BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email
info.bmjopen@bmj.com

BMJ Open

A Simplified Framework to Extract Social Determinants: A Data Science Approach

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-048397
Article Type:	Original research
Date Submitted by the Author:	19-Jan-2021
Complete List of Authors:	Teng, Andrew; University of Washington, Biomedical Informatics and Medical Education Wilcox, Adam ; University of Washington,
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, SOCIAL MEDICINE, HISTORY (see Medical History), BIOTECHNOLOGY & BIOINFORMATICS

A Simplified Framework to Extract Social Determinants: A Data Science Approach

Andrew K. Teng[1], Adam B. Wilcox[1]

Affiliation:
[1] Biomedical Informatics and Medical Education
University of Washington
Seattle, WA, United States

Corresponding Author:
Andrew K. Teng
akteng@uw.edu
Biomedical Informatics and Medical Education, University of Washington
850 Republican Street, Box 358047
Seattle, WA 98195-0005, United States

Keywords: Text extraction, text classification, housing stability, social determinants, data science

Word Count: 4,228

Abstract

Objectives
We aim to extract a subset of social factors from clinical notes using common text classification methods.
Setting
We collaborated with a local Level I trauma hospital located in an underserved area that has a housing unstable patient population of about 6.5% and extracted text notes related to various social determinants for acute care patients.
Participants
Notes were retrospectively extracted from 43,798 acute care patients.
Methods
We solely utilize open source Python packages to test simple text classification methods that can potentially be easily generalizable and implemented. We extracted social history text from various sources, such as admission and emergency department notes, over a five-year timeframe and performed manual chart reviews to ensure data quality. We manually labelled the sentiment of the notes, treating each text entry independently. Four different models with two different feature selection methods (bag of words (BOW) and bigrams) were used to classify and predict housing stability, tobacco use, and alcohol use status for the extracted clinical text.
Results
From our analysis, we found overall positive results and metrics in applying open-source classification techniques; the accuracy scores were 91.2%, 84.7%, 82.8% for housing stability, tobacco use, and alcohol use respectively. There were many limitations in our analysis including social factors not present due to patient condition, multiple copy-forward entries and shorthand. Additionally, it was difficult to translate usage degrees for tobacco and alcohol use. However, when compared to structured data sources, our classification approach on unstructured notes yielded more results for housing and alcohol use; tobacco use proved less fruitful for unstructured notes.

Article Summary
Strengths and limitations of this study
- From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status.
- Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use.
- Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social determinants and can supplement current structured sources to provide a more complete social history for patients.
- However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

I. INTRODUCTION
The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 established guidelines to help improve patient safety and efficacy by laying the framework for electronic health record (EHR) adoption in the United States through financial incentives.[1] With the HITECH Act and incentives through Meaningful Use, EHR adoption skyrocketed and large databases of clinical information were implemented.[2] These large databases can contain simple information such as patient demographics and vital signs, but it can also contain more qualitative or descriptive data such as clinical notes and images. With Meaningful Use the completeness of the data being collected has increased. Currently, many institutions have large amounts of underutilized data that are now ripe for biomedical exploration and discovery to aid in patient care creating new opportunities to explore.

Most data can be generally categorized as structured or unstructured, where structured data can consist of items such as vital signs and lab results and unstructured data can consist of items such as text notes, images, or multimedia.[3] Structured data have been essential in modern databases as they are significantly easier to query, merge, or filter when sifting through the data. They have two parts which simplifies the search process: (1) variable name and (2) value.[4] Structured data can be easily added and expanded and has proven critical in modern clinical databases, especially for data such as patient vitals and demographics.

On the other hand, although structured data can generally be easier to extract and analyze, unstructured data can potentially provide an array of information not present or easily identifiable in structured data. Challenges arise with unstructured data as they are not as easily interpretable as or categorizable as a numeric structured value. Images and text often contain many levels of metadata that would need manual review to decode or interpret. Additionally, clinicians have recently expanded intake data and social determinants of health (SDoH) information are starting to become more readily available. Furthermore, there has a been a growing interest around Medicaid patients, as SDoH can drive up to 80% of health outcomes, especially within this patient demographic.[5] Therefore, SDoH and REAL (Race, Ethnicity and Language) data are now starting to be analyzed for secondary research as recent research has indicated that there is a correlation between SDoH and health outcomes and the increasing need to research health disparities across populations.[6]

SDoH and REAL can include housing stability, access jobs and health care services, education level, language, and socioeconomic conditions.[7] These indicators are descriptors of different societies and are useful as predictors of health outcomes and the uptake of health interventions.[8] Because they can potentially be powerful indicators of health, many institutions are now starting to analyze and intake SDoH and REAL information, whether through text notes or standardized coding, such as International Classification of Diseases (ICD).[9] Additionally, SDoH can provide health teams with a greater understanding of a patient condition holistically.[10] However, there are challenges with SDoH intake as there is no standardized SDoH screening tool in the EHR[11]; additionally, coding schemes like ICD can prove to be unreliable in secondary analysis as coding can oversimplify symptoms and diagnoses leading to coding uncertainties and the fact that coding errors may be present from unintentional mistakes or even upcoding.[12,13] Past research has shown that hospital readmissions are highly influenced by patient health status and SDoH and suggest that clinical staff and researchers should consider SDoH when assessing readmission risk.[14] Housing stability is a major public health issue.

Locally, it is estimated that there are at least 22,000 homeless individuals in [redacted for review] and more than 12,000 people in the [redacted for review] region, a four percent increase over the previous year.[15] Housing instability is associated with various health inequalities, such as shorter life expectancy, higher morbidity, and increased usage of acute hospital services, "as the social determinants of homelessness and health inequities are often intertwined, and long term homelessness further exacerbates poor health".[16] It is therefore important to treat housing stability and other SDoH as a combined health issue to aid in improving health outcomes in clinical settings. Although some research has shown that patients who experience housing instability are more likely to die following admission for severe sepsis than those with insurance,[17] other research indicates that the effects of health inequalities are still unclear and need further investigation.[18] Additionally, various social habits, including tobacco and alcohol use, although may not directly be considered a SDoH, can impact health decisions and outcomes. For example, one study found that participants who drank alcohol and reported tobacco use consumed more foods higher in fat and sugar, low in vitamins and minerals as well as foods, considered by them to be less healthy and prepared in a less healthy way.[19]

Within our region, it has been noted in recent years that the smoking rate is around 13 percent; however, among Black/African-Americans or individuals with multiple races, is double the rate among white adults and four times higher than Asian adults. Additionally, it was reported that, when compared to high income households, low income households were three times more likely to be smokers.[20,21] Drug and alcohol use also shared similar metrics; within the region, "drug and alcohol-caused deaths was 22% higher among Blacks and four times greater among American Indian/Alaskan Native than among non-Hispanic

Whites" and alcohol use represented 4.97 per 100,000 deaths locally in 2015.[22,23] Therefore it may be important to look at health habits and SDoH together to better understand the patient population.[19]

Recent technological advances in machine learning and artificial intelligence have shown great potential in providing a pathway for informaticians and clinicians to better understand unstructured data. Within the clinical setting, there have been numerous approaches in adopting natural language processing (NLP) to aid with processing unstructured clinical text notes. Common uses of NLP include extracting diagnoses and chief complaints as well as grouping of information for quality improvement. There are various NLP methods that can be used in the clinical setting, such as automatic tagging of conditions or variables of interest, sentiment classification, or even text extraction. Various open source NLP and ontological tools, such as Automated Retrieval Console, Apache clinical Text Analysis and Knowledge Extraction System (Apache cTAKES), MetaMap, and HITEx, Unified Medical Language System (UMLS) Metathesaurus and BioPortal have been used to aid with text extraction or classification.[24–26] On the other hand less complex classification methods have been used as well to identify specific groups of patients, risk assessment, or aid in validating structured annotation.[27,28,29] A recent scoping review found that although practitioners collect a variety of SDOH data at point of care through EHR, the overall use of automated technology is limited to date.[30]

With the idea of implementing an easily generalizable approach to classify selected social factors, we extracted both unstructured and structured data sources related to SDoH from a local hospital to identify and generate a framework to automatically extract and classify SDoH from text notes. We focused on housing stability status, tobacco use, and alcohol use. These three social factors were chosen due to their direct impact on health outcomes and the local public health impact[15–19] and presence in the EHR. To tackle challenges associated with SDoH extraction from unstructured text notes, we aimed to create a generalizable framework using low barrier open-source tools that are commonly used in the data science field. Because notes and stylistic choices can be institution and location specific, we sought not to create a model that is generalizable but rather a simplified method that could be potentially easily implemented using common off the shelf NLP and data science tools.

II. METHODS

Study Design and Overview

A high-level overview of our workflow can be seen in Figure 1. We conducted a retrospective cohort study of patients in the acute care setting at a Level I trauma center and academic teaching hospital with the aim to create a general and easily applicable workflow to extract and classify social factors from clinical notes. We applied a two-pronged approach and collected unstructured data from a subset of patients over a 1-year timespan (Group A) to create and test the text classification model and also collected structured and unstructured data from a subset of patients over a 5-year timespan (Group B) to apply the model and compare results between the two data types. We performed automatic classification and scoring of patients via various NLP classification methods on three social factors: (1) housing stability, (2) tobacco use, and (3) alcohol use. Our general workflow for housing stability, a similar approach was also used for tobacco and alcohol use, can be seen in Figure 2. Patient data were extracted directly from the data warehouse and stored on encrypted computers and were not distributed or shared outside of the secured and closed environment.

Study Population

Data were extracted from [redacted for review], a 413-bed academic hospital that has a patient population consisting mostly from Washington, but also from a five-state area.[31] In 2014, there were 17,121 inpatient admissions, where 19 percent of the patients belong to a racial or ethnic minority and 37 percent of patients were enrolled in Medicaid.[31,32] Additionally, in 2015, the non-US born population was estimated to be around 21 percent in [redacted for review], highlighting the potential diversity that could be found with this patient population.[32]

Data Sources, Extraction, and Validation

A We extracted both structured and unstructured data sources related to housing stability, tobacco use, and alcohol use using SQL queries called directly from an integrated python-based Jupyter Notebook:

a. Structured data sources include billing and diagnostic/International Classification of Disease (ICD) 9 and 10 codes, questionnaire or Epic SmartForm responses, address fields (location), problem list (ICD 9), patient encounters, clinical events (actual encounters of care), and discharge/disposition location.
b. Unstructured data sources consisted of text notes from the emergency department (ED), admission (admit) notes, social work, and ambulance notes.

Discharge notes were not explored as they were not recorded in the same subdivided format as the admit and ED notes, making selective text extraction of SDoH difficult. From our initial list of patient identifiers over a one-year timespan from Group A, we performed manual EHR validation of a random subset of 50 patients to validate the completeness of the clinical notes and confirm the location of social history and social factors in clinical notes. Extensive research and conversations with an internal data analyst confirmed the location of these topics (housing, tobacco use, and alcohol use) within structured data sources.

Data Cleaning

After confirmation, clinical notes were extracted for both Groups A and B. The notes were cleaned (e.g. symbols removed, converted to lowercase) prior to classification and analysis in the Jupyter notebook via NLTK. Our general text extraction and cleaning workflow can be seen in Figure 3. However, housing stability notes and tobacco or alcohol use notes were stylistically and grammatically different, and both sets needed distinct additional cleaning steps. Housing stability notes that contained the phrase 'not homeless' were converted via regex to say 'housed' instead. Additionally, for housing stability, a concept dictionary was also created to substitute local facility names with more general concept (e.g. 'Union Gospel Mission' was converted to 'shelter'). This was done to explore how the algorithms handle formal nouns. For text notes in Group B, we performed an additional concept extraction step. Tobacco use and alcohol use notes often contained incomplete (lacking the subject, predicate, object format) triples or doubles (e.g. 'Denies smoking, drinking, drugs'). Due to their incomplete sentence structures, common NLP tools to parse, extract, and classify triples, such as Stanford CoreNLP, were not suitable as these tools rely on having all three parts of the triple present. These notes related to tobacco and alcohol use therefore underwent an additional step that performed a separate relation extraction that first pulls out the SDoH related objects and then would reclassify and label the negative sentiment to all components of the list. Our process can be seen in the left side of Figure 3. If the regex extraction of negative lists resulted in a different result from the text classification prediction, the regex extraction would overwrite the end result prior to scoring. Once these steps were performed, the data were considered clean and suitable for classification.

Model building

Cleaned text from Group A were used to generate and test the classification models. These notes were split in 70/30 validation and testing sets. We applied four different common NLP text classification models to the testing sets (via SciKit Learn): multinomial naïve Bayes, support vector machine, logistic regression, and random forest. Default parameters and a bag-of-words approach were used. The best performing model by accuracy was then chosen and applied to the larger corpus, Group B, with notes from patients in Group A removed, to avoid overfitting and classification bias. This process was performed for housing, tobacco use, and alcohol use.

Scoring generation

In order to create a simple method of identifying patients who are experiencing social instability, we created a scoring metric based on the classified notes. After applying the optimum model by accuracy to the entire corpus of extracted text notes, housing stability, tobacco use, and alcohol use scores were generated. Patient identifiers were mapped by patient location and those who were not in the acute care setting during this timeframe were removed. Three different scoring approaches were used to describe these social factors: (1) predictions were averaged by patient encounter, then averaged by patient identifier, (2) predictions were averaged by year, then by patient identifier, and (3) predictions were averaged by year, where each year then had a weight where the most recent year had the highest weight and the furthest year had the lowest weight (e.g. predictions from 2019 were weighted by a factor of 5 and predictions from 2015 were weighted by a factor of 1). This scoring generation process was then repeated on our structured data for all three social factors and the results were compared and analyzed. Structured data was also extracted for our list of patients in Group B.

Patient and Public Involvement
No patient involved.

III. RESULTS

Characteristics of study subjects
Clinical notes (ED, admit, social work, and ambulance) between 2015 and 2019 were extracted and included, forming Group B. Notes from the first 200 patients were included in Group A and notes from 147,457 patients were included in Group B. During the same timeframe, 61,767 patients were in acute care. After extraction and model prediction, the patient notes were cross referenced with inpatient location and only notes from those who were in acute care were retained, for a total of 43,798 patients from 2015 to 2019. The patient demographics of this final subset were 63% (*n*=27,575) male, 37% (*n*=16,223) female, 88.2% (*n*=38,634) not Hispanic or Latino, and 10.5% (*n*=4,609) Hispanic or Latino, and 1.3% (*n*=555) unknown or not answered. Further descriptive statistics can be found in Table 1.

Table 1: Population demographics

Race (*n*=43,798)	*n* (%)
White or Caucasian	31,575 (72.1%)
Black or African American	4,812 (11.0%)
Asian	3,174 (7.2%)
American Indian or Alaska Native	1,165 (2.7%)
Native Hawaiian or other Pacific Islander	524 (1.2%)
Multiple races	3 (0%)
Unavailable, unknown, or missing	2,545 (5.8%)
Age range (*n*=43,798)	***n* (%)**
0-18	1,856 (4.2%)
19-44	12,437 (28.4%)
45-64	14,863 (33.9%)
65-84	11,902 (27.2%)
85 and over	2,740 (6.3%)

Data attributes
Table 2 illustrates the amount of data for each corresponding extraction level, specifically for housing status. We first started with extracting text from the ED and admit notes, forming Group A, which consisted of 50,000 rows or text entries and covered 3,200 unique patients, over a one-year timeframe. From there, we manually labelled housing stability concepts in a binary fashion, where 0 would indicate housing stability and 1 would indicate any level of housing instability, regardless of severity. As manual labelling can be a labor-intensive process, only the first 6,000 text rows were labelled, covering 218 unique patients. However, within these first 6,000 rows, numerous notes did not contain text that alluded to housing status or were empty due to patient condition. Therefore, only 1,785 out of the 6,000 rows were labelled, covering 200 unique patients, where 995 (55.7%) were labelled as housing stable and 790 (44.3%) were labelled as housing unstable. We also found that 5.7% of the entries within this subset were duplicates or copy-forward entries. The same workflow was performed for labelling tobacco and alcohol use. However, only 1,108 rows were labelled for tobacco use and 1,220 rows for alcohol use, where in both cases 0 indicated no use, 1 indicated rare/previous/occasional use, and 2 indicated current use, regardless of degree. Tobacco use resulted in 446 labels for no use, 129 labels for rare/previous/occasional use, and 533 labels for current use. Similarly, alcohol use resulted in 595 labels for no use, 185 labels for rare/previous/occasional use, and 440 labels for current use.

Table 2: Extracted data amounts for housing status

Level of extraction	Rows (*n*)	Unique patients (*n*)	Unique encounters (*n*)	Social history entries (*n*/unique)
ED and Admit notes	49,955	3,233	15,664	21,876/21,334
'Housing (in)stability'	6,000	218	1,995	2,408/2,211
Remove nulls/missing data	1,785	200	1,361	1,785/1,684

Model performance

Four different common text classifiers, mentioned in the Methods section, were applied to the manually labelled Group A data. The statistical metrics, including accuracy, precision, and recall, can be seen in Table 3 and 4. The accuracies between the classifiers and each classification technique for housing stability were overall fairly high ranging from 84.36-92.18%. The accuracies for tobacco and alcohol use were lower, ranging from 70.87-84.68% for tobacco use and 69.95-82.79% for alcohol use. Additionally, for each top performing model, the most influential words for text classification, for each social factor, can be seen in Table 5. The best performing classification models were selected for each social factor and were used to apply the model to our entire corpus in Group B.

Table 3: Accuracies amongst text classifiers

	n=1	n=1-2
Multinomial naïve Bayes	Housing: 91.62% Tobacco: 70.87% Alcohol: 70.77%	Housing: 91.43% Tobacco: 77.18% Alcohol: 69.95%
Support vector machine	Housing: **92.18%** Tobacco: 81.08% Alcohol: 76.50%	Housing: 91.99% Tobacco: 82.88% Alcohol: 81.97%
Logistic regression	Housing: 84.36% Tobacco: 75.38% Alcohol: 77.60%	Housing: 90.13% Tobacco: **84.68%** Alcohol: **82.79%**
Random forest	Housing: 90.50% Tobacco: 76.28% Alcohol: 71.31%	Housing: 91.25% Tobacco: 78.98% Alcohol: 75.68%

Table 4: Best performing classifier detailed metrics

	Classifier	Accuracy	Recall	Precision	F1
Housing status*	Support vector machine (n=1)	0.92	0.93/0.91 (0/1)	0.94/0.90	0.93/0.91
Tobacco use**	Logistic Regression (n=1-2)	0.85	0.82/0.95/0.86 (0,1,2)	0.96/0.43/0.87 (0,1,2)	0.88/0.60/0.87 (0,1,2)
Alcohol use**	Logistic Regression (n=1-2)	0.83	0.86/0.73/0.81 (0,1,2)	0.93/0.44/0.88 (0,1,2)	0.89/0.55/0.84 (0,1,2)

* 0: no use, 1: current use
** 0: no use, 1: rare/occasional/history, 2: current use

Social factor (Classifier)	Top 20 weighted words
Housing stability (support vector machine, n=1)	`['friends' 'motel' 'stay' 'cigs' 'found' 'street' 'stays' 'streets' 'van' 'incarcerated' 'desc' 'currently' 'undomiciled' 'friend' 'respite' 'kcj'`

	'shelters' 'homelessness' 'shelter' 'homeless']
No tobacco use (logistic regression, n=1,2)	['use denies' 'deneis' 'lives' 'tobacco drug' 'seattle denies' 'use results' 'lives seattle' 'alcohol tobacco' 'tobacco drugs' 'never smoker' 'etoh tobacco' 'drinking' 'seattle tobacco' 'denies cigarettes' 'drugs tobacco' 'denies alcohol' 'tobacco alcohol' 'denies smoking' 'denies' 'denies tobacco']
No alcohol use (logistic regression, n=1,2)	['care' 'ppd' 'tobacco' 'smoking' 'etoh tobacco' 'history cocaine' 'tobacco alcohol' 'etoh illicit' 'alcohol tobacco' 'etoh drug' 'drugs etoh' 'alcohol drug' 'use none' 'alcohol drugs' 'drug etoh' 'denies alcohol' 'lives' 'denies drug' 'denies etoh' 'denies']

Scoring results and comparison

After classifying text for housing stability, tobacco use, and alcohol use for patients in Group B, we applied a scoring metric scheme, described in the Methods section. We generated three different scores that were calculated and weighted differently based on time. Our final score weighs more recent note entries and their resulting classification score higher than notes from previous years as social factors and their influence can change over time. Using the same process, we extracted and scored housing stability, tobacco use, and alcohol use with structured data sources and compared the results with the unstructured process.

I. Housing stability

Using notes, we classified 839 patients as housing unstable, a score above 0.5, and 21,370 patients as housing stable, a score of 0.5 and below. In total, we classified 22,209 patients with this text classification workflow, which covered 50.71% of the acute care patients within the same timeframe. When compared with structured data sources, only 791 (1.81%) additional patients were found.

II. Tobacco use

We classified 4,911 patients as currently using tobacco, regardless of amount or degree (1.5-2) using text notes. We classified 1,480 patients as having rare/occasional/past use of tobacco (0.5-1.5), and 7,139 patients as not using tobacco (0-0.5). In total, we classified 13,530 patients with this text classification workflow, which covered 30.9% of the acute care patients within the same timeframe. When compared with structured data sources, 17,9351 (40.9%) additional patients were captured.

III. Alcohol use

We classified 2,738 patients as currently using alcohol, regardless of amount or degree (1.5-2) using text notes. We classified 4,050 patients as having rare/occasional/past use of alcohol (0.5-1.5), and 13,885 patients as not drinking alcohol (0-0.5). In total, we classified 20,673 patients with this text classification workflow, which covered 37% of the acute care patients within the same timeframe. When compared with structured data sources, no additional patients were found.

