This is a "preproof" accepted article for *Journal of Clinical and Translational Science*.
This version may be subject to change during the production process.
10.1017/cts.2023.626

# Evidence of Housing Instability Identified by Addresses, Clinical Notes, and Diagnostic Codes in a Real-world Population with Substance Use Disorders

Daniel R. Harris[1, 2], Nicholas Anthony[1], Dana Quesinberry[2, 3], Chris Delcher[1]

[1]Institute for Pharmaceutical Outcomes & Policy, Department of Pharmacy Practice and Science, College of Pharmacy, University of Kentucky, Lexington KY 40508, USA.
[2]Kentucky Injury Prevention and Research Center, University of Kentucky, Lexington KY 40536, USA.
[3]Department of Health Management and Policy, College of Public Health, University of Kentucky, Lexington KY 40536, USA.

**Keywords (MeSH)**: Social Determinants of Health, Housing Instability, Natural Language Processing, Geocoding

**Corresponding author:** Daniel R. Harris, PhD (daniel.harris@uky.edu), The Lee Todd Jr Building (TODD), Room 353, 789 South Limestone Street, Lexington, KY 40508, 859-323-7100

The authors have no conflicts of interest to declare.

## ABSTRACT

**Introduction:** Housing instability is a social determinant of health associated with multiple negative health outcomes including substance use disorders (SUDs). Real-world evidence of housing instability is needed to improve translational research on populations with SUDs.

**Methods:** We identified evidence of housing instability by leveraging structured diagnosis codes and unstructured clinical data from electronic health records of 20,556 patients from 2017 to 2021. We applied natural language processing with named-entity recognition and pattern matching to unstructured clinical notes with free-text documentation. Additionally, we analyzed semi-structured addresses containing explicit or implicit housing-related labels. We assessed agreement on identification methods by having three experts review 300 records.

**Results**: Diagnostic codes only identified 58.5% of the population identifiable as having housing instability whereas 41.5% are identifiable from addresses only (7.1%), clinical notes only (30.4%), or both (4.0%). Reviewers unanimously agreed on 79.7% of cases reviewed; a Fleiss' Kappa score of 0.35 suggested fair agreement yet emphasized the difficulty of analyzing patients having ambiguous housing situations. Among those with poisoning episodes related to stimulants or opioids, diagnosis codes were only able to identify 63.9% of those with housing instability.

**Conclusions**: All three data sources yield valid evidence of housing instability; each has their own inherent practical use and limitations. Translational researchers requiring comprehensive real-world evidence of housing instability should optimize and implement use of structured and unstructured data. Understanding the role of housing instability and temporary housing facilities is salient in populations with SUDs.

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

## INTRODUCTION

Social determinants of health (SDOH), such as living environment and housing stability, heavily influence an individual's general well-being and unstable housing itself can have severe negative consequences[1]. Housing deprivation, or homelessness, is the most extreme form of housing instability and can lower life expectancy by 12 years[2]. Approximately 580,000 people experienced homelessness on a single night in 2020 in the United States[3]. The Kentucky Housing Corporation manually counted 3,984 people in 2022 who were unsheltered, living in emergency shelters, or living in some type of transitional housing in Kentucky[4]. Homelessness is associated with significantly higher hospital readmission rates[5], longer hospitalizations[6], disproportionately higher use of emergency medical services and ambulance transports[7], and higher rates of illness and disability[8]. Adults experiencing homelessness have far higher rates of substance use and mental health disorders[9,10]. An in-depth analysis of California's unhoused population, the largest in the United States, identified that 65% had ever used amphetamines, 56% used amphetamines regularly, and 33% regularly used cocaine[11].

Accurate identification of housing instability in healthcare data is essential for research, as coordination of treatment programs for mental health, substance use disorders (SUDs), and social services for housing results in better health outcomes[12]. In examining populations with stimulants and opioid use disorders, we previously observed significant variation in rates of homelessness using structured data across populations with opioid use disorders (lowest), stimulant use disorders, and concurrent stimulant/opioid use disorders (highest)[13]. Those with SUDs and concurrent housing issues are at a higher risk for overdose[14].

Substantial variation exists in how SDOHs, such as housing instability, are documented within electronic health records[15]. There is no consensus definition of homelessness, no best practices for documenting homelessness, and the low usage of housing-related codes within the International Classification of Disease, Clinical Modification (ICD-10-CM) vocabulary for medical diagnoses (diagnostic billing codes) indicate a general need to improve documentation of housing instability[16]. In a previous study on the impact of SDOH on overdose[13], we found that diagnostic billing codes for housing instability were underutilized within our electronic health record after observing cases of patients experiencing homelessness using residential

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

address data or clinical notes that were not diagnostically coded; this limitation required us to further develop comprehensive methods of obtaining real-world evidence of housing instability for our population with SUDs. Other studies have shown that address data can be used to determine homelessness[17,18]. A systematic review found homelessness was the third most frequent SDOH category actively researched within clinical text (behind smoking status and substance use status)[19]. Clinical text has shown promise in identifying housing issues when paired with natural language processing and data mining techniques, where both lexical approaches requiring a lexicon of housing-related terms and machine learning methods requiring labeled training performed well[20–22]. A rule-based model was useful in identifying housing issues when using unstructured data from multiple hospital systems[23]. We later describe how we use both lexical methods (for identifying housing instability and locations) and model-based methods for recognizing housing-related concepts. This work measures the concordance of housing instability evidence identified by structured, semi-structured, and unstructured clinical data in real-world patient populations having stimulant use disorders, opioid use disorders, and concurrent stimulant/opioid use disorders.

