;**4**(3):ooaa069. doi: 10.1093/jamiaopen/ooaa069.

10. **Gundlapalli AV, Carter ME, Palmer M, et al.** Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. *AMIA Annu Symp Proc*. 2013;**2013**:537–546.

11. **Hatef E, Rouhizadeh M, Nau C, et al.** Development and assessment of a natural language processing model to identify residential instability in electronic health records' unstructured data: a comparison of 3 integrated healthcare delivery systems. *JAMIA Open*. 2022;**5**(1):ooac006. doi: 10.1093/jamiaopen/ooac006.

12. **Chapman AB, Jones A, Kelley AT, et al.** ReHouSED: a novel measurement of veteran housing stability using natural language processing. *J Biomed Inform*. 2021;**122**:103903. doi: 10.1016/j.jbi.2021.103903.

13. **Morrato EH, Lennox LA, Dearing JW, et al.** The evolve to next-gen ACT network: an evolving open-access, real-world data resource primed for real-world evidence research across the clinical and translational science award consortium. *J Clin Transl Sci*. 2023;**7**(1):e224. doi: 10.1017/cts.2023.617.

14. **Rollings KA, Kunnath N, Ryus CR, Janke AT, Ibrahim AM.** Association of coded housing instability and hospitalization in the US. *JAMA Netw Open*. 2022;**5**(11):e2241951. doi: 10.1001/jamanetworkopen.2022.41951.

15. **Richards J, Kuhn R.** Unsheltered homelessness and health: a literature review. *AJPM Focus*. 2022;**2**(1):100043. doi: 10.1016/j.focus.2022.100043.

16. **Zhang Z, Huang M, Zong N, et al.** Knowledge engineering of lexicon resources of food and housing insecurity from multi-source evidence. *AMIA Jt Inform Summit Proc*. 2024;**2024**.

17. **Wen A, Fu S, Moon S**, *et al.* Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *Npj Digit Med*. 2019;**2**(1):1–7. doi: 10.1038/s41746-019-0208-8.

18. **Liu H, Bielinski SJ, Sohn S**, *et al.* An information extraction framework for cohort identification using electronic health records. *AMIA Summits Transl Sci Proc*. 2013;**2013**:149.

19. **Harkema H, Dowling JN, Thornblade T, Chapman WW.** ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*. 2009;**42**(5):839–851. doi: 10.1016/j.jbi.2009.05.002.

20. **Doran KM, Rahai N, McCormack RP**, *et al.* Substance use and homelessness among emergency department patients. *Drug Alcohol Depend*. 2018;**188**:328–333. doi: 10.1016/j.drugalcdep.2018.04.021.