IV. DISCUSSION

Our approach to a simple text classification method for various social determinants of health have shown positive results. The selected classification models were chosen as they were the most commonly used classification models when researching text classification techniques. Furthermore, these models were robust enough to curtail the need for more complex machine learning based text classification methods, which may be harder to interpret in the clinical space as the weights and decisions can be confiscated due to the black box nature of these more complex classification methods. Generally, linear models are fast to train, can work well with sparse data, and offer interpretability.[33] Additionally, recent research has also suggested that more complex machine learning approaches may not yield statistically significant improvements in predictive power to justify the time and effort necessary to implement and test these more complex methods. Although promising, more advanced methods of NLP, such as convoluted neural networks, may not provide a significant tradeoff in improvement or accuracy versus transparent understanding of rule-based approaches. In fact, Yao et al. found that the F1 scores for CNN via TensorFlow did not improve significantly for interested features when compared to logistic regression and support vector machine implementations.[34] Finally, generalizable methods to create institution-specific models can be better for the healthcare system as a whole as each institution records clinical information with variances.

Although SDoH information and other social factors can be indicative of overall health, collection of SDoH heavily relies on clinical staff to screen and document SDoH. Furthermore, it also assumes that patients will respond accurately and truthfully. Various financial incentives from the federal level have propelled collection of social factors, such as tobacco use and tobacco cessation. However, other social factors, which can be equally as important, such as alcohol use are not incentivized to be captured; rather only more severe instances are incentivized, such as alcohol dependence or alcohol addiction or disorder.[35,36] Due to this discrepancy, we found that structured data sources were less reliable, and that text classification aided in detailing a patient more holistically.

Our text classification of unstructured data relied solely on ED, admit, social work, and ambulatory notes. Social factors and other social history could also be recorded in other locations. Furthermore, social work and ambulatory notes used for housing status only and were only extracted if the notes contained a word or phrase related to housing instability. This approach was used as the notes were typically stored in a more unstructured format compared to the ED and admit notes; there were no section headers. The lack of section headers increased the difficulty to extract the notes and the notes would often verbiage that would interfere with the simple text classification approach that we used. Therefore, we decided to extract notes that contained words relating to housing instability. Additionally, tobacco and alcohol use notes had stylistic and grammatical challenges. These social factors were often grouped together in incomplete triples (e.g. "denies drinking, smoking, illicit drug use"). The classification algorithms often had trouble reciprocating the negative connotation to all components of the triple. Therefore, we used regex to specifically extract these triples and classify the note based on the presence of words related to tobacco or alcohol. These results would then override the text classification algorithm, if there was a discrepancy. Therefore, the scoring metrics for these cases would not necessarily reflect the accuracy or performance of our scoring method.

Limitations

Our study has numerous limitations. There were two distinct areas in our workflow that required manual attention: (1) EHR review and (2) labelling of features. Manual EHR review was performed to ensure that the notes contained social history information in a consistent location prior to widespread text extraction. We initially validated this with a random set of 10 patients, but later expanded our validation to 25 patients. We felt that having consistent results with the 25 patients indicated a high level of confidence. Manual labelling of features was time consuming and taxing. Although only one author performed the feature labelling, having multiple team members would provide better and possibly more consistent classification.

This approach, although we aim to create a generalizable workflow, is still stunted by local customizations due to unique nuances in note taking language. Patients can downplay or lie about their social challenges, making text classification harder to perform due to incorrect incoming data streams. Our approach relies on the fact that the patient has been seen within the healthcare system at some point in the past five years. This approach would not be applicable to those who are new to the institution or those who are not immediately identifiable. Classification levels for unstructured notes are not concrete as descriptive wording is also not concrete and can vary (e.g. "patient was a former smoker", "patient quit last week", "patient is an occasional smoker", etc.). Structured data sources can add a more concrete sense to the classification. There were 5.7% copy-forward entries present as data collection of social factors may not always be appropriate (e.g. patient is inebriated, in an altered mental state, etc.). We did not incorporate outside ontologies, such as UMLS or MetaMap, as we were interested in creating a simple text classification approach that did not need to rely on outside entities. Furthermore, we believe that these ontologies would not have added a significant improvement in our approach due to the social factors (housing, alcohol, tobacco) that were investigated. Although minimized, applying NLP to clinical notes will always present limitations and risks with biased models, biased data, and data privacy.[37]

V. CONCLUSION

From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status. Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use. Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social determinants and can supplement current structured sources to provide a more complete social history for patients. However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

Contributor statement:
AT performed the data extraction, tool building, and analysis. AB provided guidance and verification when needed.

Competing interests:
There are no competing interests.

Funding:
This work was supported by the U.S. Department of Health and Human Services, National Library of Medicine Training Grant T15LM007442.

Data sharing statement
The data used are unable to be shared due to patient privacy, confidentiality and United States healthcare laws.

VI. REFERENCES

1. Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. *Health Affairs*. 2017;36(8):1416-1422. doi:10.1377/hlthaff.2016.1651

2. Mennemeyer ST, Menachemi N, Rahurkar S, Ford EW. Impact of the HITECH Act on physicians' adoption of electronic health records. *Journal of the American Medical Informatics Association*. 2016;23(2):375-379. doi:10.1093/jamia/ocv103

3. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*. 2014;2(1). doi:10.1186/2047-2501-2-3

4. Structured and Unstructured Health Data: Challenges and Opportunities. Published October 29, 2018. https://healthsnap.io/structured-unstructured-health-data/

5. Hood CM, Gennuso KP, Swain GR, Catlin BB. County Health Rankings. *American Journal of Preventive Medicine*. 2016;50(2):129-135. doi:10.1016/j.amepre.2015.08.024

6. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving Electronic Medical Records Upstream. *American Journal of Preventive Medicine*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

7. Social Determinants of Health. HealthyPeople.gov. Accessed February 1, 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

8. Social Determinants. Institute for Health Metrics and Evaluation. Accessed February 1, 2020. http://www.healthdata.org/social-determinants

9. Nerenz DR. Health Care Organizations' Use Of Race/Ethnicity Data To Address Quality Disparities. *Health Affairs*. 2005;24(2):409-416. doi:10.1377/hlthaff.24.2.409

10. Andermann A. Taking action on the social determinants of health in clinical practice: a framework for health professionals. *Canadian Medical Association Journal*. 2016;188(17-18):E474-E483. doi:10.1503/cmaj.160177

11. Olson DP, Oldfield BJ, Navarro SM. Standardizing Social Determinants Of Health Assessments. Published March 18, 2019. https://www.healthaffairs.org/do/10.1377/hblog20190311.823116/full/

12. Wockenfuss R, Frese T, Herrmann K, Claussnitzer M, Sandholzer H. Three- and four-digit ICD-10 is not a reliable classification system in primary care. *Scand J Prim Health Care*. 2009;27(3):131-136. doi:10.1080/02813430903072215

13. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40(5 Pt 2):1620-1639. doi:10.1111/j.1475-6773.2005.00444.x

14. Lax Y, Martinez M, Brown NM. Social Determinants of Health and Hospital Readmission. *Pediatrics*. 2017;140(5):e20171427. doi:10.1542/peds.2017-1427

15. Meghan Henry, Anna Mahathey, Tyler Morrill, Anna Robinson, Azim Shivji, and Rian Watt, Abt Associates. The 2018 Annual Homeless Assessment Report (AHAR) to Congress. Published online December 2018. https://files.hudexchange.info/resources/documents/2018-AHAR-Part-1.pdf

16. Stafford A, Wood L. Tackling Health Disparities for People Who Are Homeless? Start with Social Determinants. *International Journal of Environmental Research and Public Health*. 2017;14(12):1535. doi:10.3390/ijerph14121535

17. Ahmad S, Baig S, Taneja A, Nanchal R, Kumar G. The Outcomes of Severe Sepsis in Homeless. *Chest*. 2014;146(4):230A. doi:10.1378/chest.1995140

18. Bambra C, Gibson M, Sowden A, Wright K, Whitehead M, Petticrew M. Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. *Journal of Epidemiology & Community Health*. 2010;64(4):284-291. doi:10.1136/jech.2008.082743

19. K. Papadopoulou S, N. Hassapidou M, Katsiki N, et al. Relationships Between Alcohol Consumption, Smoking Status and Food Habits in Greek Adolescents. Vascular Implications for the Future. *Current Vascular Pharmacology*. 2017;15(2):167-173. doi:10.2174/1570161114666161024123357

20. Eva Wong. Tobacco Use in King County. Published online May 2012. https://www.kingcounty.gov/depts/health/data/~/media/depts/health/data/documents/tobacco-use-in-king-county-may-2012.ashx

21. King County Community Health Needs Assessment 2018/2019. Presented at the: https://www.kingcounty.gov/depts/health/data/community-health-indicators/~/media/depts/health/data/documents/2018-2019-Joint-CHNA-Report.ashx

22. Bogan S, Donohue B. King County drug and alcohol deaths rose 9.5% in 2018. https://newsroom.uw.edu/news/king-county-drug-and-alcohol-deaths-rose-95-2018

23. Drug-caused deaths in King County. Published online February 21, 2017. https://adai.washington.edu/WAdata/KingCountyDrugDeaths.htm

24. 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.

25. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513. doi:10.1136/jamia.2009.001560

26. Gundlapalli AV, Carter ME, Divita G, et al. Extracting Concepts Related to Homelessness from the Free Text of VA Electronic Medical Records. *AMIA Annu Symp Proc*. 2014;2014:589-598.

27. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE*. Published online 2017. doi:10.1371/journal.pone.0174708

28. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment: *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2018;77(2):160-166. doi:10.1097/QAI.0000000000001580

29. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying Patients with Significant Problems Related to Social Determinants of Health with Natural Language Processing. *Stud Health Technol Inform*. 2019;264:1456-1457. doi:10.3233/SHTI190482

30. Berg K, Doktorchik C, Quan H, Saini V. *Meaningful Information in the Age of Big Data: A Scoping Review on Social Determinants of Health Data Collection for Electronic Health Records*. In Review; 2019. doi:10.21203/rs.2.16433/v1

31. 2015 CDC HA-VTE PREVENTION CHALLENGE CHAMPION. https://www.cdc.gov/ncbddd/dvt/documents/champ-fact-sheet-harborview.pdf

32. Bulger EM, Kastl JG, Maier RV. The history of Harborview Medical Center and the Washington State Trauma System. *Trauma Surgery & Acute Care Open*. 2017;2(1):e000091. doi:10.1136/tsaco-2017-000091

33. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*. 2017;105:110-120. doi:10.1016/j.ijmedinf.2017.06.004

34. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*. 2019;19(S3). doi:10.1186/s12911-019-0781-4

35. Medicare & Medicaid EHR Incentive Program: Meaningful Use Stage 1 Requirements Overview. Presented at the: 2010. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/MU_Stage1_ReqOverview.pdf

36. Eligible Professional Meaningful Use Core Measures Measure 9 of 13. Published online May 2014. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/9_Record_Smoking_Status.pdf

37. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H. Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*. Published online June 4, 2020:161-168. doi:10.14745/ccdr.v46i06a02

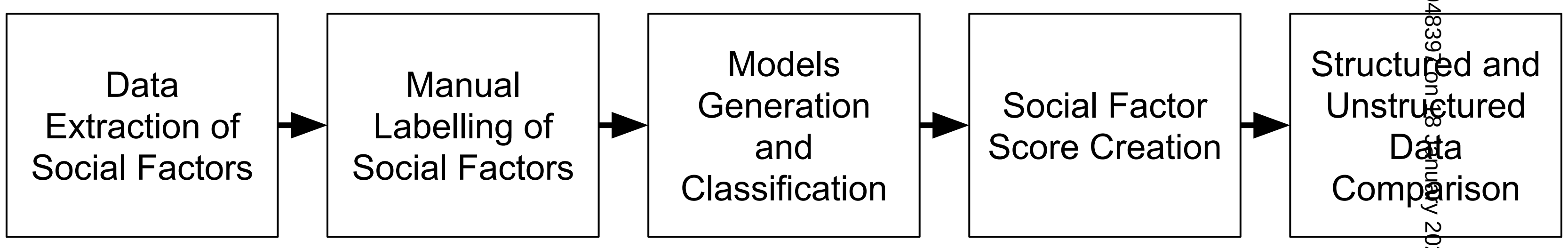
Acknowledgements:
Sally Lee
Abdelhak Abdou
Marion Granich
David Carlbom

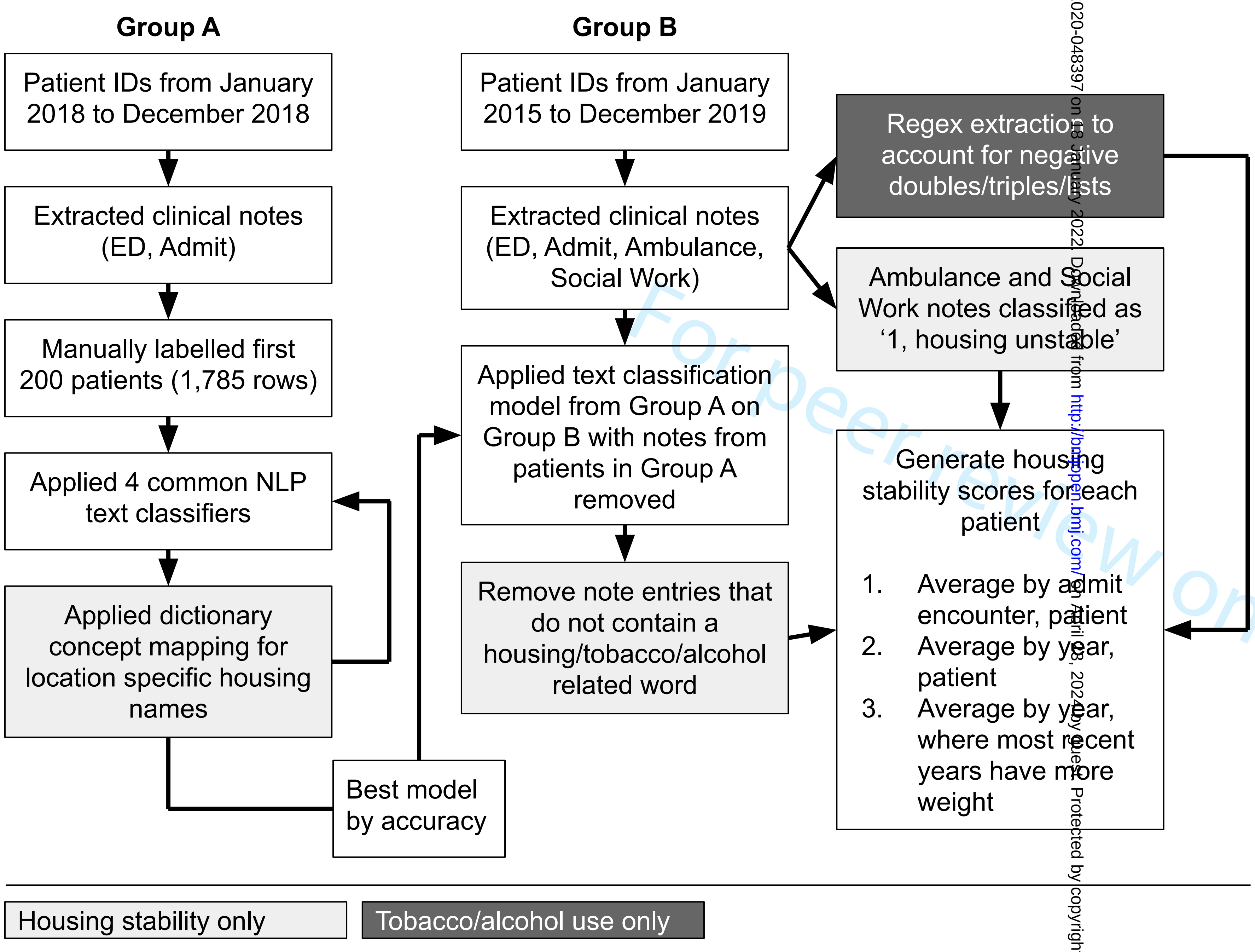
Figure legend:

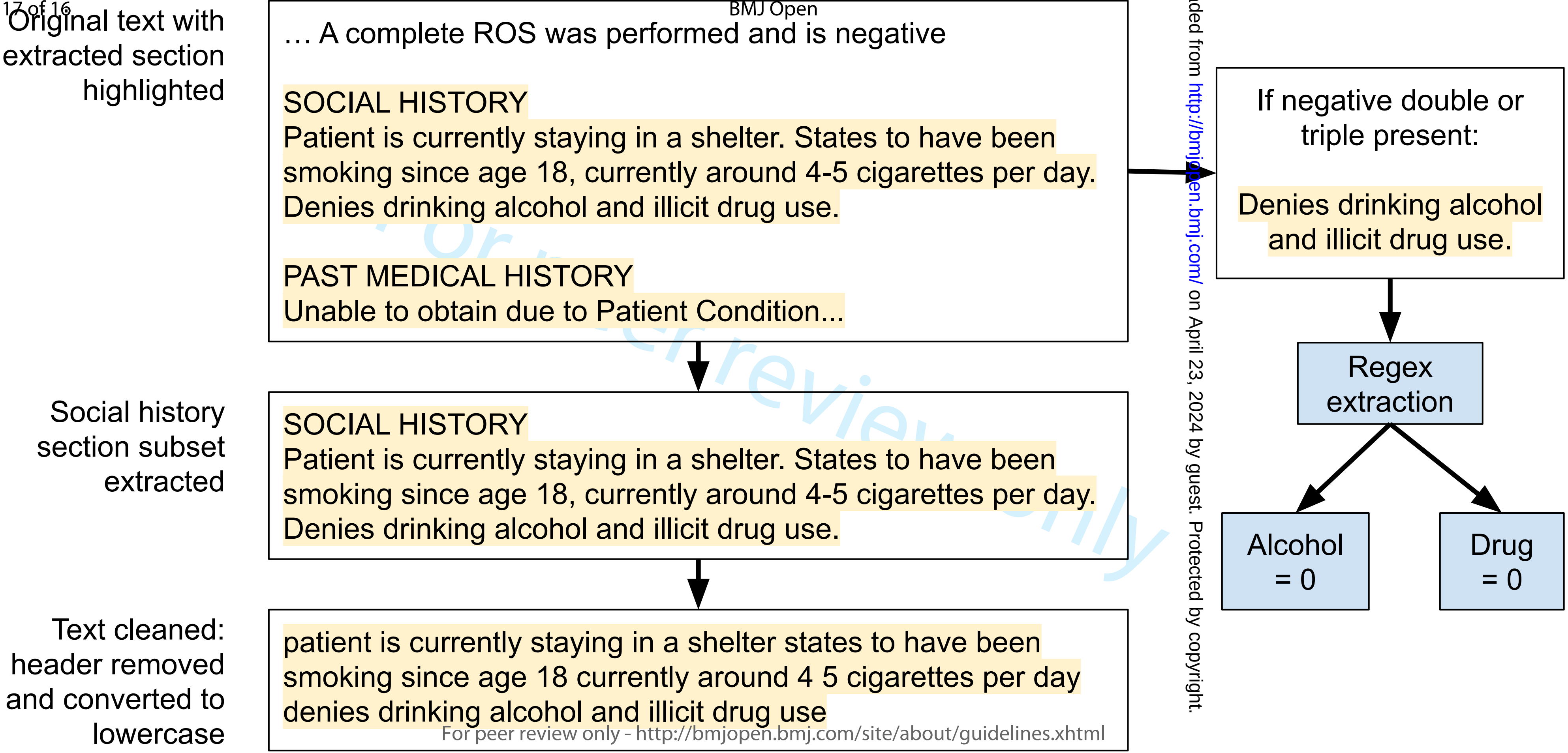
Figure 1: High-level overview of the workflow process

Figure 2: Text extraction, classification, and scoring workflow

Figure 3: Text extraction and cleaning process. Additional steps were performed for notes when classifying text related to tobacco and alcohol use to extract negative sentiment doubles or triples.