## MATERIALS AND METHODS

### Patient Population

We extracted records for adult patients who had an encounter with diagnosis codes related to stimulant and opioid use disorders, including poisoning episodes, from the University of Kentucky HealthCare (UKHC) network, which serves central Kentucky with two hospitals, two emergency departments, multiple outpatient clinics, and regional satellite clinics. UKHC is primarily located in Fayette County, Kentucky; the Kentucky Housing Corporation manually counted 715 individuals experiencing housing issues in Fayette County in 2022, which accounts for 17.95% of the entire state's population facing housing issues[4]. Another 315 (7.9% of the state) individuals were counted in UKHC's secondary service area, which collectively accounts for 25.85% of the state's homeless popuation[4]. Our study included 20,556 patients from January 1, 2017 to May 31, 2021. The stimulant-related group was identified with ICD-10-CM diagnosis codes for cocaine use disorders (F14.*), other stimulant use disorders (F15.*), poisoning by cocaine (T40.5*), and poisoning by psychostimulants (T43.6*). The opioid-related group was

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

identified as those with codes for opioid-related disorders (F11.*) and opioid-related poisoning codes (T40.0*, T40.1*, T40.2*, T40.3*, T40.4*, T40.6*). This study was approved by the University of Kentucky IRB (#74501).

Structured and Semi-structured Data

Our EHR contains structured ICD-10-CM codes associated with every patient encounter for billing purposes. For problems related to housing and/or low-income economic circumstances, we use ICD-10-CM code Z59*. Semi-structured address data is captured by the EHR as a collection of free-text fields, which includes two address lines, city, state, and zip code. Manual review of addresses revealed patients with addresses corresponding to local supportive housing shelters, including homeless shelters and residential SUD treatment facilities. To categorize patients as having housing issues, we curated a list of addresses for housing-related resources available to our community[24,25].

Unstructured Data

We extracted 18,847,299 notes from patient visits (mode=1, median=208, average=945 notes per patient). We deployed three strategies for identifying housing issues using the notes: mentions by keyword, mentions by shelter name, and concepts extracted using named-entity recognition with a biomedical model. Keywords and phrases were constructed using common knowledge of housing-related words (e.g., “homeless”, “unhoused”, “unstable housing”). We reused the same list of community resources in the address analysis to explicitly look for mentions of local shelters by name; for example, we observed phrases such as “discharged to Shelter-X”, “lives at Shelter-X”, “transported from Shelter-X”. Our third strategy was to extract concepts from the text using named-entity recognition (NER) methods available in scispaCy[26]; we deployed scispaCy’s large scientific model (“en_core_sci_lg”), which was previously trained to recognize biomedical text; the NER pipeline is responsible for tokenizing, tagging, parsing and ultimately generating important pieces of text as named entities. This strategy avoids the need to curate a lexicon of terms related to housing that are needed by keyword matching techniques. We developed a custom knowledgebase linker to the 2022 US SNOMED-CT vocabulary that processes entities recognized by scispaCy and yields structured, coded terms in SNOMED-CT[27]; we provide our contributions as open-source software[28].

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

To evaluate our identifications when using the unstructured clinical notes, three adjudicators manually reviewed a sample of notes pulled by keyword matching, shelter-by-name matching, and name-entity recognition. 300 notes were randomly sampled, where 100 had positive keyword matches, 100 had positive NER matches, and 100 had positive matches for shelters by name. We analyzed agreement of the three adjudicators by calculating Fleiss' Kappa.

## RESULTS

Table 1 gives descriptive characteristics of our populations with SUDs, which we subdivide into patients with distinct or combined stimulant- and opioid-related codes; sex, race, and age were statistically significant using chi-squared tests (p-value <0.001) per SUD type (stimulant, opioid, or both). The entry for the SUD type with the highest percentage of representation per demographic is emphasized in bold in Table 1. We also give demographics for the state of Kentucky in 2021 from the U.S. Centers for Disease Control and Prevention for comparison to the study demographics[29]. Notable shifts in demographics include having more males in the stimulant group compared to the opioid group (59.7% versus 47.2%) and far more Black patients in the stimulant group compared to the opioid group (19% versus 5%). These shifts are also important in understanding risk and protective factors surrounding both overdose and housing issues. We analyzed adjusted standardized residuals to examine differences between observed and expected numbers. For the stimulant cohort, there were a larger number of male, Black, or ages 18-24 than expected, while a smaller number of female, white, or ages 65+ than expected. For the opioid cohort, there were a larger number of female, white, or ages 65+ than expected, while a smaller number of male, Black, or ages 18-24 than expected. More discussion on how social determinants impact overdose can be found in our prior work,[13] which motivated this study and our development of methods for identifying comprehensive evidence of housing issues.

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

**Table 1: Demographics for Populations with Substance Use Disorders in the UK Health Care system, 2017 to 2021**

| | | Overall | Stimulants | Opioids | Co-DX | P-Value | KY Reference (2021) |
|---|---|---|---|---|---|---|---|
| **Number of Unique Patients** | | 20,556 | 6,165 (30%) | **9,667 (47%)** | 4,724 (23%) | | |
| **Sex** | Male | 10,754 (52.3%) | **3,683 (59.7%)** | 4,568 (47.2%) | 2,503 (52.9%) | <0.001 | 49.3% |
| | Female | 9,801 (47.7%) | 2,482 (40.3%) | **5,099 (52.7%)** | 2,220 (46.9%) | | 50.7% |
| | Other/Unknown | 1 (< 1%) | 0 (0%) | 0 (0%) | **1 (< 1%)** | | 0 % |
| **Race** | White | 18,132 (88.2%) | 4,860 (78.8%) | 8,905 (92.1%) | **4,367 (92.4%)** | <0.001 | 84.9% |
| | Black | 1,953 (9.5%) | **1,169 (19%)** | 485 (5%) | 299 (6.3%) | | 8.9% |
| | Other | 51 (< 1%) | 20 (< 1%) | **27 (< 1%)** | 4 (< 1%) | | 6.1% |
| | Unknown | 420 (2%) | 116 (1.8%) | **250 (2.5%)** | 54 (1.1%) | | 0% |
| **Age Group** | 18-24 | 545 (2.6%) | **284 (4.6%)** | 152 (1.6%) | 109 (2.3%) | <0.001 | 9.2% |
| | 25-34 | 4,533 (22.1%) | 1,324 (21.5%) | 1,839 (19%) | **1370 (29%)** | | 9.2% |
| | 35-44 | 5,806 | 1,566 | 2,549 | **1691** | | 12.3% |

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | (28.2%) | (25.4%) | (26.4%) | **(35.8%)** | | |
| | 45-54 | 4,213 (20.5%) | **1,498 (24.3%)** | 1,727 (17.9%) | 988 (20.9%) | | 12.4% |
| | 55-64 | 3,249 (15.8%) | **1,083 (17.6%)** | 1,694 (17.5%) | 472 (10%) | | 13.3% |
| | 65+ | 2,210 (10.8%) | 410 (6.7%) | **1,706 (17.6%)** | 94 (1.9%) | | 17.2% |
| **Stimulant or Opioid Poisoning** | False | 16,129 (78.5%) | **5,803 (94.1%)** | 6,587 (68.1%) | 3,739 (79.1%) | <0.001 | |
| | True | 4,427 (21.5%) | 362 (5.9%) | **3,080 (31.9%)** | 985 (20.9%) | | |

Table 2 summarizes the results using multiple methods for identifying housing issues. 14,545 patients (70.8%) had no evidence of housing issues. Using any data source, 29.2% (n=6,011) of our population had evidence of housing instability. 54% with both stimulant-related and opioid-related codes had evidence of unstable housing, compared to 30.7% for our stimulant-only group and 16.2% for our opioid-only group. There is a significant relationship between data source (diagnosis codes, address, and notes) and SUD diagnosis type (Fisher's exact test p-values < 0.001). This relationship implies that ignoring a data source would disregard important information about patients with particular SUD diagnosis types.

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

**Table 2: Housing Instability by Data Source for Populations with Substance Use Disorders**

| | | Overall | Stimulants | Opioids | Co-DX | P-Value |
|---|---|---|---|---|---|---|
| Total Population | | 6,011 | 1,891 | 1,569 | **2,551** | |
| **Housing Diagnosis Code** | False | 2,496 (41.5%) | 808 (42.7%) | 782 (49.8%) | **906 (35.5%)** | <0.001 |
| | True | 3,515 (58.5%) | 1,083 (57.3%) | 787 (50.2%) | **1,645 (64.5%)** | |
| **Housing Address** | False | 5,030 (83.7%) | 1,544 (81.6%) | 1,288 (82.1%) | **2,198 (86.2%)** | <0.001 |
| | True | 981 (16.3%) | **347 (18.4%)** | 281 (17.9%) | 353 (13.8%) | |
| **Housing in Notes** | False | 1,798 (29.9%) | 552 (29.2%) | **682 (43.5%)** | 564 (22.1%) | <0.001 |
| | True | 4,213 (70.1%) | 1,339 (70.8%) | 887 (56.5%) | **1,987 (77.9%)** | |

Only 3,515 patients (17.1%) in our population had billing codes indicating housing instability; this only represents 58.5% (n=3,515) of patients with housing instability identified from any of our data sources. There were 65 patients who had 'homeless' as the first address line, which was 6.6% of our address-based results. 8.2% of our population had a generic ‘Lexington, KY’ or equivalent address that did not specify an address which does not necessarily imply a housing issue**.** 51.2% of patients with these generic addresses had housing issues identified by other methods. We detected 16.3% of those with unstable housing as having addresses directly corresponding to a community resource from our curated list.