Data Extraction of Social Factors
Manual Labelling of Social Factors
Models Generation and Classification
Social Factor Score Creation
Structured and Unstructured Data Comparison


Group A
Group B
Patient IDs from January 2018 to December 2018
Extracted clinical notes (ED, Admit)
Manually labelled first 200 patients (1,785 rows)
Applied 4 common NLP text classifiers
Applied dictionary concept mapping for location specific housing names
Best model by accuracy
Patient IDs from January 2015 to December 2019
Extracted clinical notes (ED, Admit, Ambulance, Social Work)
Applied text classification model from Group A on Group B with notes from patients in Group A removed
Remove note entries that do not contain a housing/tobacco/alcohol related word
Regex extraction to account for negative doubles/triples/lists
Ambulance and Social Work notes classified as '1, housing unstable'
Generate housing stability scores for each patient
1. Average by admit encounter, patient
2. Average by year, patient
3. Average by year, where most recent years have more weight
Housing stability only
Tobacco/alcohol use only


Original text with extracted section highlighted
… A complete ROS was performed and is negative
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
PAST MEDICAL HISTORY
Unable to obtain due to Patient Condition...
Social history section subset extracted
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
Text cleaned: header removed and converted to lowercase
patient is currently staying in a shelter states to have been smoking since age 18 currently around 4 5 cigarettes per day denies drinking alcohol and illicit drug use
If negative double or triple present:
Denies drinking alcohol and illicit drug use.
Regex extraction
Alcohol = 0
Drug = 0

BMJ Open

A Simplified Framework to Extract Social and Behavioral Determinants: A Data Science Approach

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-048397.R1
Article Type:	Original research
Date Submitted by the Author:	31-Aug-2021
Complete List of Authors:	Teng, Andrew; University of Washington, Biomedical Informatics and Medical Education Wilcox, Adam ; University of Washington,
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Health informatics
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, SOCIAL MEDICINE, HISTORY (see Medical History), BIOTECHNOLOGY & BIOINFORMATICS

A Simplified Framework to Extract Social and Behavioral Determinants: A Data Science Approach

Andrew K. Teng[1], Adam B. Wilcox[1]

Affiliation:
[1] Biomedical Informatics and Medical Education
University of Washington
Seattle, WA, United States

Corresponding Author:
Andrew K. Teng
akteng@uw.edu
Biomedical Informatics and Medical Education, University of Washington
850 Republican Street, Box 358047
Seattle, WA 98195-0005, United States

Keywords: Text extraction, text classification, housing stability, social determinants, data science

Word Count: 4,228

Abstract

Objectives
We aim to extract a subset of social factors from clinical notes using common text classification methods.
Setting
We collaborated with a local Level I trauma hospital located in an underserved area that has a housing unstable patient population of about 6.5% and extracted text notes related to various social determinants for acute care patients.
Participants
Notes were retrospectively extracted from 43,798 acute care patients.
Methods
We solely utilize open source Python packages to test simple text classification methods that can potentially be easily generalizable and implemented. We extracted social history text from various sources, such as admission and emergency department notes, over a five-year timeframe and performed manual chart reviews to ensure data quality. We manually labelled the sentiment of the notes, treating each text entry independently. Four different models with two different feature selection methods (bag of words (BOW) and bigrams) were used to classify and predict housing stability, tobacco use, and alcohol use status for the extracted clinical text.
Results
From our analysis, we found overall positive results and metrics in applying open-source classification techniques; the accuracy scores were 91.2%, 84.7%, 82.8% for housing stability, tobacco use, and alcohol use respectively. There were many limitations in our analysis including social factors not present due to patient condition, multiple copy-forward entries and shorthand. Additionally, it was difficult to translate usage degrees for tobacco and alcohol use. However, when compared to structured data sources, our classification approach on unstructured notes yielded more results for housing and alcohol use; tobacco use proved less fruitful for unstructured notes.

Article Summary
Strengths and limitations of this study

- From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status.
- Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use.
- Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social and behavioral determinants and can supplement current structured sources to provide a more complete social history for patients.
- However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

I. INTRODUCTION

Most data can be generally categorized as structured or unstructured, where structured data can consist of items such as vital signs and lab results and unstructured data can consist of items such as text notes, images, or multimedia.[1] Although structured data can generally be easier to extract and analyze, unstructured data can potentially provide an array of information not present or easily identifiable in structured data. Challenges arise with unstructured data as they are not as easily interpretable as or categorizable as a numeric structured value. Images and text often contain many levels of metadata that would need manual review to decode or interpret. Additionally, clinicians have recently expanded intake data and social determinants of health (SDoH) information are starting to become more readily available. Furthermore, there has a been a growing interest around Medicaid patients, as SDoH can drive up to

80% of health outcomes, especially within this patient demographic.[2] Therefore, SDoH and REAL (Race, Ethnicity and Language) data are now starting to be analyzed for secondary research as recent research has indicated that there is a correlation between SDoH and health outcomes and the increasing need to research health disparities across populations.[3]

SDoH and REAL can include housing stability, access jobs and health care services, education level, language, and socioeconomic conditions.[4] These indicators are descriptors of different societies and are useful as predictors of health outcomes and the uptake of health interventions.[5] Because they can potentially be powerful indicators of health, many institutions are now starting to analyze and intake SDoH and REAL information, whether through text notes or standardized coding, such as International Classification of Diseases (ICD).[6] Additionally, SDoH can provide health teams with a greater understanding of a patient condition holistically.[7] However, there are challenges with SDoH intake as there is no standardized SDoH screening tool in the EHR across institutions[8]; additionally, coding schemes like ICD can prove to be unreliable in secondary analysis as coding can oversimplify symptoms and diagnoses leading to coding uncertainties and the fact that coding errors may be present from unintentional mistakes or even upcoding.[9,10] Additionally certain SDoH data may be more complete than others due to reimbursement incentives or other priorities.(cite) Past research has shown that hospital readmissions are highly influenced by patient health status and SDoH and suggest that clinical staff and researchers should consider SDoH when assessing readmission risk.[11]

The 2018-2019 [redacted for review] Community Health Needs Assessment (CHNA) reported the results from a health needs assessment survey given to residents to identify regional perceived healthcare issues. It was determined that housing affordability and housing stability were major challenges dominating overall health.[12] Mental health was also highlighted as a challenge for healthcare providers; mental illness can be caused by depression, schizophrenia, and alcohol and substance-related disorders.[12] The CHNA reported that adults in the lowest income tier were about 15 times more likely to experience severe psychological distress compared to their high-income counterparts. Additionally, it noted that part of the region had continued challenges with adult smoking rates.[12] Locally, it is estimated that there are at least 22,000 homeless individuals in [redacted for review] and more than 12,000 people in the [redacted for review] region, a four percent increase over the previous year.[13] Housing instability is associated with various health inequalities, such as shorter life expectancy, higher morbidity, and increased usage of acute hospital services, “as the social determinants of homelessness and health inequities are often intertwined, and long term homelessness further exacerbates poor health”.[14] It is therefore important to treat housing stability and other SDoH as a combined health issue to aid in improving health outcomes in clinical settings. Although some research has shown that patients who experience housing instability are more likely to die following admission for severe sepsis than those with insurance,[15] other research indicates that the effects of health inequalities are still unclear and need further investigation.[16] Additionally, various behavioral habits, including tobacco and alcohol use, although may not directly be considered a SDoH, can impact health decisions and outcomes. For example, one study found that participants who drank alcohol and reported tobacco use consumed more foods higher in fat and sugar, low in vitamins and minerals as well as foods, considered by them to be less healthy and prepared in a less healthy way.[17]

Within our region, it has been noted in recent years that the smoking rate is around 13 percent; however, among Black/African-Americans or individuals with multiple races, is double the rate among white adults and four times higher than Asian adults. Additionally, it was reported that, when compared to high income households, low income households were three times more likely to be smokers.[12,18] Drug and alcohol use also shared similar metrics; within the region, “drug and alcohol-caused deaths was 22% higher among Blacks and four times greater among American Indian/Alaskan Native than among non-Hispanic Whites” and alcohol use represented 4.97 per 100,000 deaths locally in 2015.[19,20] Therefore it may be important to look at social determinants and health behaviors, together known as social and behavioral determinants of health (SBDH) to better understand the patient population.[17]

Recent technological advances in machine learning and artificial intelligence have shown great potential in providing a pathway for informaticians and clinicians to better understand unstructured data.

Within the clinical setting, there have been numerous approaches in adopting natural language processing (NLP) to aid with processing unstructured clinical text notes. Common uses of NLP include extracting diagnoses and chief complaints as well as grouping of information for quality improvement. There are various NLP methods that can be used in the clinical setting, such as automatic tagging of conditions or variables of interest, sentiment classification, or even text extraction. Various open source NLP and ontological tools, such as Automated Retrieval Console, Apache clinical Text Analysis and Knowledge Extraction System (Apache cTAKES), MetaMap, and HITEx, Unified Medical Language System (UMLS) Metathesaurus and BioPortal have been used to aid with text extraction or classification.[21–23] On the other hand less complex classification methods have been used as well to identify specific groups of patients, risk assessment, or aid in validating structured annotation.[24,25,26] A recent scoping review found that although practitioners collect a variety of SBDH data at point of care through EHR, the overall use of automated technology is limited to date.[27]

With the idea of implementing an easily generalizable approach to classify selected social factors, we extracted both unstructured and structured data sources related to SBDH from a local hospital to identify and generate a framework to automatically extract and classify SBDH from text notes. We focused on housing stability status, tobacco use, and alcohol use. These three social factors were chosen due to their direct impact on health outcomes and the local public health impact[13–17] and presence in the EHR. To tackle challenges associated with SBDH extraction from unstructured text notes, we aimed to create a generalizable framework using low barrier open-source tools that are commonly used in the data science field. Because notes and stylistic choices can be institution and location specific, we sought not to create a model that is generalizable but rather a simplified method that could be potentially easily implemented using common off the shelf NLP and data science tools.

II. METHODS

Study Design and Overview

A high-level overview of our workflow can be seen in Figure 1. We retrospectively extracted patient data from the acute care setting at a Level I trauma center and academic teaching hospital with the aim to create a general and easily applicable workflow to extract and classify SBDH factors from clinical notes. We applied a two-pronged approach and collected unstructured data from a subset of patients over a 1-year timespan (Group A) to create and test the text classification model and also collected structured and unstructured data from a subset of patients over a 5-year timespan (Group B) to apply the best model created from Group A and compare results between the two data types. We performed automatic classification and scoring of patients via various NLP classification methods on three social factors: (1) housing stability, (2) tobacco use, and (3) alcohol use. Our general workflow for housing stability, a similar approach was also used for tobacco and alcohol use, can be seen in Figure 2.

Study Population

Data were extracted from [redacted for review], a 413-bed academic hospital that has a patient population consisting mostly from Washington, but also from a five-state area.[28] In 2014, there were 17,121 inpatient admissions, where 19 percent of the patients belong to a racial or ethnic minority and 37 percent of patients were enrolled in Medicaid.[28,29] Additionally, in 2015, the non-US born population was estimated to be around 21 percent in [redacted for review], highlighting the potential diversity that could be found with this patient population.[29]

Data Sources, Extraction, and Validation

We extracted both structured and unstructured data sources related to housing stability, tobacco use, and alcohol use using SQL queries called directly from an integrated python-based Jupyter Notebook:

a. Structured data sources include billing and diagnostic/International Classification of Disease (ICD) 9 and 10 codes, questionnaire or Epic SmartForm responses, address fields (location), problem list (ICD 9), patient encounters, clinical events (actual encounters of care), and discharge/disposition location.
b. Unstructured data sources consisted of text notes from the emergency department (ED), admission (admit) notes, social work, and ambulance notes.

Discharge notes were not explored as they were not recorded in the same subdivided format as the admit and ED notes, making selective text extraction of SBDH difficult. From our initial list of patient identifiers over a one-year timespan from Group A, we performed manual EHR validation of a random subset of 50 patients to validate the completeness of the clinical notes and confirm the location of social history and social factors in clinical notes. Extensive research and conversations with an internal data analyst confirmed the location of these topics (housing, tobacco use, and alcohol use) within structured data sources.

Data Cleaning
After confirmation, clinical notes were extracted for both Groups A and B. The notes were cleaned (e.g. symbols removed, converted to lowercase) prior to classification and analysis in the python Jupyter notebook via NLTK. Our general text extraction and cleaning workflow can be seen in Figure 3. However, housing stability notes and tobacco or alcohol use notes were stylistically and grammatically different, and both sets needed distinct additional cleaning steps. Housing stability notes that contained the phrase 'not homeless' were converted via regex to say 'housed' instead. Additionally, for housing stability, a concept dictionary was also created to substitute local facility names with more general concept (e.g. 'Union Gospel Mission' was converted to 'shelter'). This was done to explore how the algorithms handle formal nouns.

For text notes in Group B, we performed an additional concept extraction step. Tobacco use and alcohol use notes often contained incomplete (lacking the subject, predicate, object format) triples or doubles (e.g. 'Denies smoking, drinking, drugs'). Due to their incomplete sentence structures, common NLP tools to parse, extract, and classify triples, such as Stanford CoreNLP, were not suitable as these tools rely on having all three parts of the triple present. These notes related to tobacco and alcohol use therefore underwent an additional step that performed a separate relation extraction that first pulls out the SBDH related objects and then would reclassify and label the negative sentiment to all components of the list. Our process can be seen in the left side of Figure 3. If the regex extraction of negative lists resulted in a different result from the text classification prediction, the regex extraction would overwrite the end result prior to scoring. Once these steps were performed, the data were considered clean and suitable for classification.

Model building
Cleaned text from Group A were used to generate and test the classification models. These notes were split in 70/30 validation and testing sets. We applied four different common NLP text classification models to the testing sets (via SciKit Learn): multinomial naïve Bayes, support vector machine, logistic regression, and random forest. Default parameters and a bag-of-words approach were used. The best performing model by accuracy was then chosen and applied to the larger corpus, Group B, with notes from patients in Group A removed, to avoid overfitting and classification bias. This process was performed for housing, tobacco use, and alcohol use.

Scoring generation
In order to create a simple method of identifying patients who are experiencing social instability, we created a scoring metric based on the classified notes. After applying the optimum model by accuracy to the entire corpus of extracted text notes, housing stability, tobacco use, and alcohol use scores were generated. Patient identifiers were mapped by patient location and those who were not in the acute care setting during this timeframe were removed. Three different scoring approaches were used to describe these social factors: (1) predictions were averaged by patient encounter, then averaged by patient identifier, (2) predictions were averaged by year, then by patient identifier, and (3) predictions were averaged by year, where each year then had a weight where the most recent year had the highest weight and the furthest year had the lowest weight (e.g. predictions from 2019 were weighted by a factor of 5 and predictions from 2015 were weighted by a factor of 1). This scoring generation process was then repeated on our structured data for all three social factors and the results were compared and analyzed. Structured data was also extracted for our list of patients in Group B.

Patient and Public Involvement

No patient involved. The retrospective exploration is a part of a larger study and was approved by the [redacted for review] Institutional Review Board #STUDY00006723. Patient data elements, including encounter identifiers, race, age, and notes with SBDH, were extracted directly from the data warehouse and stored on encrypted computers and were not distributed or shared outside of the secured and closed environment. No patient identifiers or names were stored in this analysis.

III. RESULTS

Characteristics of study subjects

Clinical notes (ED, admit, social work, and ambulance) between 2015 and 2019 were extracted and included, forming Group B. Notes from the first 200 patients were included in Group A and notes from 147,457 patients were included in Group B. During the same timeframe, 61,767 patients were in acute care. After extraction and model prediction, the patient notes were cross referenced with inpatient location and only notes from those who were in acute care were retained, for a total of 43,798 patients from 2015 to 2019. The patient demographics of this final subset were 63% (*n*=27,575) male, 37% (*n*=16,223) female, 88.2% (*n*=38,634) not Hispanic or Latino, and 10.5% (*n*=4,609) Hispanic or Latino, and 1.3% (*n*=555) unknown or not answered. Further descriptive statistics can be found in Table 1.

Table 1: Population demographics

Race (*n*=43,798)	*n* (%)
White or Caucasian	31,575 (72.1%)
Black or African American	4,812 (11.0%)
Asian	3,174 (7.2%)
American Indian or Alaska Native	1,165 (2.7%)
Native Hawaiian or other Pacific Islander	524 (1.2%)
Multiple races	3 (0%)
Unavailable, unknown, or missing	2,545 (5.8%)
Age range (*n*=43,798)	***n* (%)**
0-18	1,856 (4.2%)
19-44	12,437 (28.4%)
45-64	14,863 (33.9%)
65-84	11,902 (27.2%)
85 and over	2,740 (6.3%)

Data attributes

Table 2 illustrates the amount of data for each corresponding extraction level, specifically for housing status. We first started with extracting text from the ED and admit notes, forming Group A, which consisted of 50,000 rows or text entries and covered 3,200 unique patients, over a one-year timeframe. From there, we manually labelled housing stability concepts in a binary fashion, where 0 would indicate housing stability and 1 would indicate any level of housing instability, regardless of severity. As manual labelling can be a labor-intensive process, only the first 6,000 text rows were labelled, covering 218 unique patients. However, within these first 6,000 rows, numerous notes did not contain text that alluded to housing status or were empty due to patient condition. Therefore, only 1,785 out of the 6,000 rows were labelled, covering 200 unique patients, where 995 (55.7%) were labelled as housing stable and 790 (44.3%) were labelled as housing unstable. We also found that 5.7% of the entries within this subset were duplicates or copy-forward entries. The same workflow was performed for labelling tobacco and alcohol use. However, only 1,108 rows were labelled for tobacco use and 1,220 rows for alcohol use, where in both cases 0 indicated no use, 1 indicated rare/previous/occasional use, and 2 indicated current use, regardless of degree. Tobacco use resulted in 446 (40.3%) labels for no use, 129 (11.6%) labels for rare/previous/occasional use, and 533 (48.1%) labels for current use. Similarly, alcohol use resulted in 595 (48.8%) labels for no use, 185 (15.2%) labels for rare/previous/occasional use, and 440 (36%) labels for current use.

Table 2: Extracted data amounts for housing status

Level of extraction	Rows (*n*)	Unique patients (*n*)	Unique encounters (*n*)	Social history entries (*n*/unique)
ED and Admit notes	49,955	3,233	15,664	21,876/21,334
Housing, Tobacco, Alcohol Information	6,000	218	1,995	2,408/2,211
Remove nulls/missing data	Housing: 1,785 Tobacco: 1,108 Alcohol: 1,220	Housing: 200 Tobacco: 179 Alcohol: 181	1,361	1,785/1,684

Model performance

Four different common text classifiers, mentioned in the Methods section, were applied to the manually labelled Group A data. The statistical metrics, including accuracy, precision, and recall, can be seen in Table 3 and 4. The accuracies between the classifiers and each classification technique for housing stability were overall fairly high ranging from 84.36-92.18%. The accuracies for tobacco and alcohol use were lower, ranging from 70.87-84.68% for tobacco use and 69.95-82.79% for alcohol use. Additionally, for each top performing model, the most influential words for text classification, for each social factor, can be seen in Table 5. The best performing classification models were selected for each social factor and were used to apply the model to our entire corpus in Group B.

Table 3: Accuracies amongst text classifiers

	n=1	n=1-2
Multinomial naïve Bayes	Housing: 91.62% Tobacco: 70.87% Alcohol: 70.77%	Housing: 91.43% Tobacco: 77.18% Alcohol: 69.95%
Support vector machine	Housing: **92.18%** Tobacco: 81.08% Alcohol: 76.50%	Housing: 91.99% Tobacco: 82.88% Alcohol: 81.97%
Logistic regression	Housing: 84.36% Tobacco: 75.38% Alcohol: 77.60%	Housing: 90.13% Tobacco: **84.68%** Alcohol: **82.79%**
Random forest	Housing: 90.50% Tobacco: 76.28% Alcohol: 71.31%	Housing: 91.25% Tobacco: 78.98% Alcohol: 75.68%

Table 4: Best performing classifier detailed metrics

	Classifier	Accuracy	Recall	Precision	F1
Housing status*	Support vector machine (n=1)	0.92	0.93/0.91 (0/1)	0.94/0.90	0.93/0.91
Tobacco use**	Logistic Regression (n=1-2)	0.85	0.82/0.95/0.86 (0,1,2)	0.96/0.43/0.87 (0,1,2)	0.88/0.60/0.87 (0,1,2)
Alcohol use**	Logistic Regression (n=1-2)	0.83	0.86/0.73/0.81 (0,1,2)	0.93/0.44/0.88 (0,1,2)	0.89/0.55/0.84 (0,1,2)

* 0: no use, 1: current use

** 0: no use, 1: rare/occasional/history, 2: current use

Table 5: Word or phrase importance ranking

Social factor (Classifier)	Top 20 weighted words
Housing stability (support vector machine, n=1)	`['friends' 'motel' 'stay' 'cigs' 'found' 'street' 'stays' 'streets' 'van'`

	'incarcerated' 'desc' 'currently' 'undomiciled' 'friend' 'respite' 'kcj' 'shelters' 'homelessness' 'shelter' 'homeless']
No tobacco use (logistic regression, n=1,2)	['use denies' 'deneis' 'lives' 'tobacco drug' 'seattle denies' 'use results' 'lives seattle' 'alcohol tobacco' 'tobacco drugs' 'never smoker' 'etoh tobacco' 'drinking' 'seattle tobacco' 'denies cigarettes' 'drugs tobacco' 'denies alcohol' 'tobacco alcohol' 'denies smoking' 'denies' 'denies tobacco']
No alcohol use (logistic regression, n=1,2)	['care' 'ppd' 'tobacco' 'smoking' 'etoh tobacco' 'history cocaine' 'tobacco alcohol' 'etoh illicit' 'alcohol tobacco' 'etoh drug' 'drugs etoh' 'alcohol drug' 'use none' 'alcohol drugs' 'drug etoh' 'denies alcohol' 'lives' 'denies drug' 'denies etoh' 'denies']

Scoring results and comparison

After classifying text for housing stability, tobacco use, and alcohol use for patients in Group B, we applied a scoring metric scheme, described in the Methods section. We generated three different scores that were calculated and weighted differently based on time. Our final score weighs more recent note entries and their resulting classification score higher than notes from previous years as social factors and their influence can change over time. Using the same process, we extracted and scored housing stability, tobacco use, and alcohol use with structured data sources and compared the results with the unstructured process.

I. Housing stability

Using notes, we classified 839 patients as housing unstable, a score above 0.5, and 21,370 patients as housing stable, a score of 0.5 and below. In total, we classified 22,209 patients with this text classification workflow, which covered 50.71% of the acute care patients within the same timeframe. When compared with structured data sources, only 791 (1.81%) additional patients were found.

II. Tobacco use

We classified 4,911 patients as currently using tobacco, regardless of amount or degree (1.5-2) using text notes. We classified 1,480 patients as having rare/occasional/past use of tobacco (0.5-1.5), and 7,139 patients as not using tobacco (0-0.5). In total, we classified 13,530 patients with this text classification workflow, which covered 30.9% of the acute care patients within the same timeframe. When compared with structured data sources, 17,9351 (40.9%) additional patients were captured.