6,011 patients in our study population (29.2%) had housing issues when merging signals from all three data sources. Figure 1 demonstrates intersecting results of each method. 59.9% of patients with housing issues have documentation originating from a single source; only 286 (3.4%) had evidence in all three sources and less than a third had evidence from more than one source. This

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

suggests that all three data sources are needed to understand housing issues; 41.5% of our population with housing issues are identifiable by analyzing addresses only (7.1%), clinical notes only (30.4%), or either one (4%). Figure 2 visualizes how these differences are distributed within our community and shows the magnitude of underrepresentation for those with housing issues when only considering diagnosis codes. Similar results were observed within the stimulant-related group (17.5% to 30.7%), the opioid group (8.1% to 16.2%), and the group with both (34.8% to 54%). As demonstrated in Figure 2, there are implications for linking patient records to geographic units such as census tracts. If only diagnosis codes were used, 111 census tracts (7.8%) would be missed and considered absent of individuals with housing instability. 478 census tracts (33.6%) saw increases in the number of individuals. Within these tracts, increases of up to 171 additional individuals were counted as having housing instability using results from address and note data; on average, 8.7 additional patients were added to each census tract.

Our validation review sample corresponded to 300 distinct individuals (4.9% of our population with housing instability). The reviewers unanimously agreed with the extractions in 239 out of 300 cases (79.7%) but the consensus varied by extraction method. NER and keyword methods had the highest total agreement (93% and 86%, respectively), but shelters were only 60% concordant; a Fleiss' Kappa score of 0.35 suggested only fair agreement, which largely stemmed from ambiguity around what role a shelter or SUD treatment facility was serving in a patient's life.

Table 3 describes how different methods of utilizing the clinical notes contributed to our housing instability totals from Table 2; keyword matching was the largest contributor. Approximately half of this population had addresses corresponding to known housing resources. Method was statistically significant ($p < 0.001$) for SUD type using Fisher's exact tests.

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

**Table 3: Evidence of Housing Instability in Notes by Method**

| | | Overall | Stimulants | Opioids | Co-DX | P-Value |
|---|---|---|---|---|---|---|
| Total Population | | 4,213 | 1,339 | 887 | 1,987 | |
| **Shelters by Name** | False | 2,084 (49.5%) | **742 (55.4%)** | 439 (49.5%) | 903 (45.4%) | <0.001 |
| | True | 2,129 (50.5%) | 597 (44.6%) | 448 (50.5%) | **1,084 (54.6%)** | |
| **Keywords** | False | 610 (14.5%) | 126 (9.4%) | **219 (24.7%)** | 265 (13.3%) | <0.001 |
| | True | 3,603 (85.5%) | **1,213 (90.6%)** | 668 (75.3%) | 1,722 (86.7%) | |
| **Named-Entity Recognition** | False | 3,996 (94.8%) | 1,281 (95.7%) | **867 (97.7%)** | 1,848 (93.0%) | <0.001 |
| | True | 217 (5.2%) | 58 (4.3%) | 20 (2.3%) | **139 (7.0%)** | |

Table 4 describes demographics for those identified as having housing issues by data source. Sex and age were statistically significant ($p < 0.001$) for all data sources (diagnosis codes, addresses, and notes). Race was only significant ($p < 0.001$) for those with diagnosis codes and notes. Poisoning events related to stimulants or opioids (ICD10 T40.* and T43.6) were similarly identified in 17.1% to 20% of the population across data sources, however poisonings were only statistically significant for those with housing issues identified from diagnosis codes and notes ($p < 0.001$) using Fisher's exact tests. 1,099 of our population's 4,427 poisoning episodes (24.8%) were among those with evidence of housing instability; this represents 18.2% of our population with housing instability having had prior poisoning episodes. Only 703 of these 1,099 (63.9%) had a housing-related diagnosis code, which further highlights the importance of address and note data sources.

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

**Table 4: Housing Instability Demographics by Data Source**

| | | Overall | Diagnosis Code | | Address | | Notes | |
|---|---|---|---|---|---|---|---|---|
| **Number of Unique Patients** | | 6,011 | 3,515 | | 981 | | 4,213 | |
| **Sex** | Male | 3,297 (55.3%) | 2,041 (58.8%) | p < 0.001 | **626 (64.3%)** | p < 0.001 | 2,258 (53.6%) | p < 0.001 |
| | Female | 2,666 (44.7%) | 1,429 (41.2%) | | 348 (35.7%) | | **1,955 (46.4%)** | |
| | Other or Unknown | 1 (0.0%) | **1 (0.0%)** | | 0 (0%) | | 0 (0%) | |
| **Race** | White | 5,207 (86.6%) | 3,039 (86.5%) | p < 0.001 | 828 (84.4%) | p > 0.05 | **3,661 (86.9%)** | p < 0.001 |
| | Black | 664 (11.0%) | 378 (10.8%) | | **129 (13.1%)** | | 495 (11.7%) | |
| | Other | 10 (0.2%) | 4 (0.1%) | | **3 (0.3%)** | | 7 (0.2%) | |
| | Unknown | 130 (2.2%) | **94 (2.7%)** | | 21 (2.1%) | | 50 (1.2%) | |
| **Age Group** | 18-24 | 554 (9.3%) | 245 (7.1%) | p < 0.001 | **105 (10.8%)** | p < 0.001 | 429 (10.2%) | p < 0.001 |
| | 25-34 | 1,897 (31.8%) | 967 (27.9%) | | 323 (33.2%) | | **1,473 (35.0%)** | |
| | 35-44 | 1,651 | **998** | | 265 | | 1,172 | |

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | (27.7%) | **(28.8%)** | | (27.2%) | | (27.8%) | |
| | 45-54 | 1,128 (18.9%) | **751 (21.6%)** | | 190 (19.5%) | | 735 (17.4%) | |
| | 55-64 | 592 (9.9%) | **412 (11.9%)** | | 68 (7.0%) | | 345 (8.2%) | |
| | 65+ | 142 (2.4%) | **98 (2.8%)** | | 23 (2.4%) | | 59 (1.4%) | |
| **Stimulant or Opioid Poisoning** | False | 4,912 (81.7%) | 2,812 (80.0%) | p < 0.001 | 802 (81.8%) | p > 0.05 | **3,491 (82.9%)** | p < 0.001 |
| | True | 1099 (18.3%) | **703 (20.0%)** | | 179 (18.2%) | | 722 (17.1%) | |

## DISCUSSION

The nuances of housing stability are dynamic and complex by nature in healthcare data; our overarching goal was to demonstrate how multiple real-world data sources contribute to the identification of housing instability. Adding unstructured data nearly doubled the number of patients identified as having unstable housing.

The largest increase in housing status identification occurred from analyzing clinical notes. Both matching by keyword and by shelter names produced a substantial number of patients that were not detected otherwise. Our NER method underperformed, and all matches were available through other means; upon review, we learned that not all extracted named entities mapped to concepts in SNOMED CT due to the limitations of using a dictionary approach to concept matching. If the concept did not exist in SNOMED CT, the named entity would not produce a corresponding match, despite being detected. This creates bias in our NER results in that successful matches are limited to our target vocabulary, SNOMED CT, which may lack a

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

comprehensive vocabulary for housing issues. We leave improving this model for future work, but we acknowledge that NER matches demonstrated higher accuracy and agreement between our manual adjudicators, which suggests that this method may give results with high precision at the cost of lowering recall.

We demonstrate that real-world local context is important in identifying those with disrupted living environments and that shelters providing temporary housing can be explicitly identified by name. We interpreted clinical notes that mention a shelter by name as an implication of the patient having a housing need. Many of these are in the form of 'discharged to Shelter-X' as mentioned above, but there are indirect mentions that we also assumed implied a housing need. For example, 'Social worker printed out information about two facilities, Shelter-X and Shelter-Y, and gave to patient'; this sentence is ambiguous on whether the patient stayed at either shelter, but it does imply that a provider perceived a housing need. Differences in demographics and risk of housing instability among SUD cohorts may be confounded by other factors. For example, more men than women have evidence of homelessness from address data (64.3% vs 35.7%); the stimulant cohort has a higher percentage of men than others (59.7%) and has a higher number of individuals with address data implying housing issues. Men stay homeless longer than women on average[30] which increases their chance to stay at a shelter during our study period and more likely to use a shelter address for correspondence; additionally, our community has specific male-only shelters with recovery programs. We wish to further explore discrepancies in clinical documentation of housing instability across SUD cohorts as future work.

The most difficult cases when identifying by shelter names emerged when shelters served multiple roles, such as providing temporary housing and residential substance use treatment, which does not require homelessness as a condition for admittance; manual review did reveal some instances where shelters were used strictly for substance rehabilitation by the context given in the note. For this reason, searching for shelter names may overestimate housing instability unless residential treatment is included in the semantics of having a disrupted living environment. Address methods require curation of locations for community housing resources; for our study, this was a manual process using known data sources for our service area. Obtaining this information may be difficult for larger jurisdictions; however, if this information

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

exists online, it could potentially be requested or scraped as part of a larger automation effort to improve the ease of implementation.