III. Alcohol use

We classified 2,738 patients as currently using alcohol, regardless of amount or degree (1.5-2) using text notes. We classified 4,050 patients as having rare/occasional/past use of alcohol (0.5-1.5), and 13,885 patients as not drinking alcohol (0-0.5). In total, we classified 20,673 patients with this text classification workflow, which covered 37% of the acute care patients within the same timeframe. When compared with structured data sources, no additional patients were found.

IV. DISCUSSION

Our approach to a simple text classification method for various social determinants of health have shown positive results. The selected classification models were chosen as they were the most commonly used classification models when researching text classification techniques. Furthermore, these models were robust enough to curtail the need for more complex machine learning based text classification methods, which may be harder to interpret in the clinical space as the weights and decisions can be confiscated due to the black box nature of these more complex classification methods. Generally, linear models are fast to train, can work well with sparse data, and offer interpretability.[30] Additionally, recent research has also suggested that more complex machine learning approaches may not yield statistically significant improvements in predictive power to justify the time and effort necessary to implement and test these more complex methods. Although promising, more advanced methods of NLP, such as convoluted neural networks, may not provide a significant tradeoff in improvement or accuracy versus transparent understanding of rule-based approaches. In fact, Yao et al. found that the F1 scores for CNN via TensorFlow did not improve significantly for interested features when compared to logistic regression and support vector machine implementations.[31] Finally, generalizable methods to create institution-specific models can be better for the healthcare system as a whole as each institution records clinical information with variances.

Although SBDH information and other social factors can be indicative of overall health, collection of SBDH heavily relies on clinical staff to screen and document SBDH. Furthermore, it also assumes that patients will respond accurately and truthfully. Various financial incentives from the federal level have propelled collection of social factors, such as tobacco use and tobacco cessation. However, other social factors, which can be equally as important, such as alcohol use are not incentivized to be captured; rather only more severe instances are incentivized, such as alcohol dependence or alcohol addiction or disorder.[32,33] Due to this discrepancy, we found that structured data sources were less reliable, and that text classification aided in detailing a patient more holistically.

Our text classification of unstructured data relied solely on ED, admit, social work, and ambulatory notes. Social factors and other social history could also be recorded in other locations. Furthermore, social work and ambulatory notes used for housing status only and were only extracted if the notes contained a word or phrase related to housing instability. This approach was used as the notes were typically stored in a more unstructured format compared to the ED and admit notes; there were no section headers. The lack of section headers increased the difficulty to extract the notes and the notes would often verbiage that would interfere with the simple text classification approach that we used. Therefore, we decided to extract notes that contained words relating to housing instability. Additionally, tobacco and alcohol use notes had stylistic and grammatical challenges. These social factors were often grouped together in incomplete triples (e.g. “denies drinking, smoking, illicit drug use”). The classification algorithms often had trouble reciprocating the negative connotation to all components of the triple. Therefore, we used regex to specifically extract these triples and classify the note based on the presence of words related to tobacco or alcohol. These results would then override the text classification algorithm, if there was a discrepancy. Therefore, the scoring metrics for these cases would not necessarily reflect the accuracy or performance of our scoring method.

It was interesting to find that tobacco use was recorded significantly more often in structured data sources compared to alcohol use and housing stability. However, because tobacco use is a (Centers for Medicare and Medicare Services) CMS core quality measure, it can be expected that this feature is more available in structured form as it is often directly asked to the patient on intake forms, screeners, or during cessation treatment.[33] Furthermore, the Joint Commission created the Tobacco Performance Measure Set, which are three standardized performance measures addressing tobacco screening and cessation counseling: (1) Tobacco use screening of patients 18 years and over, (2) Tobacco use treatment, including counseling and medication during hospitalization, and (3) Tobacco use treatment management plan at discharge. CMS began using these performance measures in 2016.[34] Because alcohol consumption is not a recommended CMS core quality measure for adults, the amount of data regarding alcohol use is not complete in structured form as it may not be consistently collected during intake procedures.

Past research has consistently pointed towards SBDH impacting patient health and outcomes. However, collection of SBDH can be a major limiting factor in the ability to model and integrate these data. There has not been a standardized collection process for SBDH data across the institution, whether it is recorded through notes or electronic forms. Additionally, many times, SBDH data may not be asked due to patient condition or it might not be updated regularly. Providers and healthcare institutions should strive to collect SBDH data more regularly even if the data fields are not empty as SBDH status can change. These intake procedures should be present and not optional; currently, only language preference must be completed due to translation laws in place. Additionally, educating patients to utilize patient portals and update information via these portals can provide more current SBDH information. However, we should note that vulnerable populations would most likely not be the primary audience to utilize this feature, and this is the subpopulation that arguably needs more attention.

Limitations

Our study has numerous limitations. There were two distinct areas in our workflow that required manual attention: (1) EHR review and (2) labelling of features. Manual EHR review was performed to ensure that the notes contained social history information in a consistent location prior to widespread text extraction. We initially validated this with a random set of 10 patients, but later expanded our validation to 25 patients. We felt that having consistent results with the 25 patients indicated a high level of confidence. Manual labelling of features was time consuming and taxing. Although only one author performed the feature labelling, having multiple team members would provide better and possibly more consistent classification.

This approach, although we aim to create a generalizable workflow, is still stunted by local customizations due to unique nuances in note taking language. Patients can withhold information about their social challenges, making text classification harder to perform due to incorrect incoming data streams. Our approach relies on the fact that the patient has been seen within the healthcare system at some point in the past five years. This approach would not be applicable to those who are new to the institution or those who are not immediately identifiable. Classification levels for unstructured notes are not concrete as descriptive wording is also not concrete and can vary (e.g. “patient was a former smoker”, “patient quit last week”, “patient is an occasional smoker”, etc.). Structured data sources can add a more concrete sense to the classification. There were 5.7% copy-forward entries present as data collection of social factors may not always be appropriate (e.g. patient is inebriated, in an altered mental state, etc.). We did not incorporate outside ontologies, such as UMLS or MetaMap, as we were interested in creating a simple text classification approach that did not need to rely on outside entities. Furthermore, we believe that these ontologies would not have added a significant improvement in our approach due to the social factors (housing, alcohol, tobacco) that were investigated. Although minimized, applying NLP to clinical notes will always present limitations and risks with biased models, biased data, and data privacy.[35]

Community needs are constantly changing as the health of the community is not static. Currently, the King County CHNA has identified obesity, healthcare access, insurance status and drug use as other potential SBDH information to explore. These data types would be stored in different areas of the EHR and within different notes. It would be interesting to see if our designed workflow presented could be applicable and generalized to meet the needs of other SBDH data. Although we aimed to create a simplified framework to extract SBDH data from clinical notes, more complex methods such as convoluted neural networks and more advanced NLP part of speech tagging may be worth exploring as they may help improve accuracy and precision of the classification. As more notes become available for patients, it will also be important to keep in mind the potential bias of having more notes present from sicker patients and evaluating ways to reduce this bias.

We sourced data from solely one medical center. Patients might have had encounters or other visit types in neighboring hospitals and healthcare systems in the region. The lack of data sharing between institutions prevents holistic collection of SBDH data. Data completeness is vitally important to the quality and accuracy of models that are dependent on big data. Poor data quality and completeness lead to lower utilization and the lack of data can potentially lead to mistakes in the decision-making process; additionally, since there is no single or standardized source for SBDH data, the diversity of data and complexity of the associated data structures increase the difficulty and bottlenecks for data integration.[36]

The lack of a standardized methodology to collect and store all SBDH data will limit the potential of this research field. Additionally, SBDH factors are constantly changing for patients as their behaviors can change depending on their circumstance. Being able to aggregate these data and create adaptable models is crucial as these features are never static. Furthermore, public health and outreach services fluctuate over time. Creating a method or utilizing an API to update the list of community shelters and other places for homeless services would be necessary to maintain an accurate understanding of a patients housing status.

V. CONCLUSION

From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status. Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use. Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social determinants and can supplement current structured sources to provide a more complete social history for patients. However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

Contributor statement:
AT performed the data extraction, tool building, and analysis. AB provided guidance and verification when needed.

Competing interests:
There are no competing interests.

Funding:
This work was supported by the U.S. Department of Health and Human Services, National Library of Medicine Training Grant T15LM007442.

Data sharing statement
The data used are unable to be shared due to patient privacy, confidentiality and United States healthcare laws.

Ethics statement
This research is a part of a larger study that has been approved by the University of Washington IRB #STUDY00006723.

VI. REFERENCES

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*. 2014;2(1). doi:10.1186/2047-2501-2-3

2. Hood CM, Gennuso KP, Swain GR, Catlin BB. County Health Rankings. *American Journal of Preventive Medicine*. 2016;50(2):129-135. doi:10.1016/j.amepre.2015.08.024

3. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving Electronic Medical Records Upstream. *American Journal of Preventive Medicine*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

4. Social Determinants of Health. HealthyPeople.gov. Accessed February 1, 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

5. Social Determinants. Institute for Health Metrics and Evaluation. Accessed February 1, 2020. http://www.healthdata.org/social-determinants

6. Nerenz DR. Health Care Organizations' Use Of Race/Ethnicity Data To Address Quality Disparities. *Health Affairs*. 2005;24(2):409-416. doi:10.1377/hlthaff.24.2.409

7. Andermann A. Taking action on the social determinants of health in clinical practice: a framework for health professionals. *Canadian Medical Association Journal*. 2016;188(17-18):E474-E483. doi:10.1503/cmaj.160177

8. Olson DP, Oldfield BJ, Navarro SM. Standardizing Social Determinants Of Health Assessments. Published March 18, 2019. https://www.healthaffairs.org/do/10.1377/hblog20190311.823116/full/

9. Wockenfuss R, Frese T, Herrmann K, Claussnitzer M, Sandholzer H. Three- and four-digit ICD-10 is not a reliable classification system in primary care. *Scand J Prim Health Care*. 2009;27(3):131-136. doi:10.1080/02813430903072215

10. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40(5 Pt 2):1620-1639. doi:10.1111/j.1475-6773.2005.00444.x

11. Lax Y, Martinez M, Brown NM. Social Determinants of Health and Hospital Readmission. *Pediatrics*. 2017;140(5):e20171427. doi:10.1542/peds.2017-1427

12. King County Community Health Needs Assessment 2018/2019. Presented at the: https://www.kingcounty.gov/depts/health/data/community-health-indicators/~/media/depts/health/data/documents/2018-2019-Joint-CHNA-Report.ashx

13. Henry M, Mahathey A, Morrill T, et al. *The 2018 Annual Homeless Assessment Report (AHAR) to Congress*. The U.S. Department of Housing and Urban Development OFFICE OF COMMUNITY PLANNING AND DEVELOPMENT; 2018. https://files.hudexchange.info/resources/documents/2018-AHAR-Part-1.pdf

14. Stafford A, Wood L. Tackling Health Disparities for People Who Are Homeless? Start with Social Determinants. *International Journal of Environmental Research and Public Health*. 2017;14(12):1535. doi:10.3390/ijerph14121535

15. Ahmad S, Baig S, Taneja A, Nanchal R, Kumar G. The Outcomes of Severe Sepsis in Homeless. *Chest*. 2014;146(4):230A. doi:10.1378/chest.1995140

16. Bambra C, Gibson M, Sowden A, Wright K, Whitehead M, Petticrew M. Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. *Journal of Epidemiology & Community Health*. 2010;64(4):284-291. doi:10.1136/jech.2008.082743

17. Papadopoulou S, Hassapidou M, Katsiki N, et al. Relationships Between Alcohol Consumption, Smoking Status and Food Habits in Greek Adolescents. Vascular

Implications for the Future. *Current Vascular Pharmacology*. 2017;15(2):167-173. doi:10.2174/1570161114666161024123357

18. Wong E. *Tobacco Use in King County*. Public Health Seattle & King County; 2012. https://www.kingcounty.gov/depts/health/data/~/media/depts/health/data/documents/tobacco-use-in-king-county-may-2012.ashx

19. Bogan S, Donohue B. King County drug and alcohol deaths rose 9.5% in 2018. https://newsroom.uw.edu/news/king-county-drug-and-alcohol-deaths-rose-95-2018

20. Drug-caused deaths in King County. Published online February 21, 2017. https://adai.washington.edu/WAdata/KingCountyDrugDeaths.htm

21. 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.

22. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513. doi:10.1136/jamia.2009.001560

23. Gundlapalli AV, Carter ME, Divita G, et al. Extracting Concepts Related to Homelessness from the Free Text of VA Electronic Medical Records. *AMIA Annu Symp Proc*. 2014;2014:589-598.

24. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE*. Published online 2017. doi:10.1371/journal.pone.0174708

25. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment: *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2018;77(2):160-166. doi:10.1097/QAI.0000000000001580

26. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying Patients with Significant Problems Related to Social Determinants of Health with Natural Language Processing. *Stud Health Technol Inform*. 2019;264:1456-1457. doi:10.3233/SHTI190482

27. Berg K, Doktorchik C, Quan H, Saini V. *Meaningful Information in the Age of Big Data: A Scoping Review on Social Determinants of Health Data Collection for Electronic Health Records*. In Review; 2019. doi:10.21203/rs.2.16433/v1

28. 2015 CDC HA-VTE PREVENTION CHALLENGE CHAMPION. https://www.cdc.gov/ncbddd/dvt/documents/champ-fact-sheet-harborview.pdf

29. Bulger EM, Kastl JG, Maier RV. The history of Harborview Medical Center and the Washington State Trauma System. *Trauma Surgery & Acute Care Open*. 2017;2(1):e000091. doi:10.1136/tsaco-2017-000091

30. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*. 2017;105:110-120. doi:10.1016/j.ijmedinf.2017.06.004

31. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*. 2019;19(S3). doi:10.1186/s12911-019-0781-4

32. Medicare & Medicaid EHR Incentive Program: Meaningful Use Stage 1 Requirements Overview. Presented at the: 2010. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/MU_Stage1_ReqOverview.pdf

33. Eligible Professional Meaningful Use Core Measures Measure 9 of 13. Published online May 2014. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/9_Record_Smoking_Status.pdf

34. Quality Measures and Tobacco Cessation. https://www.bhthechange.org/wp-content/uploads/2017/12/Quality-Measures-and-Tobacco-Cessation.pdf

35. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H. Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*. Published online June 4, 2020:161-168. doi:10.14745/ccdr.v46i06a02

36. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *CODATA*. 2015;14(0):2. doi:10.5334/dsj-2015-002

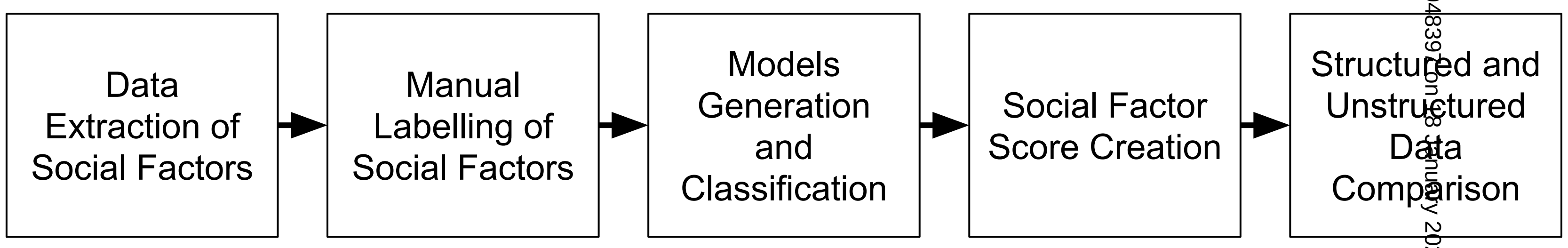
Acknowledgements:
Sally Lee
Abdelhak Abdou
Marion Granich
David Carlbom

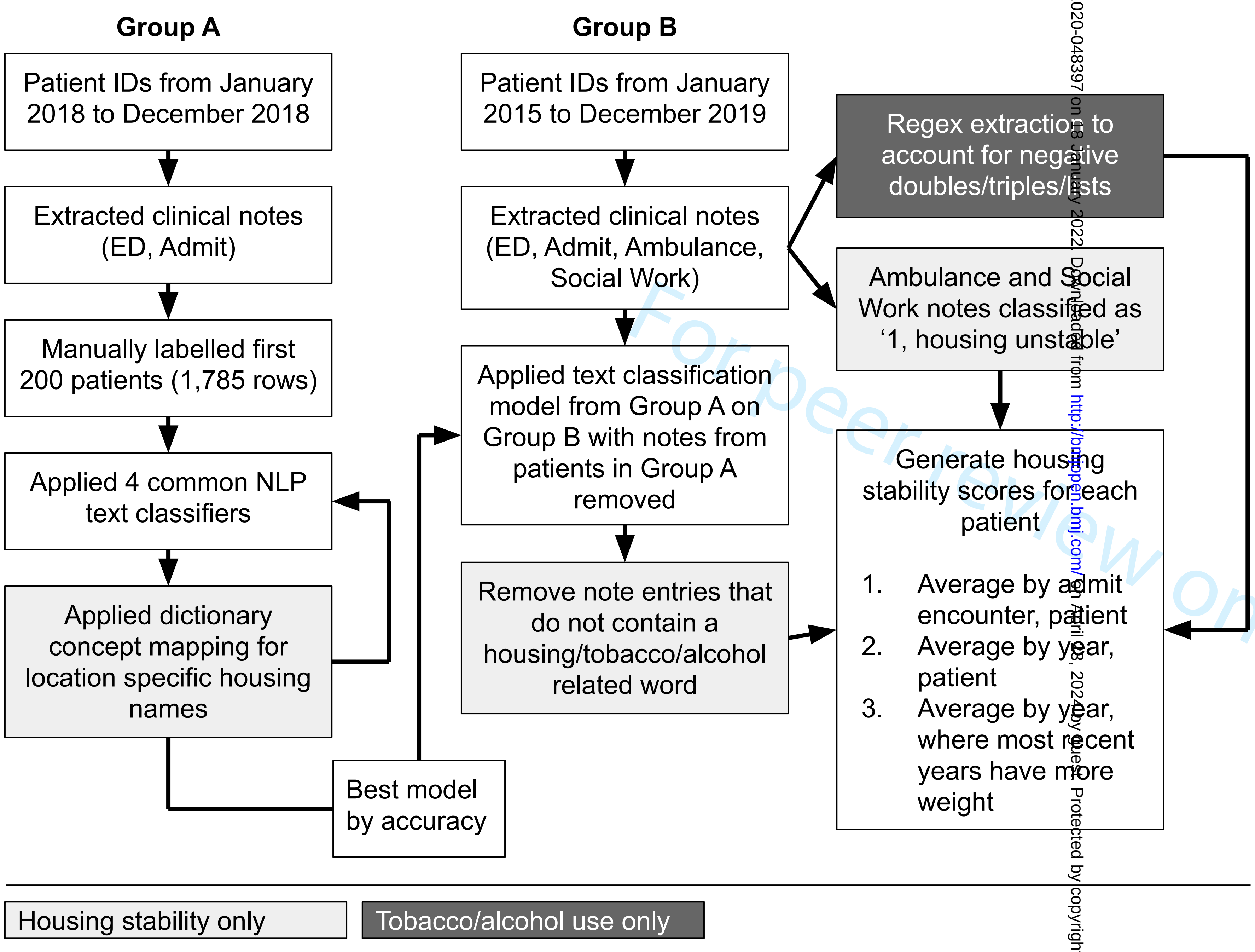
Figure legend:

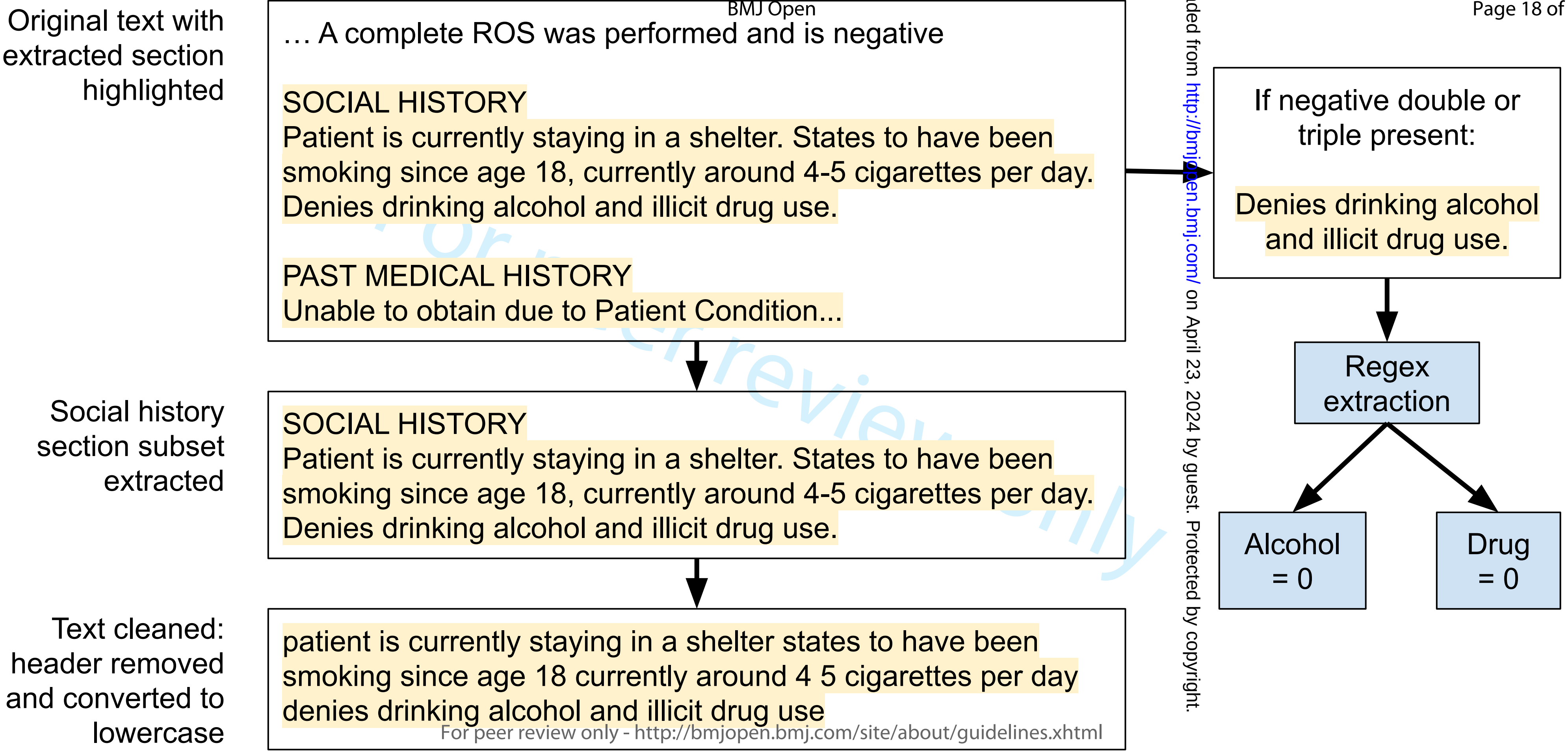
Figure 1: High-level overview of the workflow process

Figure 2: Text extraction, classification, and scoring workflow

Figure 3: Text extraction and cleaning process. Additional steps were performed for notes when classifying text related to tobacco and alcohol use to extract negative sentiment doubles or triples.