There are several vignettes from our study that are instructive. First, negations did impact the success of keyword matching and led to incorrect housing status assignment. For example, "asked if she was homeless and she denied," but this phrase did not occur otherwise outside of the sample. Second, we found true false positives, such as 'found a homeless person sleeping in her bathroom', but these examples were uncommon and represented less than one percent of our sample and were so narrow in phrasing that they did not manifest otherwise. Third, our manual review found that language around housing is difficult and non-standardized, which is far more problematic than false positives. For example, our team considers 'has stable housing including homeless shelter' as paradoxical, as homeless shelters imply housing instability, and this provides data to advocate for better and more consistent clinical documentation. Fourth, we observed clinical documentation of "elective" homelessness, such as 'patient was living on the streets but does have a home and multiple dogs'; this situation is relatively rare but highlights the complexity of housing circumstances. Fifth, the manual review demonstrated the difficulty of considering prior history of housing issues; one example documented an episode of homelessness that occurred several years ago.

All sources of evidence for housing issues have limitations. Diagnosis codes underrepresent the homeless population[13,15], which is confirmed in our study as 41.5% of individuals identified as homeless through other means lacked a diagnosis code; we did not validate the diagnostic accuracy of the ICD-10-CM diagnostic codes. Our address data was limited to the most recent address. Furthermore, our list of homeless resources was taken as a snapshot in time and may not reflect resources available during the entire four-year study period. For these reasons, we were unable to examine the temporal relationship between the EHR address and housing issues. Our clinical notes are inherently limited to only what was documented within the note; we observed that 38.3% of patients having diagnosis codes for homelessness had no accompanying clinical documentation within the unstructured notes**.**

Many of the homeless shelters that serve our community's housing needs also serve other roles, such as transitional living support or substance-related rehabilitation. This limitation does not

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

impact our original goal of wanting to identify those with higher risk factors for SUDs and overdose; patients interacting with a shelter are already at a higher risk for substance-related issues regardless of *why* that interaction occurs because of the known association between homelessness, SUDs, and overdose[13]. Because of this association, we need comprehensive, real-world evidence of housing issues using multiple data sources. Our address and NER methods could be adopted by clinical data warehouse (CDW) teams to improve the identification of those with housing issues. In fact, our team is responsible for geocoding UKHC records on behalf of its CDW team and UK's Center for Clinical and Translational Sciences; this data, in turn, is made available to others for enterprise reporting and research. We see an opportunity for a quality improvement project; our methods depend upon reliable data, either accurate addresses or clinical documentation. Table 3 suggested that method of documentation is inconsistent across different SUD cohorts; Table 4 suggested demographic characteristics of patients are related to how housing is documented. For example, it is not immediately clear why males have a higher proportion of housing documented as address data, but it suggests that care is needed when collecting address information during clinical administration to avoid bias. The importance of consistent documentation is further demonstrated by 18.3% of our population with housing issues having experienced a poisoning related to stimulants or opioids; clinical documentation potentially leads to better coordination of follow-up care and appropriate social services.

## CONCLUSION

The number of patients identified as having housing issues nearly doubled when including data sources for structured and semi-structured data; therefore, it is abundantly clear that translational use cases needing real-world evidence must consider diverse data sources. We advocate that real-world local context is paramount when processing unstructured data due to either the large occurrence of homeless shelters mentioned by name in clinical notes or the large number of patients with residential addresses corresponding to a shelter. Our study underscores the importance of analyzing multiple facets of text data from multiple data sources to get a comprehensive understanding of a patient's social determinants of health.

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

## ACKNOWLEDGEMENTS

This project is fully supported by the Centers for Disease Control and Prevention of the U.S. Department of Health and Human Services (HHS) as part of grant 1R01CE003360-01-00. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by CDC/HHS, or the U.S. Government.

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

3. 2020 AHAR: Part 1 - PIT Estimates of Homelessness in the U.S. | HUD USER. Accessed November 22, 2022. https://www.huduser.gov/portal/datasets/ahar/2020-ahar-part-1-pit-estimates-of-homelessness-in-the-us.html

4. Kentucky Housing Corporation. Accessed July 5, 2023. https://www.kyhousing.org:443/Data-Library/Pages/K-Count-Results.aspx

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. Salit SA, Kuhn EM, Hartz AJ, Vu JM, Mosso AL. Hospitalization Costs Associated with Homelessness in New York City. *New England Journal of Medicine*. 1998;338(24):1734-1740. doi:10.1056/NEJM199806113382406

7. Abramson TM, Sanko S, Eckstein M. Emergency Medical Services Utilization by Homeless Patients. *Prehospital Emergency Care*. 2021;25(3):333-340. doi:10.1080/10903127.2020.1777234

8. 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

9. Fischer PJ, Breakey WR. The epidemiology of alcohol, drug, and mental disorders among homeless persons. *American Psychologist*. 1991;46(11):1115-1128. doi:10.1037/0003-066X.46.11.1115

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

11. California Statewide Study of People Experiencing Homelessness | Benioff Homelessness and Housing Initiative. Published July 5, 2023. Accessed July 6, 2023. https://homelessness.ucsf.edu/our-impact/our-studies/california-statewide-study-people-experiencing-homelessness