Data Extraction of Social Factors
Manual Labelling of Social Factors
Models Generation and Classification
Social Factor Score Creation
Structured and Unstructured Data Comparison


Group A
Group B
Patient IDs from January 2018 to December 2018
Extracted clinical notes (ED, Admit)
Manually labelled first 200 patients (1,785 rows)
Applied 4 common NLP text classifiers
Applied dictionary concept mapping for location specific housing names
Best model by accuracy
Patient IDs from January 2015 to December 2019
Extracted clinical notes (ED, Admit, Ambulance, Social Work)
Applied text classification model from Group A on Group B with notes from patients in Group A removed
Remove note entries that do not contain a housing/tobacco/alcohol related word
Regex extraction to account for negative doubles/triples/lists
Ambulance and Social Work notes classified as '1, housing unstable'
Generate housing stability scores for each patient
1. Average by admit encounter, patient
2. Average by year, patient
3. Average by year, where most recent years have more weight
Housing stability only
Tobacco/alcohol use only


Original text with extracted section highlighted
… A complete ROS was performed and is negative
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
PAST MEDICAL HISTORY
Unable to obtain due to Patient Condition...
Social history section subset extracted
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
Text cleaned: header removed and converted to lowercase
patient is currently staying in a shelter states to have been smoking since age 18 currently around 4 5 cigarettes per day denies drinking alcohol and illicit drug use
If negative double or triple present:
Denies drinking alcohol and illicit drug use.
Regex extraction
Alcohol = 0
Drug = 0

BMJ Open

A Simplified Framework to Extract Social and Behavioral Determinants: A Data Science Approach

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-048397.R2
Article Type:	Original research
Date Submitted by the Author:	25-Sep-2021
Complete List of Authors:	Teng, Andrew; University of Washington, Biomedical Informatics and Medical Education Wilcox, Adam ; University of Washington,
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Health informatics
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, SOCIAL MEDICINE, HISTORY (see Medical History), BIOTECHNOLOGY & BIOINFORMATICS

A Simplified Framework to Extract Social and Behavioral Determinants: A Data Science Approach

Andrew K. Teng[1], Adam B. Wilcox[1]

Affiliation:
[1] Biomedical Informatics and Medical Education
University of Washington
Seattle, WA, United States

Corresponding Author:
Andrew K. Teng
akteng@uw.edu
Biomedical Informatics and Medical Education, University of Washington
850 Republican Street, Box 358047
Seattle, WA 98195-0005, United States

Keywords: Text extraction, text classification, housing stability, social determinants, data science

Word Count: 4,985

Abstract

Objectives
We aim to extract a subset of social factors from clinical notes using common text classification methods.
Setting
We collaborated with a local Level I trauma hospital located in an underserved area that has a housing unstable patient population of about 6.5% and extracted text notes related to various social determinants for acute care patients.
Participants
Notes were retrospectively extracted from 43,798 acute care patients.
Methods
We solely utilize open source Python packages to test simple text classification methods that can potentially be easily generalizable and implemented. We extracted social history text from various sources, such as admission and emergency department notes, over a five-year timeframe and performed manual chart reviews to ensure data quality. We manually labelled the sentiment of the notes, treating each text entry independently. Four different models with two different feature selection methods (bag of words (BOW) and bigrams) were used to classify and predict housing stability, tobacco use, and alcohol use status for the extracted clinical text.
Results
From our analysis, we found overall positive results and metrics in applying open-source classification techniques; the accuracy scores were 91.2%, 84.7%, 82.8% for housing stability, tobacco use, and alcohol use respectively. There were many limitations in our analysis including social factors not present due to patient condition, multiple copy-forward entries and shorthand. Additionally, it was difficult to translate usage degrees for tobacco and alcohol use. However, when compared to structured data sources, our classification approach on unstructured notes yielded more results for housing and alcohol use; tobacco use proved less fruitful for unstructured notes.

Article Summary
Strengths and limitations of this study

- From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status.
- Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use.
- Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social and behavioral determinants and can supplement current structured sources to provide a more complete social history for patients.
- However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

I. INTRODUCTION

Most data can be generally categorized as structured or unstructured, where structured data can consist of items such as vital signs and lab results and unstructured data can consist of items such as text notes or images.[1] Although structured data can generally be easier to extract and analyze, unstructured data can potentially provide an array of information not present or easily identifiable in structured data. As healthcare institutions expand data collection to include non-clinical features, more unstructured data surrounding behavioral health and social determinants of health (SDoH) information, are starting to become more readily available. Furthermore, there has a been a growing interest around Medicaid patients, as SDoH can drive up to 80% of health outcomes, especially within this patient demographic.[2] Therefore, SDoH and REAL (Race, Ethnicity and Language) data are now being used for secondary analysis as recent research has indicated that there is a correlation between SDoH and health outcomes and the increasing need to research health disparities across populations.[3]

SDoH and REAL can include housing stability, access jobs and health care services, education level, language, and socioeconomic conditions.[4] These indicators are descriptors of different societies and are useful as predictors of health outcomes and the uptake of health interventions.[5] Because they can potentially be powerful indicators of health, many institutions are now starting to analyze and intake SDoH and REAL information, whether through text notes or standardized coding, such as International Classification of Diseases (ICD).[6] Additionally, SDoH can provide health teams with a greater understanding of a patient condition holistically.[7] However, there are challenges with SDoH intake as there is no standardized SDoH screening tool in the EHR across institutions[8]; additionally, coding schemes like ICD can prove to be unreliable in secondary analysis as coding can oversimplify symptoms and diagnoses leading to coding uncertainties and the fact that coding errors may be present from unintentional mistakes or even upcoding.[9,10] Additionally certain SDoH data may be more complete than others due to reimbursement incentives or other priorities.[11] Past research has shown that hospital readmissions are highly influenced by patient health status and SDoH and suggest that clinical staff and researchers should consider SDoH when assessing readmission risk.[12]

The 2018-2019 [redacted for review] Community Health Needs Assessment (CHNA) reported the results from a health needs assessment survey given to residents to identify regional perceived healthcare issues. It was determined that housing affordability and housing stability were major challenges dominating overall health.[13] Mental health was also highlighted as a challenge for healthcare providers; mental illness can be caused by depression, schizophrenia, and alcohol and substance-related disorders.[13] The CHNA reported that adults in the lowest income tier were about 15 times more likely to experience severe psychological distress compared to their high-income counterparts. Additionally, it noted that part of the region had continued challenges with adult smoking rates.[13] Locally, it is estimated that there are at least 22,000 homeless individuals in [redacted for review] and more than 12,000 people in the [redacted for review] region, a four percent increase over the previous year.[14] Housing instability is associated with various health inequalities, such as shorter life expectancy, higher morbidity, and increased usage of acute hospital services, “as the social determinants of homelessness and health inequities are often intertwined, and long term homelessness further exacerbates poor health”.[15] It is therefore important to treat housing stability and other SDoH as a combined health issue to aid in improving health outcomes in clinical settings. Although some research has shown that patients who experience housing instability are more likely to die following admission for severe sepsis than those with insurance,[16] other research indicates that the effects of health inequalities are still unclear and need further investigation.[17] Additionally, various behavioral habits, including tobacco and alcohol use, although may not directly be considered a SDoH, can impact health decisions and outcomes. For example, one study found that participants who drank alcohol and reported tobacco use consumed more foods higher in fat and sugar, low in vitamins and minerals as well as foods, considered by them to be less healthy and prepared in a less healthy way.[18]

Within our region, it has been noted in recent years that the smoking rate is around 13 percent; however, among Black/African-Americans or individuals with multiple races, is double the rate among white adults and four times higher than Asian adults. Additionally, it was reported that, when compared to high income households, low income households were three times more likely to be smokers.[13,19] Drug and alcohol use also shared similar metrics; within the region, “drug and alcohol-caused deaths was 22% higher among Blacks and four times greater among American Indian/Alaskan Native than among non-Hispanic Whites” and alcohol use represented 4.97 per 100,000 deaths locally in 2015.[20,21] Therefore it may be important to look at social determinants and health behaviors, together known as social and behavioral determinants of health (SBDH) to better understand the patient population.[18]

Recent technological advances in machine learning and artificial intelligence have shown great potential in providing a pathway for informaticians and clinicians to better understand unstructured data.
Within the clinical setting, there have been numerous approaches in adopting natural language processing (NLP) to aid with processing unstructured clinical text notes. Common uses of NLP include

extracting diagnoses and chief complaints as well as grouping of information for quality improvement. There are various NLP methods that can be used in the clinical setting, such as automatic tagging of conditions or variables of interest, sentiment classification, or even text extraction. Various open source NLP and ontological tools, such as Automated Retrieval Console, Apache clinical Text Analysis and Knowledge Extraction System (Apache cTAKES), MetaMap, and HITEx, Unified Medical Language System (UMLS) Metathesaurus and BioPortal have been used to aid with text extraction or classification.[22–24] On the other hand less complex classification methods have been used as well to identify specific groups of patients, risk assessment, or aid in validating structured annotation.[25,26,27] A recent scoping review found that although practitioners collect a variety of SBDH data at point of care through EHR, the overall use of automated technology is limited to date.[28]

With the idea of implementing an easily generalizable approach to classify selected social factors, we extracted both unstructured and structured data sources related to SBDH from a local hospital to identify and generate a framework to automatically extract and classify SBDH from text notes. We focused on housing stability status, tobacco use, and alcohol use. These three social factors were chosen due to their direct impact on health outcomes and the local public health impact[14–18] and presence in the EHR. To tackle challenges associated with SBDH extraction from unstructured text notes, we aimed to create a generalizable framework using low barrier open-source tools that are commonly used in the data science field. Because notes and stylistic choices can be institution and location specific, we sought not to create a model that is generalizable but rather a simplified method that could be potentially easily implemented using common off the shelf NLP and data science tools.

II. METHODS

Study Design and Overview

A high-level overview of our workflow can be seen in Figure 1. We retrospectively extracted patient data from the acute care setting at a Level I trauma center and academic teaching hospital with the aim to create a general and easily applicable workflow to extract and classify SBDH factors from clinical notes. We applied a two-pronged approach and collected unstructured data from a subset of patients over a 1-year timespan (Group A) to create and test the text classification model and also collected structured and unstructured data from a subset of patients over a 5-year timespan (Group B) to apply the best model created from Group A and compare results between the two data types. We performed automatic classification and scoring of patients via various NLP classification methods on three social factors: (1) housing stability, (2) tobacco use, and (3) alcohol use. Our general workflow for housing stability, a similar approach was also used for tobacco and alcohol use, can be seen in Figure 2.

Study Population

Data were extracted from [redacted for review], a 413-bed academic hospital that has a patient population consisting mostly from Washington, but also from a five-state area.[29] In 2014, there were 17,121 inpatient admissions, where 19 percent of the patients belong to a racial or ethnic minority and 37 percent of patients were enrolled in Medicaid.[29,30] Additionally, in 2015, the non-US born population was estimated to be around 21 percent in [redacted for review], highlighting the potential diversity that could be found with this patient population.[30]

Data Sources, Extraction, and Validation

We extracted both structured and unstructured data sources related to housing stability, tobacco use, and alcohol use using SQL queries called directly from an integrated python-based Jupyter Notebook:

a. Structured data sources include billing and diagnostic/International Classification of Disease (ICD) 9 and 10 codes, questionnaire or Epic SmartForm responses, address fields (location), problem list (ICD 9), patient encounters, clinical events (actual encounters of care), and discharge/disposition location.
b. Unstructured data sources consisted of text notes from the emergency department (ED), admission (admit) notes, social work, and ambulance notes.

Discharge notes were not explored as they were not recorded in the same subdivided format as the admit and ED notes, making selective text extraction of SBDH difficult. From our initial list of patient identifiers over a one-year timespan from Group A, we performed manual EHR validation of a random subset of 50 patients to validate the completeness of the clinical notes and confirm the location of social history and social factors in clinical notes. Extensive research and conversations with an internal data analyst confirmed the location of these topics (housing, tobacco use, and alcohol use) within structured data sources.

Data Cleaning

After confirmation, clinical notes were extracted for both Groups A and B. The notes were cleaned (e.g. symbols removed, converted to lowercase) prior to classification and analysis in the Python Jupyter notebook via NLTK. Our general text extraction and cleaning workflow can be seen in Figure 3. However, housing stability notes and tobacco or alcohol use notes were stylistically and grammatically different, and both sets needed distinct additional cleaning steps. Housing stability notes that contained the phrase ‘not homeless’ were converted via regex to say ‘housed’ instead. Additionally, for housing stability, a concept dictionary was also created to substitute local facility names with more general concept (e.g. ‘Union Gospel Mission’ was converted to ‘shelter’). This was done to explore how the algorithms handle formal nouns.

For text notes in Group B, we performed an additional concept extraction step. Tobacco use and alcohol use notes often contained incomplete (lacking the subject, predicate, object format) triples or doubles (e.g. ‘Denies smoking, drinking, drugs’). Due to their incomplete sentence structures, common NLP tools to parse, extract, and classify triples, such as Stanford CoreNLP, were not suitable as these tools rely on having all three parts of the triple present. These notes related to tobacco and alcohol use therefore underwent an additional step that performed a separate relation extraction that would first identify a negative sentiment word (e.g. denies), then individually extract the following SBDH related objects in the list by commas or conjunctions (e.g. and, or), and then label, or reclassify if necessary, the negative sentiment to all components of the list. Our process can be seen in the left side of Figure 3. If the regex extraction of negative lists resulted in a different result from the text classification prediction, the regex extraction would overwrite the end result prior to scoring. Once these steps were performed, the data were considered clean and suitable for classification.

Model building

Cleaned text from Group A were used to generate and test the classification models. These notes were split in 70/30 validation and testing sets. We applied four different common NLP text classification models to the testing sets (via SciKit Learn): multinomial naïve Bayes, support vector machine, logistic regression, and random forest. Default parameters and a bag-of-words approach were used. The best performing model by accuracy was then chosen and applied to the larger corpus, Group B, with notes from patients in Group A removed, to avoid overfitting and classification bias. This process was performed for housing, tobacco use, and alcohol use.

Scoring generation

In order to create a simple method of identifying patients who are experiencing social instability, we created a scoring metric based on the classified notes. After applying the optimum model by accuracy to the entire corpus of extracted text notes, housing stability, tobacco use, and alcohol use scores were generated. Patient identifiers were mapped by patient location and those who were not in the acute care setting during this timeframe were removed. Three different scoring approaches were used to describe these social factors: (1) predictions were averaged by patient encounter, then averaged by patient identifier, (2) predictions were averaged by year, then by patient identifier, and (3) predictions were averaged by year, where each year then had a weight where the most recent year had the highest weight and the furthest year had the lowest weight (e.g. predictions from 2019 were weighted by a factor of 5 and predictions from 2015 were weighted by a factor of 1). This scoring generation process was then repeated on our structured data for all three social factors and the results were compared and analyzed. Structured data was also extracted for our list of patients in Group B.

Patient and Public Involvement
No patients were involved. The retrospective exploration is a part of a larger study and was approved by the [redacted for review] Institutional Review Board #STUDY00006723. Patient data elements, including encounter identifiers, race, age, and notes with SBDH, were extracted directly from the data warehouse and stored on encrypted computers and were not distributed or shared outside of the secured and closed environment. No patient identifiers or names were stored in this analysis.

III. RESULTS

Characteristics of study subjects
Clinical notes (ED, admit, social work, and ambulance) between 2015 and 2019 were extracted and included, forming Group B. Notes from the first 200 patients were included in Group A and notes from 147,457 patients were included in Group B. During the same timeframe, 61,767 patients were in acute care. After extraction and model prediction, the patient notes were cross referenced with inpatient location and only notes from those who were in acute care were retained, for a total of 43,798 patients from 2015 to 2019. The patient demographics of this final subset were 63% (*n*=27,575) male, 37% (*n*=16,223) female, 88.2% (*n*=38,634) not Hispanic or Latino, and 10.5% (*n*=4,609) Hispanic or Latino, and 1.3% (*n*=555) unknown or not answered. Further descriptive statistics can be found in Table 1.

Table 1: Population demographics

Race (*n*=43,798)	*n* (%)
White or Caucasian	31,575 (72.1%)
Black or African American	4,812 (11.0%)
Asian	3,174 (7.2%)
American Indian or Alaska Native	1,165 (2.7%)
Native Hawaiian or other Pacific Islander	524 (1.2%)
Multiple races	3 (0%)
Unavailable, unknown, or missing	2,545 (5.8%)
Age range (*n*=43,798)	***n* (%)**
0-18	1,856 (4.2%)
19-44	12,437 (28.4%)
45-64	14,863 (33.9%)
65-84	11,902 (27.2%)
85 and over	2,740 (6.3%)

Data attributes
Table 2 illustrates the amount of data for each corresponding extraction level, specifically for housing status. We first started with extracting text from the ED and admit notes, forming Group A, which consisted of 50,000 rows or text entries and covered 3,200 unique patients, over a one-year timeframe. From there, we manually labelled housing stability concepts in a binary fashion, where 0 would indicate housing stability and 1 would indicate any level of housing instability, regardless of severity. As manual labelling can be a labor-intensive process, only the first 6,000 text rows were labelled, covering 218 unique patients. However, within these first 6,000 rows, numerous notes did not contain text that alluded to housing status or were empty due to patient condition. Therefore, only 1,785 out of the 6,000 rows were labelled, covering 200 unique patients, where 995 (55.7%) were labelled as housing stable and 790 (44.3%) were labelled as housing unstable. We also found that 5.7% of the entries within this subset were duplicates or copy-forward entries. The same workflow was performed for labelling tobacco and alcohol use. However, only 1,108 rows were labelled for tobacco use and 1,220 rows for alcohol use, where in both cases 0 indicated no use, 1 indicated rare/previous/occasional use, and 2 indicated current use, regardless of degree. Tobacco use resulted in 446 (40.3%) labels for no use, 129 (11.6%) labels for rare/previous/occasional use, and 533 (48.1%) labels for current use. Similarly, alcohol use resulted in 595 (48.8%) labels for no use, 185 (15.2%) labels for rare/previous/occasional use, and 440 (36%) labels for current use.

Table 2: Extracted data amounts for housing status

Level of extraction	Rows (*n*)	Unique patients (*n*)	Unique encounters (*n*)	Social history entries (*n*/unique)
ED and Admit notes	49,955	3,233	15,664	21,876/21,334
Housing, Tobacco, Alcohol Information	6,000	218	1,995	2,408/2,211
Remove nulls/missing data	Housing: 1,785 Tobacco: 1,108 Alcohol: 1,220	Housing: 200 Tobacco: 179 Alcohol: 181	1,361	1,785/1,684

Model performance

Four different common text classifiers, mentioned in the Methods section, were applied to the manually labelled Group A data. The statistical metrics, including accuracy, precision, and recall, can be seen in Table 3 and 4. The accuracies between the classifiers and each classification technique for housing stability were overall fairly high ranging from 84.36-92.18%. The accuracies for tobacco and alcohol use were lower, ranging from 70.87-84.68% for tobacco use and 69.95-82.79% for alcohol use. Additionally, for each top performing model, the most influential words for text classification, for each social factor, can be seen in Table 5. The best performing classification models were selected for each social factor and were used to apply the model to our entire corpus in Group B.