12. Fitzpatrick-Lewis D, Ganann R, Krishnaratne S, Ciliska D, Kouyoumdjian F, Hwang SW. Effectiveness of interventions to improve the health and housing status of homeless people: a rapid systematic review. *BMC Public Health*. 2011;11(1):638. doi:10.1186/1471-2458-11-638

13. Delcher C, Harris DR, Anthony N, Stoops WW, Thompson K, Quesinberry D. Substance use disorders and social determinants of health from electronic medical records obtained during Kentucky's "triple wave." *Pharmacology Biochemistry and Behavior*. Published online November 22, 2022:173495. doi:10.1016/j.pbb.2022.173495

14. Palis H, Xavier C, Dobrer S, et al. Concurrent use of opioids and stimulants and risk of fatal overdose: A cohort study. *BMC Public Health*. 2022;22(1):2084. doi:10.1186/s12889-022-14506-w

15. Wang M, Pantell MS, Gottlieb LM, Adler-Milstein J. Documentation and review of social determinants of health data in the EHR: measures and associated insights. *Journal of the American Medical Informatics Association*. 2021;28(12):2608-2616. doi:10.1093/jamia/ocab194

16. Bensken WP. How do we define homelessness in large health care data? Identifying variation in composition and comorbidities. *Health Serv Outcomes Res Method*. 2021;21(1):145-166. doi:10.1007/s10742-020-00225-5

17. Vickery KD, Shippee ND, Bodurtha P, et al. Identifying Homeless Medicaid Enrollees Using Enrollment Addresses. *Health Services Research*. 2018;53(3):1992-2004. doi:10.1111/1475-6773.12738

18. Zech J, Husk G, Moore T, Kuperman GJ, Shapiro JS. Identifying homelessness using health information exchange data. *Journal of the American Medical Informatics Association*. 2015;22(3):682-687. doi:10.1093/jamia/ocu005

19. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association*. 2021;28(12):2716-2727. doi:10.1093/jamia/ocab170

20. 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. *Journal of the American Medical Informatics Association*. 2018;25(1):61-71. doi:10.1093/jamia/ocx059

21. 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

22. 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.

23. 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

24. District 9 - Department of Corrections. Accessed November 28, 2022. https://corrections.ky.gov/Reentry/resources/Pages/District9.aspx

25. Community Resource Guide. Accessed November 28, 2022. https://www.kyhousing.org/Programs/Documents/Community%20Resource%20Guide.pdf

26. scispaCy · spaCy Universe. scispaCy. Accessed November 30, 2022. https://spacy.io/universe/project/scispacy

27. SNOMED Home page. SNOMED. Accessed November 30, 2022. https://www.snomed.org/

28. GitHub - UK-IPOP/STIMuLINK: Meta repository for work related to STIMuLINK grant. Accessed July 7, 2023. https://github.com/UK-IPOP/STIMuLINK

29. Single-Race Population Estimates 2020-2021 by State and Single-Year Age Request. Accessed July 5, 2023. https://wonder.cdc.gov/Single-Race-single-year-v2021.HTML

30. North CS, Smith EM. A comparison of homeless men and women: Different populations, different needs. *Community Ment Health J*. 1993;29(5):423-431. doi:10.1007/BF00754410