Table 3: Accuracies amongst text classifiers

	n=1	n=1-2
Multinomial naïve Bayes	Housing: 91.62% Tobacco: 70.87% Alcohol: 70.77%	Housing: 91.43% Tobacco: 77.18% Alcohol: 69.95%
Support vector machine	Housing: **92.18%** Tobacco: 81.08% Alcohol: 76.50%	Housing: 91.99% Tobacco: 82.88% Alcohol: 81.97%
Logistic regression	Housing: 84.36% Tobacco: 75.38% Alcohol: 77.60%	Housing: 90.13% Tobacco: **84.68%** Alcohol: **82.79%**
Random forest	Housing: 90.50% Tobacco: 76.28% Alcohol: 71.31%	Housing: 91.25% Tobacco: 78.98% Alcohol: 75.68%

Table 4: Best performing classifier detailed metrics

	Classifier	Accuracy	Recall	Precision	F1
Housing status*	Support vector machine (n=1)	0.92	0.93/0.91 (0/1)	0.94/0.90	0.93/0.91
Tobacco use**	Logistic Regression (n=1-2)	0.85	0.82/0.95/0.86 (0,1,2)	0.96/0.43/0.87 (0,1,2)	0.88/0.60/0.87 (0,1,2)
Alcohol use**	Logistic Regression (n=1-2)	0.83	0.86/0.73/0.81 (0,1,2)	0.93/0.44/0.88 (0,1,2)	0.89/0.55/0.84 (0,1,2)

* 0: no use, 1: current use
** 0: no use, 1: rare/occasional/history, 2: current use

Table 5: Word or phrase importance ranking

Social factor (Classifier)	Top 20 weighted words
Housing stability (support vector machine, n=1)	`['friends' 'motel' 'stay' 'cigs' 'found' 'street' 'stays' 'streets' 'van'`

	'incarcerated' 'desc' 'currently' 'undomiciled' 'friend' 'respite' 'kcj' 'shelters' 'homelessness' 'shelter' 'homeless']
No tobacco use (logistic regression, n=1,2)	['use denies' 'deneis' 'lives' 'tobacco drug' 'seattle denies' 'use results' 'lives seattle' 'alcohol tobacco' 'tobacco drugs' 'never smoker' 'etoh tobacco' 'drinking' 'seattle tobacco' 'denies cigarettes' 'drugs tobacco' 'denies alcohol' 'tobacco alcohol' 'denies smoking' 'denies' 'denies tobacco']
No alcohol use (logistic regression, n=1,2)	['care' 'ppd' 'tobacco' 'smoking' 'etoh tobacco' 'history cocaine' 'tobacco alcohol' 'etoh illicit' 'alcohol tobacco' 'etoh drug' 'drugs etoh' 'alcohol drug' 'use none' 'alcohol drugs' 'drug etoh' 'denies alcohol' 'lives' 'denies drug' 'denies etoh' 'denies']

Scoring results and comparison

After classifying text for housing stability, tobacco use, and alcohol use for patients in Group B, we applied a scoring metric scheme, described in the Methods section. We generated three different scores that were calculated and weighted differently based on time. Our final score weighs more recent note entries and their resulting classification score higher than notes from previous years as social factors and their influence can change over time. Using the same process, we extracted and scored housing stability, tobacco use, and alcohol use with structured data sources and compared the results with the unstructured process.

I. Housing stability

Using notes, we classified 839 patients as housing unstable, a score above 0.5, and 21,370 patients as housing stable, a score of 0.5 and below. In total, we classified 22,209 patients with this text classification workflow, which covered 50.71% of the acute care patients within the same timeframe. When compared with structured data sources, only 791 (1.81%) additional patients were found.

II. Tobacco use

We classified 4,911 patients as currently using tobacco, regardless of amount or degree (1.5-2) using text notes. We classified 1,480 patients as having rare/occasional/past use of tobacco (0.5-1.5), and 7,139 patients as not using tobacco (0-0.5). In total, we classified 13,530 patients with this text classification workflow, which covered 30.9% of the acute care patients within the same timeframe. When compared with structured data sources, 17,9351 (40.9%) additional patients were captured.

III. Alcohol use

We classified 2,738 patients as currently using alcohol, regardless of amount or degree (1.5-2) using text notes. We classified 4,050 patients as having rare/occasional/past use of alcohol (0.5-1.5), and 13,885 patients as not drinking alcohol (0-0.5). In total, we classified 20,673 patients with this text classification workflow, which covered 37% of the acute care patients within the same timeframe. When compared with structured data sources, no additional patients were found.

IV. DISCUSSION

Our approach to a simple text classification method for various social determinants of health have shown positive results. The selected classification models were chosen as they were the most commonly used classification models when researching text classification techniques. Furthermore, these models were robust enough to curtail the need for more complex machine learning based text classification methods, which may be harder to interpret in the clinical space as the weights and decisions can be confiscated due to the black box nature of these more complex classification methods. Generally, linear models are fast to train, can work well with sparse data, and offer interpretability.[31] Additionally, recent research has also suggested that more complex machine learning approaches may not yield statistically significant improvements in predictive power to justify the time and effort necessary to implement and test these more complex methods. Although promising, more advanced methods of NLP, such as convoluted neural networks, may not provide a significant tradeoff in improvement or accuracy versus transparent understanding of rule-based approaches. In fact, Yao et al. found that the F1 scores for CNN via TensorFlow did not improve significantly for interested features when compared to logistic regression and support vector machine implementations.[32] Finally, generalizable methods to create institution-specific models can be better for the healthcare system as a whole as each institution records clinical information with variances.

Although SBDH information and other social factors can be indicative of overall health, collection of SBDH heavily relies on clinical staff to screen and document SBDH. Furthermore, it also assumes that patients will respond accurately and truthfully. Various financial incentives from the federal level have propelled collection of social factors, such as tobacco use and tobacco cessation. However, other social factors, which can be equally as important, such as alcohol use are not incentivized to be captured; rather only more severe instances are incentivized, such as alcohol dependence or alcohol addiction or disorder.[33,11] Due to this discrepancy, we found that structured data sources were less reliable, and that text classification aided in detailing a patient more holistically.

Our text classification of unstructured data relied solely on ED, admit, social work, and ambulatory notes as our parsing and extraction method could only work with notes in a certain format with the social history heading. Social factors and other social history could also be recorded in other locations, but were not compatible with our approach. Furthermore, social work and ambulatory notes used for housing status only and were only extracted if the notes contained a word or phrase related to housing instability. This approach was used as the notes were typically stored in a more unstructured format compared to the ED and admit notes; there were no section headers. The lack of section headers increased the difficulty to extract the notes and the notes would often verbiage that would interfere with the simple text classification approach that we used. Therefore, we decided to extract notes that contained words relating to housing instability. Additionally, tobacco and alcohol use notes had stylistic and grammatical challenges. These social factors were often grouped together in incomplete triples (e.g. "denies drinking, smoking, illicit drug use"). The classification algorithms often had trouble reciprocating the negative connotation to all components of the triple. Therefore, we used regex to specifically extract these triples and classify the note based on the presence of words related to tobacco or alcohol. Without this additional data cleaning or manipulation step, the negative sentiment in a list would not have been applied to all elements within the list, but rather only the first element. In our example of 'denies smoking, drinking, drugs', the negative sentiment of 'denies' would have only been applied to smoking as smoking immediately follows 'denies'. However, with our additional concept extraction step, the negative sentiment of 'denies' is now also applied to 'drinking' and 'drugs'. These results would then override the text classification algorithm, if there was a discrepancy. Therefore, the scoring metrics for these cases would not necessarily reflect the accuracy or performance of our scoring method.

It was interesting to find that tobacco use was recorded significantly more often in structured data sources compared to alcohol use and housing stability. However, because tobacco use is a (Centers for Medicare and Medicare Services) CMS core quality measure, it can be expected that this feature is more available in structured form as it is often directly asked to the patient on intake forms, screeners, or during cessation treatment.[11] Furthermore, the Joint Commission created the Tobacco Performance Measure Set, which are three standardized performance measures addressing tobacco screening and cessation counseling: (1) Tobacco use screening of patients 18 years and over, (2) Tobacco use treatment, including counseling and medication during hospitalization, and (3) Tobacco use treatment management plan at discharge. CMS began using these performance measures in 2016.[34] Because alcohol consumption is not a recommended CMS core quality measure for adults, the amount of data regarding alcohol use is not complete in structured form as it may not be consistently collected during intake procedures.

Past research has consistently pointed towards SBDH impacting patient health and outcomes. However, collection of SBDH can be a major limiting factor in the ability to model and integrate these data. There has not been a standardized collection process for SBDH data across the institution, whether it is recorded through notes or electronic forms. Additionally, many times, SBDH data may not be asked due to patient condition or it might not be updated regularly. Providers and healthcare institutions should strive to collect SBDH data more regularly even if the data fields are not empty as SBDH status can change. These intake procedures should be present and not optional; currently, only language preference must be completed due to translation laws in place. Additionally, educating patients to utilize patient portals and update information via these portals can provide more current SBDH information. However, we should note that vulnerable populations would most likely not be the primary audience to utilize this feature, and this is the subpopulation that arguably needs more attention.

Limitations

Our study has numerous limitations. There were two distinct areas in our workflow that required manual attention: (1) EHR review and (2) labelling of features. Manual EHR review was performed to ensure that the notes contained social history information in a consistent location prior to widespread text extraction. We initially validated this with a random set of 10 patients, but later expanded our validation to 25 patients. We felt that having consistent results with the 25 patients indicated a high level of confidence. Manual labelling of features was time consuming and taxing. Although only one author performed the feature labelling, having multiple team members would provide better and possibly more consistent classification.

This approach, although we aim to create a generalizable workflow, is still stunted by local customizations due to unique nuances in note taking language. Patients can withhold information about their social challenges, making text classification harder to perform due to incorrect incoming data streams. Our approach relies on the fact that the patient has been seen within the healthcare system at some point in the past five years. This approach would not be applicable to those who are new to the institution or those who are not immediately identifiable. Classification levels for unstructured notes are not concrete as descriptive wording is also not concrete and can vary (e.g. “patient was a former smoker”, “patient quit last week”, “patient is an occasional smoker”, etc.). Structured data sources can add a more concrete sense to the classification. There were 5.7% copy-forward entries present as data collection of social factors may not always be appropriate (e.g. patient is inebriated, in an altered mental state, etc.). We did not incorporate outside ontologies, such as UMLS or MetaMap, as we were interested in creating a simple text classification approach that did not need to rely on outside entities. Furthermore, we believe that these ontologies would not have added a significant improvement in our approach due to the social factors (housing, alcohol, tobacco) that were investigated. Although minimized, applying NLP to clinical notes will always present limitations and risks with biased models, biased data, and data privacy.[35]

Community needs are constantly changing as the health of the community is not static. Currently, the King County CHNA has identified obesity, healthcare access, insurance status and drug use as other potential SBDH information to explore. These data types would be stored in different areas of the EHR and within different notes. It would be interesting to see if our designed workflow presented could be applicable and generalized to meet the needs of other SBDH data. Although we aimed to create a simplified framework to extract SBDH data from clinical notes, more complex methods such as convoluted neural networks and more advanced NLP part of speech tagging may be worth exploring as they may help improve accuracy and precision of the classification. As more notes become available for patients, it will also be important to keep in mind the potential bias of having more notes present from sicker patients and evaluating ways to reduce this bias.

We sourced data from solely one medical center. Patients might have had encounters or other visit types in neighboring hospitals and healthcare systems in the region. The lack of data sharing between institutions prevents holistic collection of SBDH data. Data completeness is vitally important to the quality and accuracy of models that are dependent on big data. Poor data quality and completeness lead to lower utilization and the lack of data can potentially lead to mistakes in the decision-making process; additionally, since there is no single or standardized source for SBDH data, the diversity of data and complexity of the associated data structures increase the difficulty and bottlenecks for data integration.[36] The lack of a standardized methodology to collect and store all SBDH data will limit the potential of this research field. Additionally, SBDH factors are constantly changing for patients as their behaviors can change depending on their circumstance. Being able to aggregate these data and create adaptable models is crucial as these features are never static. Furthermore, public health and outreach services fluctuate over time. Creating a method or utilizing an API to update the list of community shelters and other places for homeless services would be necessary to maintain an accurate understanding of a patients housing status.

V. CONCLUSION

From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status. Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use. Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social determinants and can supplement current structured sources to provide a more complete social history for patients. However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

Contributor statement:
AT performed the data extraction, tool building, and analysis. AB provided guidance and verification when needed.

Competing interests:
There are no competing interests.

Funding:
This work was supported by the U.S. Department of Health and Human Services, National Library of Medicine Training Grant T15LM007442.

Data sharing statement
The data used are unable to be shared due to patient privacy, confidentiality, and United States healthcare laws.

Ethics statement
This research is a part of a larger study that has been approved by the University of Washington IRB #STUDY00006723.

VI. REFERENCES

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*. 2014;2(1). doi:10.1186/2047-2501-2-3

2. Hood CM, Gennuso KP, Swain GR, Catlin BB. County Health Rankings. *American Journal of Preventive Medicine*. 2016;50(2):129-135. doi:10.1016/j.amepre.2015.08.024

3. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving Electronic Medical Records Upstream. *American Journal of Preventive Medicine*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

4. Social Determinants of Health. HealthyPeople.gov. Accessed February 1, 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

5. Social Determinants. Institute for Health Metrics and Evaluation. Accessed February 1, 2020. http://www.healthdata.org/social-determinants

6. Nerenz DR. Health Care Organizations' Use Of Race/Ethnicity Data To Address Quality Disparities. *Health Affairs*. 2005;24(2):409-416. doi:10.1377/hlthaff.24.2.409

7. Andermann A. Taking action on the social determinants of health in clinical practice: a framework for health professionals. *Canadian Medical Association Journal*. 2016;188(17-18):E474-E483. doi:10.1503/cmaj.160177

8. Olson DP, Oldfield BJ, Navarro SM. Standardizing Social Determinants Of Health Assessments. Published March 18, 2019. https://www.healthaffairs.org/do/10.1377/hblog20190311.823116/full/

9. Wockenfuss R, Frese T, Herrmann K, Claussnitzer M, Sandholzer H. Three- and four-digit ICD-10 is not a reliable classification system in primary care. *Scand J Prim Health Care*. 2009;27(3):131-136. doi:10.1080/02813430903072215

10. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40(5 Pt 2):1620-1639. doi:10.1111/j.1475-6773.2005.00444.x

11. Eligible Professional Meaningful Use Core Measures Measure 9 of 13. Published online May 2014. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/9_Record_Smoking_Status.pdf

12. Lax Y, Martinez M, Brown NM. Social Determinants of Health and Hospital Readmission. *Pediatrics*. 2017;140(5):e20171427. doi:10.1542/peds.2017-1427

13. King County Community Health Needs Assessment 2018/2019. Presented at the: https://www.kingcounty.gov/depts/health/data/community-health-indicators/~/media/depts/health/data/documents/2018-2019-Joint-CHNA-Report.ashx

14. Henry M, Mahathey A, Morrill T, et al. *The 2018 Annual Homeless Assessment Report (AHAR) to Congress*. The U.S. Department of Housing and Urban Development OFFICE OF COMMUNITY PLANNING AND DEVELOPMENT; 2018. https://files.hudexchange.info/resources/documents/2018-AHAR-Part-1.pdf

15. Stafford A, Wood L. Tackling Health Disparities for People Who Are Homeless? Start with Social Determinants. *International Journal of Environmental Research and Public Health*. 2017;14(12):1535. doi:10.3390/ijerph14121535

16. Ahmad S, Baig S, Taneja A, Nanchal R, Kumar G. The Outcomes of Severe Sepsis in Homeless. *Chest*. 2014;146(4):230A. doi:10.1378/chest.1995140

17. Bambra C, Gibson M, Sowden A, Wright K, Whitehead M, Petticrew M. Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. *Journal of Epidemiology & Community Health*. 2010;64(4):284-291. doi:10.1136/jech.2008.082743

18. Papadopoulou S, Hassapidou M, Katsiki N, et al. Relationships Between Alcohol Consumption, Smoking Status and Food Habits in Greek Adolescents. Vascular Implications for the Future. *Current Vascular Pharmacology*. 2017;15(2):167-173. doi:10.2174/1570161114666161024123357

19. Wong E. *Tobacco Use in King County*. Public Health Seattle & King County; 2012. https://www.kingcounty.gov/depts/health/data/~/media/depts/health/data/documents/tobacco-use-in-king-county-may-2012.ashx

20. Bogan S, Donohue B. King County drug and alcohol deaths rose 9.5% in 2018. https://newsroom.uw.edu/news/king-county-drug-and-alcohol-deaths-rose-95-2018

21. Drug-caused deaths in King County. Published online February 21, 2017. https://adai.washington.edu/WAdata/KingCountyDrugDeaths.htm

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. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513. doi:10.1136/jamia.2009.001560

24. Gundlapalli AV, Carter ME, Divita G, et al. Extracting Concepts Related to Homelessness from the Free Text of VA Electronic Medical Records. *AMIA Annu Symp Proc*. 2014;2014:589-598.

25. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE*. Published online 2017. doi:10.1371/journal.pone.0174708

26. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment: *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2018;77(2):160-166. doi:10.1097/QAI.0000000000001580

27. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying Patients with Significant Problems Related to Social Determinants of Health with Natural Language Processing. *Stud Health Technol Inform*. 2019;264:1456-1457. doi:10.3233/SHTI190482

28. Berg K, Doktorchik C, Quan H, Saini V. *Meaningful Information in the Age of Big Data: A Scoping Review on Social Determinants of Health Data Collection for Electronic Health Records*. In Review; 2019. doi:10.21203/rs.2.16433/v1

29. 2015 CDC HA-VTE PREVENTION CHALLENGE CHAMPION. https://www.cdc.gov/ncbddd/dvt/documents/champ-fact-sheet-harborview.pdf

30. Bulger EM, Kastl JG, Maier RV. The history of Harborview Medical Center and the Washington State Trauma System. *Trauma Surgery & Acute Care Open*. 2017;2(1):e000091. doi:10.1136/tsaco-2017-000091

31. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*. 2017;105:110-120. doi:10.1016/j.ijmedinf.2017.06.004

32. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*. 2019;19(S3). doi:10.1186/s12911-019-0781-4

33. Medicare & Medicaid EHR Incentive Program: Meaningful Use Stage 1 Requirements Overview. Presented at the: 2010. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/MU_Stage1_ReqOverview.pdf

34. Quality Measures and Tobacco Cessation. https://www.bhthechange.org/wp-content/uploads/2017/12/Quality-Measures-and-Tobacco-Cessation.pdf

35. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H. Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*. Published online June 4, 2020:161-168. doi:10.14745/ccdr.v46i06a02

36. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *CODATA*. 2015;14(0):2. doi:10.5334/dsj-2015-002

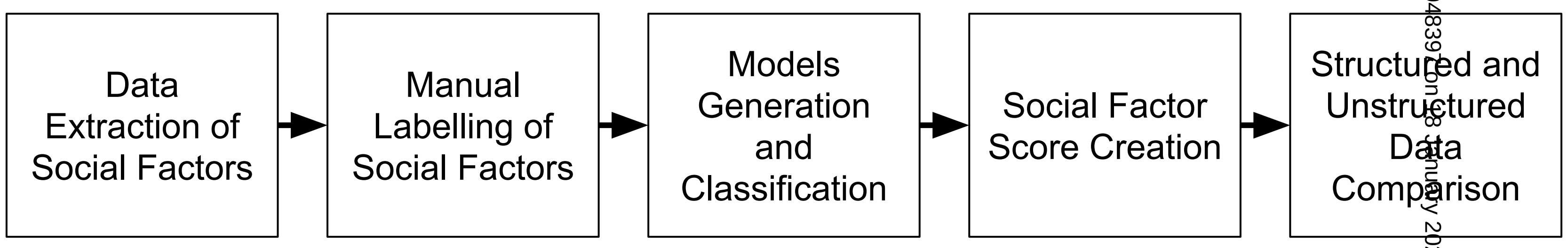
Acknowledgements:
Sally Lee
Abdelhak Abdou
Marion Granich
David Carlbom

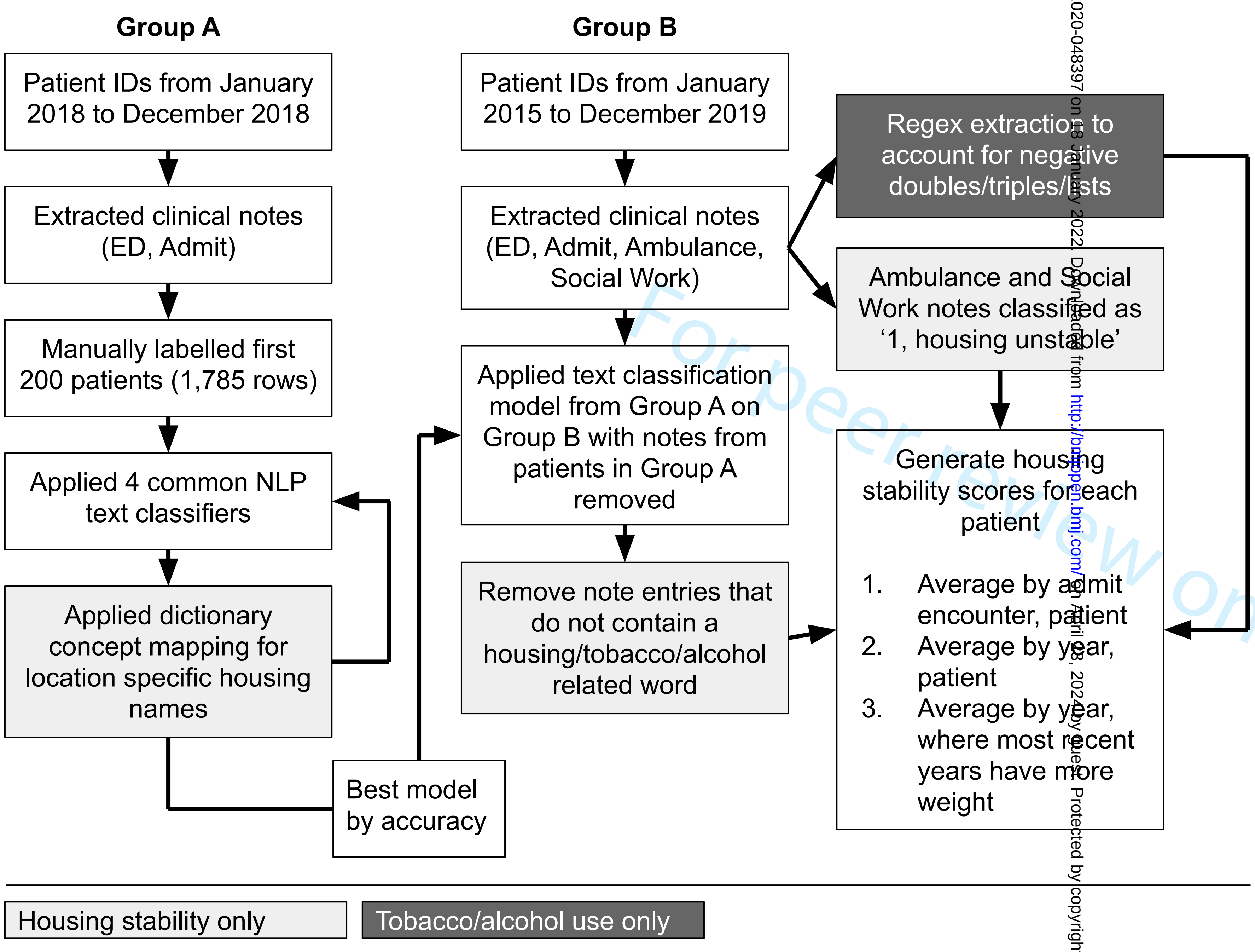
Figure legend:

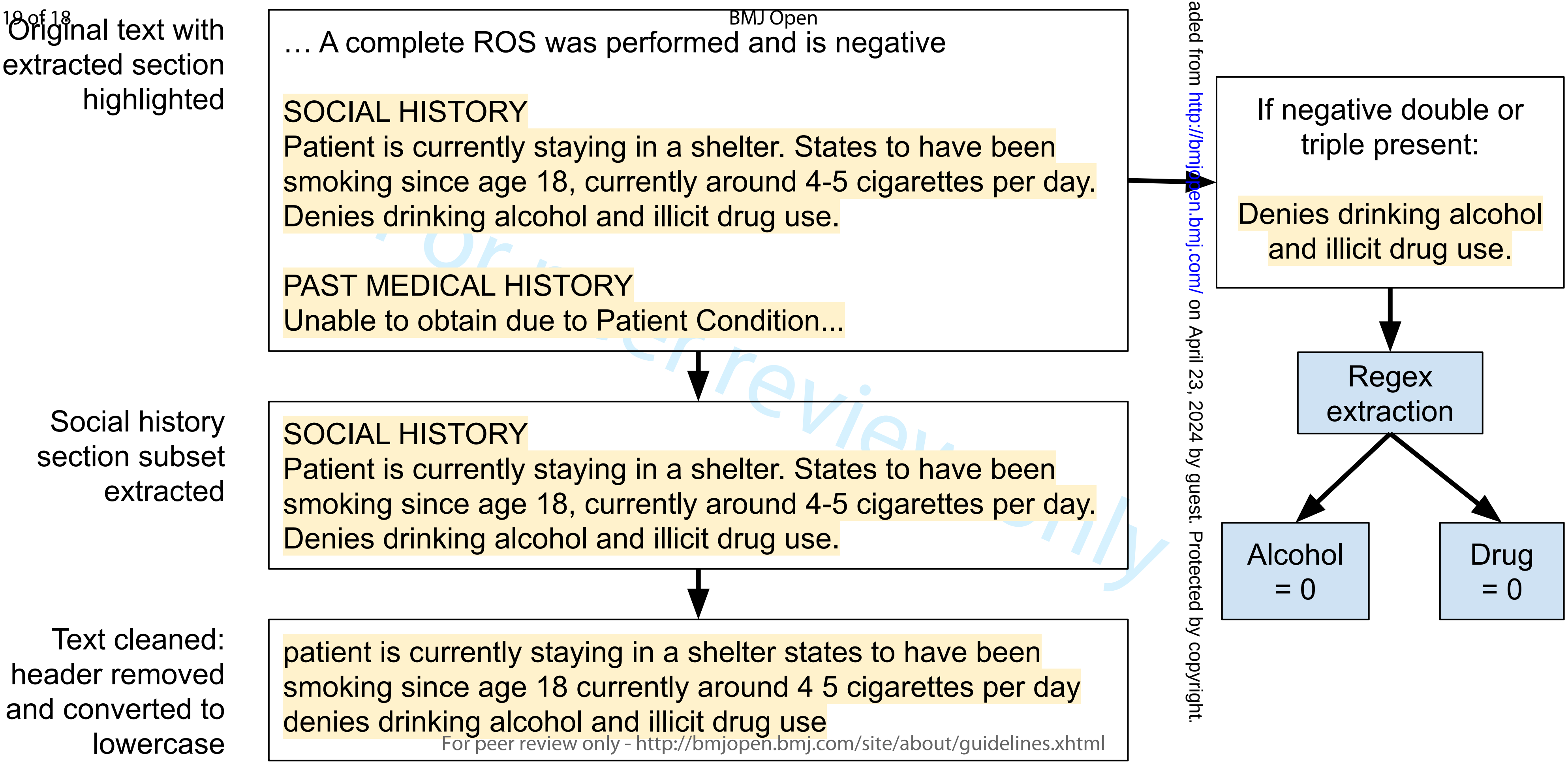
Figure 1: High-level overview of the workflow process

Figure 2: Text extraction, classification, and scoring workflow

Figure 3: Text extraction and cleaning process. Additional steps were performed for notes when classifying text related to tobacco and alcohol use to extract negative sentiment doubles or triples.


Data Extraction of Social Factors
Manual Labelling of Social Factors
Models Generation and Classification
Social Factor Score Creation
Structured and Unstructured Data Comparison


Group A
Patient IDs from January 2018 to December 2018
Extracted clinical notes (ED, Admit)
Manually labelled first 200 patients (1,785 rows)
Applied 4 common NLP text classifiers
Applied dictionary concept mapping for location specific housing names
Best model by accuracy
Group B
Patient IDs from January 2015 to December 2019
Extracted clinical notes (ED, Admit, Ambulance, Social Work)
Applied text classification model from Group A on Group B with notes from patients in Group A removed
Remove note entries that do not contain a housing/tobacco/alcohol related word
Regex extraction to account for negative doubles/triples/lists
Ambulance and Social Work notes classified as '1, housing unstable'
Generate housing stability scores for each patient
1. Average by admit encounter, patient
2. Average by year, patient
3. Average by year, where most recent years have more weight
Housing stability only
Tobacco/alcohol use only


Original text with extracted section highlighted
… A complete ROS was performed and is negative
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
PAST MEDICAL HISTORY
Unable to obtain due to Patient Condition...
Social history section subset extracted
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
Text cleaned: header removed and converted to lowercase
patient is currently staying in a shelter states to have been smoking since age 18 currently around 4 5 cigarettes per day denies drinking alcohol and illicit drug use
If negative double or triple present:
Denies drinking alcohol and illicit drug use.
Regex extraction
Alcohol = 0
Drug = 0

BMJ Open

A Simplified Data Science Approach to Extract Social and Behavioral Determinants: A Retrospective Cohort Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-048397.R3
Article Type:	Original research
Date Submitted by the Author:	12-Oct-2021
Complete List of Authors:	Teng, Andrew; University of Washington, Biomedical Informatics and Medical Education Wilcox, Adam ; University of Washington,
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Health informatics
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, SOCIAL MEDICINE, HISTORY (see Medical History), BIOTECHNOLOGY & BIOINFORMATICS

A Simplified Data Science Approach to Extract Social and Behavioral Determinants: A Retrospective Cohort Study

Andrew K. Teng[1], Adam B. Wilcox[1]

Affiliation:
[1] Biomedical Informatics and Medical Education
University of Washington
Seattle, WA, United States

Corresponding Author:
Andrew K. Teng
akteng@uw.edu
Biomedical Informatics and Medical Education, University of Washington
850 Republican Street, Box 358047
Seattle, WA 98195-0005, United States

Keywords: Text extraction, text classification, housing stability, social determinants, data science

Word Count: 4,985

Abstract

Objectives
We aim to extract a subset of social factors from clinical notes using common text classification methods.
Setting
We collaborated with a local Level I trauma hospital located in an underserved area that has a housing unstable patient population of about 6.5% and extracted text notes related to various social determinants for acute care patients.
Participants
Notes were retrospectively extracted from 43,798 acute care patients.
Methods
We solely utilize open source Python packages to test simple text classification methods that can potentially be easily generalizable and implemented. We extracted social history text from various sources, such as admission and emergency department notes, over a five-year timeframe and performed manual chart reviews to ensure data quality. We manually labelled the sentiment of the notes, treating each text entry independently. Four different models with two different feature selection methods (bag of words (BOW) and bigrams) were used to classify and predict housing stability, tobacco use, and alcohol use status for the extracted clinical text.
Results
From our analysis, we found overall positive results and metrics in applying open-source classification techniques; the accuracy scores were 91.2%, 84.7%, 82.8% for housing stability, tobacco use, and alcohol use respectively. There were many limitations in our analysis including social factors not present due to patient condition, multiple copy-forward entries and shorthand. Additionally, it was difficult to translate usage degrees for tobacco and alcohol use. However, when compared to structured data sources, our classification approach on unstructured notes yielded more results for housing and alcohol use; tobacco use proved less fruitful for unstructured notes.

Article Summary
Strengths and limitations of this study

- From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status.
- Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use.
- Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social and behavioral determinants and can supplement current structured sources to provide a more complete social history for patients.
- However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

I. INTRODUCTION

Most data can be generally categorized as structured or unstructured, where structured data can consist of items such as vital signs and lab results and unstructured data can consist of items such as text notes or images.[1] Although structured data can generally be easier to extract and analyze, unstructured data can potentially provide an array of information not present or easily identifiable in structured data. As healthcare institutions expand data collection to include non-clinical features, more unstructured data surrounding behavioral health and social determinants of health (SDoH) information, are starting to become more readily available. Furthermore, there has a been a growing interest around Medicaid patients, as SDoH can drive up to 80% of health outcomes, especially within this patient demographic.[2] Therefore, SDoH and REAL (Race, Ethnicity and Language) data are now being used for secondary analysis as recent research has indicated that there is a correlation between SDoH and health outcomes and the increasing need to research health disparities across populations.[3]

SDoH and REAL can include housing stability, access jobs and health care services, education level, language, and socioeconomic conditions.[4] These indicators are descriptors of different societies and are useful as predictors of health outcomes and the uptake of health interventions.[5] Because they can potentially be powerful indicators of health, many institutions are now starting to analyze and intake SDoH and REAL information, whether through text notes or standardized coding, such as International Classification of Diseases (ICD).[6] Additionally, SDoH can provide health teams with a greater understanding of a patient condition holistically.[7] However, there are challenges with SDoH intake as there is no standardized SDoH screening tool in the EHR across institutions[8]; additionally, coding schemes like ICD can prove to be unreliable in secondary analysis as coding can oversimplify symptoms and diagnoses leading to coding uncertainties and the fact that coding errors may be present from unintentional mistakes or even upcoding.[9,10] Additionally certain SDoH data may be more complete than others due to reimbursement incentives or other priorities.[11] Past research has shown that hospital readmissions are highly influenced by patient health status and SDoH and suggest that clinical staff and researchers should consider SDoH when assessing readmission risk.[12]

The 2018-2019 [redacted for review] Community Health Needs Assessment (CHNA) reported the results from a health needs assessment survey given to residents to identify regional perceived healthcare issues. It was determined that housing affordability and housing stability were major challenges dominating overall health.[13] Mental health was also highlighted as a challenge for healthcare providers; mental illness can be caused by depression, schizophrenia, and alcohol and substance-related disorders.[13] The CHNA reported that adults in the lowest income tier were about 15 times more likely to experience severe psychological distress compared to their high-income counterparts. Additionally, it noted that part of the region had continued challenges with adult smoking rates.[13] Locally, it is estimated that there are at least 22,000 homeless individuals in [redacted for review] and more than 12,000 people in the [redacted for review] region, a four percent increase over the previous year.[14] Housing instability is associated with various health inequalities, such as shorter life expectancy, higher morbidity, and increased usage of acute hospital services, "as the social determinants of homelessness and health inequities are often intertwined, and long term homelessness further exacerbates poor health".[15] It is therefore important to treat housing stability and other SDoH as a combined health issue to aid in improving health outcomes in clinical settings. Although some research has shown that patients who experience housing instability are more likely to die following admission for severe sepsis than those with insurance,[16] other research indicates that the effects of health inequalities are still unclear and need further investigation.[17] Additionally, various behavioral habits, including tobacco and alcohol use, although may not directly be considered a SDoH, can impact health decisions and outcomes. For example, one study found that participants who drank alcohol and reported tobacco use consumed more foods higher in fat and sugar, low in vitamins and minerals as well as foods, considered by them to be less healthy and prepared in a less healthy way.[18]

Within our region, it has been noted in recent years that the smoking rate is around 13 percent; however, among Black/African-Americans or individuals with multiple races, is double the rate among white adults and four times higher than Asian adults. Additionally, it was reported that, when compared to high income households, low income households were three times more likely to be smokers.[13,19] Drug and alcohol use also shared similar metrics; within the region, "drug and alcohol-caused deaths was 22% higher among Blacks and four times greater among American Indian/Alaskan Native than among non-Hispanic Whites" and alcohol use represented 4.97 per 100,000 deaths locally in 2015.[20,21] Therefore it may be important to look at social determinants and health behaviors, together known as social and behavioral determinants of health (SBDH) to better understand the patient population.[18]

Recent technological advances in machine learning and artificial intelligence have shown great potential in providing a pathway for informaticians and clinicians to better understand unstructured data.
Within the clinical setting, there have been numerous approaches in adopting natural language processing (NLP) to aid with processing unstructured clinical text notes. Common uses of NLP include

extracting diagnoses and chief complaints as well as grouping of information for quality improvement. There are various NLP methods that can be used in the clinical setting, such as automatic tagging of conditions or variables of interest, sentiment classification, or even text extraction. Various open source NLP and ontological tools, such as Automated Retrieval Console, Apache clinical Text Analysis and Knowledge Extraction System (Apache cTAKES), MetaMap, and HITEx, Unified Medical Language System (UMLS) Metathesaurus and BioPortal have been used to aid with text extraction or classification.[22–24] On the other hand less complex classification methods have been used as well to identify specific groups of patients, risk assessment, or aid in validating structured annotation.[25,26,27] A recent scoping review found that although practitioners collect a variety of SBDH data at point of care through EHR, the overall use of automated technology is limited to date.[28]

With the idea of implementing an easily generalizable approach to classify selected social factors, we extracted both unstructured and structured data sources related to SBDH from a local hospital to identify and generate a framework to automatically extract and classify SBDH from text notes. We focused on housing stability status, tobacco use, and alcohol use. These three social factors were chosen due to their direct impact on health outcomes and the local public health impact[14–18] and presence in the EHR. To tackle challenges associated with SBDH extraction from unstructured text notes, we aimed to create a generalizable framework using low barrier open-source tools that are commonly used in the data science field. Because notes and stylistic choices can be institution and location specific, we sought not to create a model that is generalizable but rather a simplified method that could be potentially easily implemented using common off the shelf NLP and data science tools.

II. METHODS

Study Design and Overview

A high-level overview of our workflow can be seen in Figure 1. We retrospectively extracted patient data from the acute care setting at a Level I trauma center and academic teaching hospital with the aim to create a general and easily applicable workflow to extract and classify SBDH factors from clinical notes. We applied a two-pronged approach and collected unstructured data from a subset of patients over a 1-year timespan (Group A) to create and test the text classification model and also collected structured and unstructured data from a subset of patients over a 5-year timespan (Group B) to apply the best model created from Group A and compare results between the two data types. We performed automatic classification and scoring of patients via various NLP classification methods on three social factors: (1) housing stability, (2) tobacco use, and (3) alcohol use. Our general workflow for housing stability, a similar approach was also used for tobacco and alcohol use, can be seen in Figure 2.

Study Population

Data were extracted from [redacted for review], a 413-bed academic hospital that has a patient population consisting mostly from Washington, but also from a five-state area.[29] In 2014, there were 17,121 inpatient admissions, where 19 percent of the patients belong to a racial or ethnic minority and 37 percent of patients were enrolled in Medicaid.[29,30] Additionally, in 2015, the non-US born population was estimated to be around 21 percent in [redacted for review], highlighting the potential diversity that could be found with this patient population.[30]

Data Sources, Extraction, and Validation

We extracted both structured and unstructured data sources related to housing stability, tobacco use, and alcohol use using SQL queries called directly from an integrated python-based Jupyter Notebook:

a. Structured data sources include billing and diagnostic/International Classification of Disease (ICD) 9 and 10 codes, questionnaire or Epic SmartForm responses, address fields (location), problem list (ICD 9), patient encounters, clinical events (actual encounters of care), and discharge/disposition location.
b. Unstructured data sources consisted of text notes from the emergency department (ED), admission (admit) notes, social work, and ambulance notes.

Discharge notes were not explored as they were not recorded in the same subdivided format as the admit and ED notes, making selective text extraction of SBDH difficult. From our initial list of patient identifiers over a one-year timespan from Group A, we performed manual EHR validation of a random subset of 50 patients to validate the completeness of the clinical notes and confirm the location of social history and social factors in clinical notes. Extensive research and conversations with an internal data analyst confirmed the location of these topics (housing, tobacco use, and alcohol use) within structured data sources.

Data Cleaning

After confirmation, clinical notes were extracted for both Groups A and B. The notes were cleaned (e.g. symbols removed, converted to lowercase) prior to classification and analysis in the Python Jupyter notebook via NLTK. Our general text extraction and cleaning workflow can be seen in Figure 3. However, housing stability notes and tobacco or alcohol use notes were stylistically and grammatically different, and both sets needed distinct additional cleaning steps. Housing stability notes that contained the phrase 'not homeless' were converted via regex to say 'housed' instead. Additionally, for housing stability, a concept dictionary was also created to substitute local facility names with more general concept (e.g. 'Union Gospel Mission' was converted to 'shelter'). This was done to explore how the algorithms handle formal nouns.

For text notes in Group B, we performed an additional concept extraction step. Tobacco use and alcohol use notes often contained incomplete (lacking the subject, predicate, object format) triples or doubles (e.g. 'Denies smoking, drinking, drugs'). Due to their incomplete sentence structures, common NLP tools to parse, extract, and classify triples, such as Stanford CoreNLP, were not suitable as these tools rely on having all three parts of the triple present. These notes related to tobacco and alcohol use therefore underwent an additional step that performed a separate relation extraction that would first identify a negative sentiment word (e.g. denies), then individually extract the following SBDH related objects in the list by commas or conjunctions (e.g. and, or), and then label, or reclassify if necessary, the negative sentiment to all components of the list. Our process can be seen in the left side of Figure 3. If the regex extraction of negative lists resulted in a different result from the text classification prediction, the regex extraction would overwrite the end result prior to scoring. Once these steps were performed, the data were considered clean and suitable for classification.

Model building

Cleaned text from Group A were used to generate and test the classification models. These notes were split in 70/30 validation and testing sets. We applied four different common NLP text classification models to the testing sets (via SciKit Learn): multinomial naïve Bayes, support vector machine, logistic regression, and random forest. Default parameters and a bag-of-words approach were used. The best performing model by accuracy was then chosen and applied to the larger corpus, Group B, with notes from patients in Group A removed, to avoid overfitting and classification bias. This process was performed for housing, tobacco use, and alcohol use.

Scoring generation

In order to create a simple method of identifying patients who are experiencing social instability, we created a scoring metric based on the classified notes. After applying the optimum model by accuracy to the entire corpus of extracted text notes, housing stability, tobacco use, and alcohol use scores were generated. Patient identifiers were mapped by patient location and those who were not in the acute care setting during this timeframe were removed. Three different scoring approaches were used to describe these social factors: (1) predictions were averaged by patient encounter, then averaged by patient identifier, (2) predictions were averaged by year, then by patient identifier, and (3) predictions were averaged by year, where each year then had a weight where the most recent year had the highest weight and the furthest year had the lowest weight (e.g. predictions from 2019 were weighted by a factor of 5 and predictions from 2015 were weighted by a factor of 1). This scoring generation process was then repeated on our structured data for all three social factors and the results were compared and analyzed. Structured data was also extracted for our list of patients in Group B.

Patient and Public Involvement

No patients were involved. The retrospective exploration is a part of a larger study and was approved by the [redacted for review] Institutional Review Board #STUDY00006723. Patient data elements, including encounter identifiers, race, age, and notes with SBDH, were extracted directly from the data warehouse and stored on encrypted computers and were not distributed or shared outside of the secured and closed environment. No patient identifiers or names were stored in this analysis.

III. RESULTS

Characteristics of study subjects

Clinical notes (ED, admit, social work, and ambulance) between 2015 and 2019 were extracted and included, forming Group B. Notes from the first 200 patients were included in Group A and notes from 147,457 patients were included in Group B. During the same timeframe, 61,767 patients were in acute care. After extraction and model prediction, the patient notes were cross referenced with inpatient location and only notes from those who were in acute care were retained, for a total of 43,798 patients from 2015 to 2019. The patient demographics of this final subset were 63% (*n*=27,575) male, 37% (*n*=16,223) female, 88.2% (*n*=38,634) not Hispanic or Latino, and 10.5% (*n*=4,609) Hispanic or Latino, and 1.3% (*n*=555) unknown or not answered. Further descriptive statistics can be found in Table 1.