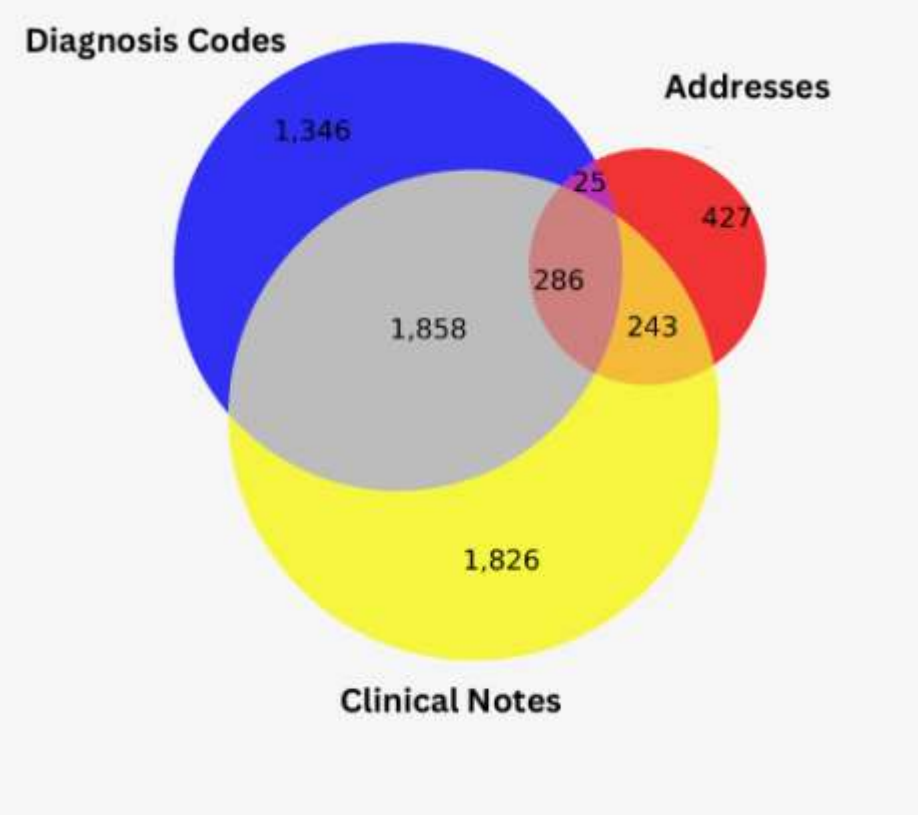


**Figure 1:** Unique patients (6,011 total) identified having housing issues by intersection of data source: diagnosis codes (3,515 total or 58.5%), addresses (981 total or 16.3%), and clinical notes (4,213 total or 70.1%).

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press

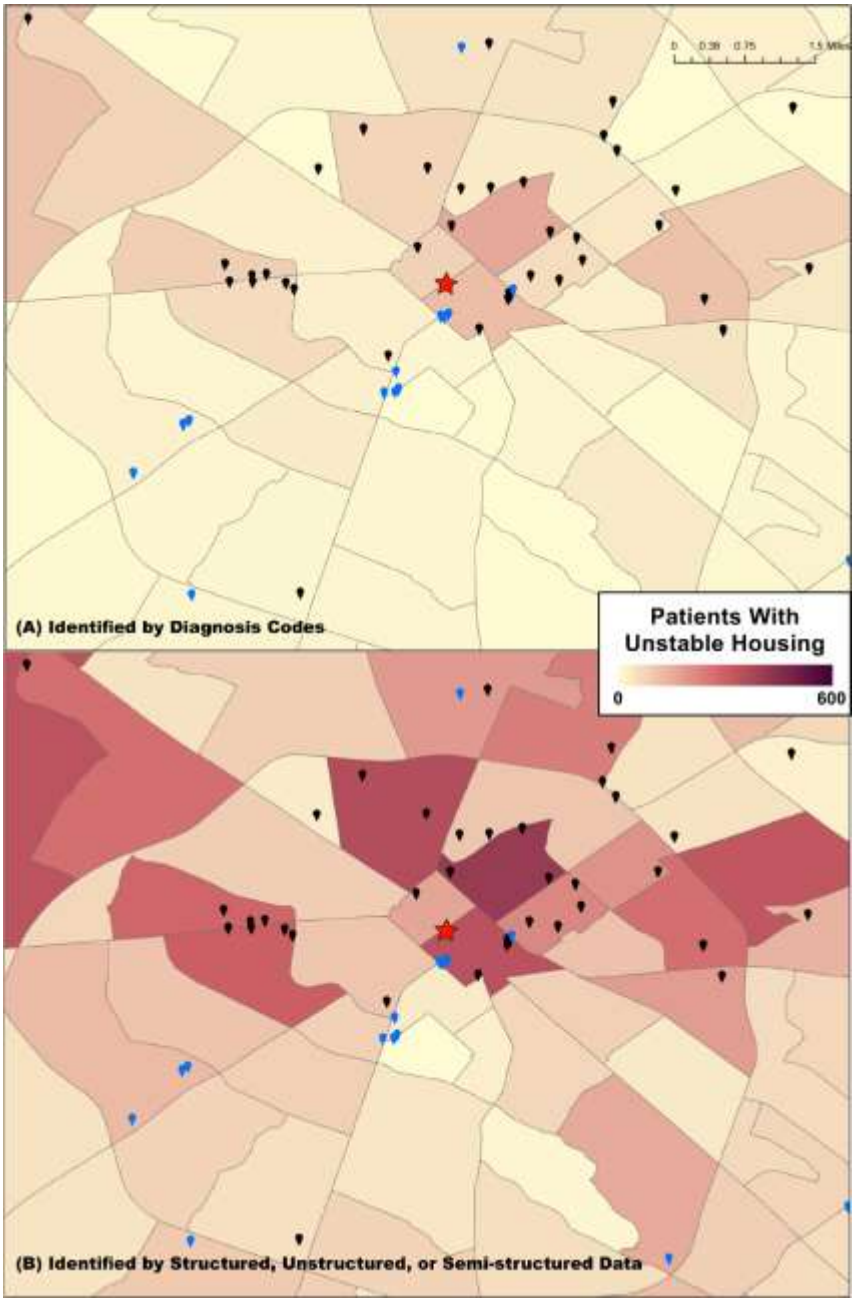


**Figure 2:** Patients with unstable housing in Fayette County, Kentucky when (A) using only diagnosis codes or (B) when using diagnosis codes, clinical notes, or address data. Black pins are locations of housing-related community resources; blue pins are locations of hospitals, clinics, and emergency departments in our healthcare network; administrative boundaries are census tracts. The red star is the city center of downtown Lexington.

https://doi.org/10.1017/cts.2023.626 Published online by Cambridge University Press