Table 1: Population demographics

Race (*n*=43,798)	*n* (%)
White or Caucasian	31,575 (72.1%)
Black or African American	4,812 (11.0%)
Asian	3,174 (7.2%)
American Indian or Alaska Native	1,165 (2.7%)
Native Hawaiian or other Pacific Islander	524 (1.2%)
Multiple races	3 (0%)
Unavailable, unknown, or missing	2,545 (5.8%)
Age range (*n*=43,798)	***n* (%)**
0-18	1,856 (4.2%)
19-44	12,437 (28.4%)
45-64	14,863 (33.9%)
65-84	11,902 (27.2%)
85 and over	2,740 (6.3%)

Data attributes

Table 2 illustrates the amount of data for each corresponding extraction level, specifically for housing status. We first started with extracting text from the ED and admit notes, forming Group A, which consisted of 50,000 rows or text entries and covered 3,200 unique patients, over a one-year timeframe. From there, we manually labelled housing stability concepts in a binary fashion, where 0 would indicate housing stability and 1 would indicate any level of housing instability, regardless of severity. As manual labelling can be a labor-intensive process, only the first 6,000 text rows were labelled, covering 218 unique patients. However, within these first 6,000 rows, numerous notes did not contain text that alluded to housing status or were empty due to patient condition. Therefore, only 1,785 out of the 6,000 rows were labelled, covering 200 unique patients, where 995 (55.7%) were labelled as housing stable and 790 (44.3%) were labelled as housing unstable. We also found that 5.7% of the entries within this subset were duplicates or copy-forward entries. The same workflow was performed for labelling tobacco and alcohol use. However, only 1,108 rows were labelled for tobacco use and 1,220 rows for alcohol use, where in both cases 0 indicated no use, 1 indicated rare/previous/occasional use, and 2 indicated current use, regardless of degree. Tobacco use resulted in 446 (40.3%) labels for no use, 129 (11.6%) labels for rare/previous/occasional use, and 533 (48.1%) labels for current use. Similarly, alcohol use resulted in 595 (48.8%) labels for no use, 185 (15.2%) labels for rare/previous/occasional use, and 440 (36%) labels for current use.

Table 2: Extracted data amounts for housing status

Level of extraction	Rows (*n*)	Unique patients (*n*)	Unique encounters (*n*)	Social history entries (*n*/unique)
ED and Admit notes	49,955	3,233	15,664	21,876/21,334
Housing, Tobacco, Alcohol Information	6,000	218	1,995	2,408/2,211
Remove nulls/missing data	Housing: 1,785 Tobacco: 1,108 Alcohol: 1,220	Housing: 200 Tobacco: 179 Alcohol: 181	1,361	1,785/1,684

Model performance

Four different common text classifiers, mentioned in the Methods section, were applied to the manually labelled Group A data. The statistical metrics, including accuracy, precision, and recall, can be seen in Table 3 and 4. The accuracies between the classifiers and each classification technique for housing stability were overall fairly high ranging from 84.36-92.18%. The accuracies for tobacco and alcohol use were lower, ranging from 70.87-84.68% for tobacco use and 69.95-82.79% for alcohol use. Additionally, for each top performing model, the most influential words for text classification, for each social factor, can be seen in Table 5. The best performing classification models were selected for each social factor and were used to apply the model to our entire corpus in Group B.

Table 3: Accuracies amongst text classifiers

	n=1	n=1-2
Multinomial naïve Bayes	Housing: 91.62% Tobacco: 70.87% Alcohol: 70.77%	Housing: 91.43% Tobacco: 77.18% Alcohol: 69.95%
Support vector machine	Housing: **92.18%** Tobacco: 81.08% Alcohol: 76.50%	Housing: 91.99% Tobacco: 82.88% Alcohol: 81.97%
Logistic regression	Housing: 84.36% Tobacco: 75.38% Alcohol: 77.60%	Housing: 90.13% Tobacco: **84.68%** Alcohol: **82.79%**
Random forest	Housing: 90.50% Tobacco: 76.28% Alcohol: 71.31%	Housing: 91.25% Tobacco: 78.98% Alcohol: 75.68%

Table 4: Best performing classifier detailed metrics

	Classifier	Accuracy	Recall	Precision	F1
Housing status*	Support vector machine (n=1)	0.92	0.93/0.91 (0/1)	0.94/0.90	0.93/0.91
Tobacco use**	Logistic Regression (n=1-2)	0.85	0.82/0.95/0.86 (0,1,2)	0.96/0.43/0.87 (0,1,2)	0.88/0.60/0.87 (0,1,2)
Alcohol use**	Logistic Regression (n=1-2)	0.83	0.86/0.73/0.81 (0,1,2)	0.93/0.44/0.88 (0,1,2)	0.89/0.55/0.84 (0,1,2)

* 0: no use, 1: current use
** 0: no use, 1: rare/occasional/history, 2: current use

Table 5: Word or phrase importance ranking

Social factor (Classifier)	Top 20 weighted words
Housing stability (support vector machine, n=1)	`['friends' 'motel' 'stay' 'cigs' 'found' 'street' 'stays' 'streets' 'van'`

	'incarcerated' 'desc' 'currently' 'undomiciled' 'friend' 'respite' 'kcj' 'shelters' 'homelessness' 'shelter' 'homeless']
No tobacco use (logistic regression, n=1,2)	['use denies' 'deneis' 'lives' 'tobacco drug' 'seattle denies' 'use results' 'lives seattle' 'alcohol tobacco' 'tobacco drugs' 'never smoker' 'etoh tobacco' 'drinking' 'seattle tobacco' 'denies cigarettes' 'drugs tobacco' 'denies alcohol' 'tobacco alcohol' 'denies smoking' 'denies' 'denies tobacco']
No alcohol use (logistic regression, n=1,2)	['care' 'ppd' 'tobacco' 'smoking' 'etoh tobacco' 'history cocaine' 'tobacco alcohol' 'etoh illicit' 'alcohol tobacco' 'etoh drug' 'drugs etoh' 'alcohol drug' 'use none' 'alcohol drugs' 'drug etoh' 'denies alcohol' 'lives' 'denies drug' 'denies etoh' 'denies']

Scoring results and comparison

After classifying text for housing stability, tobacco use, and alcohol use for patients in Group B, we applied a scoring metric scheme, described in the Methods section. We generated three different scores that were calculated and weighted differently based on time. Our final score weighs more recent note entries and their resulting classification score higher than notes from previous years as social factors and their influence can change over time. Using the same process, we extracted and scored housing stability, tobacco use, and alcohol use with structured data sources and compared the results with the unstructured process.

I. Housing stability

Using notes, we classified 839 patients as housing unstable, a score above 0.5, and 21,370 patients as housing stable, a score of 0.5 and below. In total, we classified 22,209 patients with this text classification workflow, which covered 50.71% of the acute care patients within the same timeframe. When compared with structured data sources, only 791 (1.81%) additional patients were found.

II. Tobacco use

We classified 4,911 patients as currently using tobacco, regardless of amount or degree (1.5-2) using text notes. We classified 1,480 patients as having rare/occasional/past use of tobacco (0.5-1.5), and 7,139 patients as not using tobacco (0-0.5). In total, we classified 13,530 patients with this text classification workflow, which covered 30.9% of the acute care patients within the same timeframe. When compared with structured data sources, 17,9351 (40.9%) additional patients were captured.

III. Alcohol use

We classified 2,738 patients as currently using alcohol, regardless of amount or degree (1.5-2) using text notes. We classified 4,050 patients as having rare/occasional/past use of alcohol (0.5-1.5), and 13,885 patients as not drinking alcohol (0-0.5). In total, we classified 20,673 patients with this text classification workflow, which covered 37% of the acute care patients within the same timeframe. When compared with structured data sources, no additional patients were found.

IV. DISCUSSION

Our approach to a simple text classification method for various social determinants of health have shown positive results. The selected classification models were chosen as they were the most commonly used classification models when researching text classification techniques. Furthermore, these models were robust enough to curtail the need for more complex machine learning based text classification methods, which may be harder to interpret in the clinical space as the weights and decisions can be confiscated due to the black box nature of these more complex classification methods. Generally, linear models are fast to train, can work well with sparse data, and offer interpretability.[31] Additionally, recent research has also suggested that more complex machine learning approaches may not yield statistically significant improvements in predictive power to justify the time and effort necessary to implement and test these more complex methods. Although promising, more advanced methods of NLP, such as convoluted neural networks, may not provide a significant tradeoff in improvement or accuracy versus transparent understanding of rule-based approaches. In fact, Yao et al. found that the F1 scores for CNN via TensorFlow did not improve significantly for interested features when compared to logistic regression and support vector machine implementations.[32] Finally, generalizable methods to create institution-specific models can be better for the healthcare system as a whole as each institution records clinical information with variances.

Although SBDH information and other social factors can be indicative of overall health, collection of SBDH heavily relies on clinical staff to screen and document SBDH. Furthermore, it also assumes that patients will respond accurately and truthfully. Various financial incentives from the federal level have propelled collection of social factors, such as tobacco use and tobacco cessation. However, other social factors, which can be equally as important, such as alcohol use are not incentivized to be captured; rather only more severe instances are incentivized, such as alcohol dependence or alcohol addiction or disorder.[33,11] Due to this discrepancy, we found that structured data sources were less reliable, and that text classification aided in detailing a patient more holistically.

Our text classification of unstructured data relied solely on ED, admit, social work, and ambulatory notes as our parsing and extraction method could only work with notes in a certain format with the social history heading. Social factors and other social history could also be recorded in other locations, but were not compatible with our approach. Furthermore, social work and ambulatory notes used for housing status only and were only extracted if the notes contained a word or phrase related to housing instability. This approach was used as the notes were typically stored in a more unstructured format compared to the ED and admit notes; there were no section headers. The lack of section headers increased the difficulty to extract the notes and the notes would often verbiage that would interfere with the simple text classification approach that we used. Therefore, we decided to extract notes that contained words relating to housing instability. Additionally, tobacco and alcohol use notes had stylistic and grammatical challenges. These social factors were often grouped together in incomplete triples (e.g. "denies drinking, smoking, illicit drug use"). The classification algorithms often had trouble reciprocating the negative connotation to all components of the triple. Therefore, we used regex to specifically extract these triples and classify the note based on the presence of words related to tobacco or alcohol. Without this additional data cleaning or manipulation step, the negative sentiment in a list would not have been applied to all elements within the list, but rather only the first element. In our example of 'denies smoking, drinking, drugs', the negative sentiment of 'denies' would have only been applied to smoking as smoking immediately follows 'denies'. However, with our additional concept extraction step, the negative sentiment of 'denies' is now also applied to 'drinking' and 'drugs'. These results would then override the text classification algorithm, if there was a discrepancy. Therefore, the scoring metrics for these cases would not necessarily reflect the accuracy or performance of our scoring method.

It was interesting to find that tobacco use was recorded significantly more often in structured data sources compared to alcohol use and housing stability. However, because tobacco use is a (Centers for Medicare and Medicare Services) CMS core quality measure, it can be expected that this feature is more available in structured form as it is often directly asked to the patient on intake forms, screeners, or during cessation treatment.[11] Furthermore, the Joint Commission created the Tobacco Performance Measure Set, which are three standardized performance measures addressing tobacco screening and cessation counseling: (1) Tobacco use screening of patients 18 years and over, (2) Tobacco use treatment, including counseling and medication during hospitalization, and (3) Tobacco use treatment management plan at discharge. CMS began using these performance measures in 2016.[34] Because alcohol consumption is not a recommended CMS core quality measure for adults, the amount of data regarding alcohol use is not complete in structured form as it may not be consistently collected during intake procedures.

Past research has consistently pointed towards SBDH impacting patient health and outcomes. However, collection of SBDH can be a major limiting factor in the ability to model and integrate these data. There has not been a standardized collection process for SBDH data across the institution, whether it is recorded through notes or electronic forms. Additionally, many times, SBDH data may not be asked due to patient condition or it might not be updated regularly. Providers and healthcare institutions should strive to collect SBDH data more regularly even if the data fields are not empty as SBDH status can change. These intake procedures should be present and not optional; currently, only language preference must be completed due to translation laws in place. Additionally, educating patients to utilize patient portals and update information via these portals can provide more current SBDH information. However, we should note that vulnerable populations would most likely not be the primary audience to utilize this feature, and this is the subpopulation that arguably needs more attention.

Limitations
Our study has numerous limitations. There were two distinct areas in our workflow that required manual attention: (1) EHR review and (2) labelling of features. Manual EHR review was performed to ensure that the notes contained social history information in a consistent location prior to widespread text extraction. We initially validated this with a random set of 10 patients, but later expanded our validation to 25 patients. We felt that having consistent results with the 25 patients indicated a high level of confidence. Manual labelling of features was time consuming and taxing. Although only one author performed the feature labelling, having multiple team members would provide better and possibly more consistent classification.

This approach, although we aim to create a generalizable workflow, is still stunted by local customizations due to unique nuances in note taking language. Patients can withhold information about their social challenges, making text classification harder to perform due to incorrect incoming data streams. Our approach relies on the fact that the patient has been seen within the healthcare system at some point in the past five years. This approach would not be applicable to those who are new to the institution or those who are not immediately identifiable. Classification levels for unstructured notes are not concrete as descriptive wording is also not concrete and can vary (e.g. “patient was a former smoker”, “patient quit last week”, “patient is an occasional smoker”, etc.). Structured data sources can add a more concrete sense to the classification. There were 5.7% copy-forward entries present as data collection of social factors may not always be appropriate (e.g. patient is inebriated, in an altered mental state, etc.). We did not incorporate outside ontologies, such as UMLS or MetaMap, as we were interested in creating a simple text classification approach that did not need to rely on outside entities. Furthermore, we believe that these ontologies would not have added a significant improvement in our approach due to the social factors (housing, alcohol, tobacco) that were investigated. Although minimized, applying NLP to clinical notes will always present limitations and risks with biased models, biased data, and data privacy.[35]

Community needs are constantly changing as the health of the community is not static. Currently, the King County CHNA has identified obesity, healthcare access, insurance status and drug use as other potential SBDH information to explore. These data types would be stored in different areas of the EHR and within different notes. It would be interesting to see if our designed workflow presented could be applicable and generalized to meet the needs of other SBDH data. Although we aimed to create a simplified framework to extract SBDH data from clinical notes, more complex methods such as convoluted neural networks and more advanced NLP part of speech tagging may be worth exploring as they may help improve accuracy and precision of the classification. As more notes become available for patients, it will also be important to keep in mind the potential bias of having more notes present from sicker patients and evaluating ways to reduce this bias.

We sourced data from solely one medical center. Patients might have had encounters or other visit types in neighboring hospitals and healthcare systems in the region. The lack of data sharing between institutions prevents holistic collection of SBDH data. Data completeness is vitally important to the quality and accuracy of models that are dependent on big data. Poor data quality and completeness lead to lower utilization and the lack of data can potentially lead to mistakes in the decision-making process; additionally, since there is no single or standardized source for SBDH data, the diversity of data and complexity of the associated data structures increase the difficulty and bottlenecks for data integration.[36] The lack of a standardized methodology to collect and store all SBDH data will limit the potential of this research field. Additionally, SBDH factors are constantly changing for patients as their behaviors can change depending on their circumstance. Being able to aggregate these data and create adaptable models is crucial as these features are never static. Furthermore, public health and outreach services fluctuate over time. Creating a method or utilizing an API to update the list of community shelters and other places for homeless services would be necessary to maintain an accurate understanding of a patients housing status.

V. CONCLUSION

From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status. Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use. Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social determinants and can supplement current structured sources to provide a more complete social history for patients. However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

Contributor statement:
AT performed the data extraction, tool building, and analysis. AB provided guidance and verification when needed.

Competing interests:
There are no competing interests.

Funding:
This work was supported by the U.S. Department of Health and Human Services, National Library of Medicine Training Grant T15LM007442.

Data sharing statement
The data used are unable to be shared due to patient privacy, confidentiality, and United States healthcare laws.

Ethics statement
This research is a part of a larger study that has been approved by the University of Washington IRB #STUDY00006723.

VI. REFERENCES

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*. 2014;2(1). doi:10.1186/2047-2501-2-3

2. Hood CM, Gennuso KP, Swain GR, Catlin BB. County Health Rankings. *American Journal of Preventive Medicine*. 2016;50(2):129-135. doi:10.1016/j.amepre.2015.08.024

3. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving Electronic Medical Records Upstream. *American Journal of Preventive Medicine*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

4. Social Determinants of Health. HealthyPeople.gov. Accessed February 1, 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

5. Social Determinants. Institute for Health Metrics and Evaluation. Accessed February 1, 2020. http://www.healthdata.org/social-determinants

6. Nerenz DR. Health Care Organizations' Use Of Race/Ethnicity Data To Address Quality Disparities. *Health Affairs*. 2005;24(2):409-416. doi:10.1377/hlthaff.24.2.409

7. Andermann A. Taking action on the social determinants of health in clinical practice: a framework for health professionals. *Canadian Medical Association Journal*. 2016;188(17-18):E474-E483. doi:10.1503/cmaj.160177

8. Olson DP, Oldfield BJ, Navarro SM. Standardizing Social Determinants Of Health Assessments. Published March 18, 2019. https://www.healthaffairs.org/do/10.1377/hblog20190311.823116/full/

9. Wockenfuss R, Frese T, Herrmann K, Claussnitzer M, Sandholzer H. Three- and four-digit ICD-10 is not a reliable classification system in primary care. *Scand J Prim Health Care*. 2009;27(3):131-136. doi:10.1080/02813430903072215

10. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40(5 Pt 2):1620-1639. doi:10.1111/j.1475-6773.2005.00444.x

11. Eligible Professional Meaningful Use Core Measures Measure 9 of 13. Published online May 2014. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/9_Record_Smoking_Status.pdf

12. Lax Y, Martinez M, Brown NM. Social Determinants of Health and Hospital Readmission. *Pediatrics*. 2017;140(5):e20171427. doi:10.1542/peds.2017-1427

13. King County Community Health Needs Assessment 2018/2019. Presented at the: https://www.kingcounty.gov/depts/health/data/community-health-indicators/~/media/depts/health/data/documents/2018-2019-Joint-CHNA-Report.ashx

14. Henry M, Mahathey A, Morrill T, et al. *The 2018 Annual Homeless Assessment Report (AHAR) to Congress*. The U.S. Department of Housing and Urban Development OFFICE OF COMMUNITY PLANNING AND DEVELOPMENT; 2018. https://files.hudexchange.info/resources/documents/2018-AHAR-Part-1.pdf

15. Stafford A, Wood L. Tackling Health Disparities for People Who Are Homeless? Start with Social Determinants. *International Journal of Environmental Research and Public Health*. 2017;14(12):1535. doi:10.3390/ijerph14121535

16. Ahmad S, Baig S, Taneja A, Nanchal R, Kumar G. The Outcomes of Severe Sepsis in Homeless. *Chest*. 2014;146(4):230A. doi:10.1378/chest.1995140

17. Bambra C, Gibson M, Sowden A, Wright K, Whitehead M, Petticrew M. Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. *Journal of Epidemiology & Community Health*. 2010;64(4):284-291. doi:10.1136/jech.2008.082743

18. Papadopoulou S, Hassapidou M, Katsiki N, et al. Relationships Between Alcohol Consumption, Smoking Status and Food Habits in Greek Adolescents. Vascular Implications for the Future. *Current Vascular Pharmacology*. 2017;15(2):167-173. doi:10.2174/1570161114666161024123357

19. Wong E. *Tobacco Use in King County*. Public Health Seattle & King County; 2012. https://www.kingcounty.gov/depts/health/data/~/media/depts/health/data/documents/tobacco-use-in-king-county-may-2012.ashx

20. Bogan S, Donohue B. King County drug and alcohol deaths rose 9.5% in 2018. https://newsroom.uw.edu/news/king-county-drug-and-alcohol-deaths-rose-95-2018

21. Drug-caused deaths in King County. Published online February 21, 2017. https://adai.washington.edu/WAdata/KingCountyDrugDeaths.htm

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. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513. doi:10.1136/jamia.2009.001560

24. Gundlapalli AV, Carter ME, Divita G, et al. Extracting Concepts Related to Homelessness from the Free Text of VA Electronic Medical Records. *AMIA Annu Symp Proc*. 2014;2014:589-598.

25. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE*. Published online 2017. doi:10.1371/journal.pone.0174708

26. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment: *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2018;77(2):160-166. doi:10.1097/QAI.0000000000001580

27. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying Patients with Significant Problems Related to Social Determinants of Health with Natural Language Processing. *Stud Health Technol Inform*. 2019;264:1456-1457. doi:10.3233/SHTI190482

28. Berg K, Doktorchik C, Quan H, Saini V. *Meaningful Information in the Age of Big Data: A Scoping Review on Social Determinants of Health Data Collection for Electronic Health Records*. In Review; 2019. doi:10.21203/rs.2.16433/v1

29. 2015 CDC HA-VTE PREVENTION CHALLENGE CHAMPION. https://www.cdc.gov/ncbddd/dvt/documents/champ-fact-sheet-harborview.pdf

30. Bulger EM, Kastl JG, Maier RV. The history of Harborview Medical Center and the Washington State Trauma System. *Trauma Surgery & Acute Care Open*. 2017;2(1):e000091. doi:10.1136/tsaco-2017-000091

31. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*. 2017;105:110-120. doi:10.1016/j.ijmedinf.2017.06.004

32. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*. 2019;19(S3). doi:10.1186/s12911-019-0781-4

33. Medicare & Medicaid EHR Incentive Program: Meaningful Use Stage 1 Requirements Overview. Presented at the: 2010. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/MU_Stage1_ReqOverview.pdf

34. Quality Measures and Tobacco Cessation. https://www.bhthechange.org/wp-content/uploads/2017/12/Quality-Measures-and-Tobacco-Cessation.pdf

35. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H. Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*. Published online June 4, 2020:161-168. doi:10.14745/ccdr.v46i06a02

36. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *CODATA*. 2015;14(0):2. doi:10.5334/dsj-2015-002

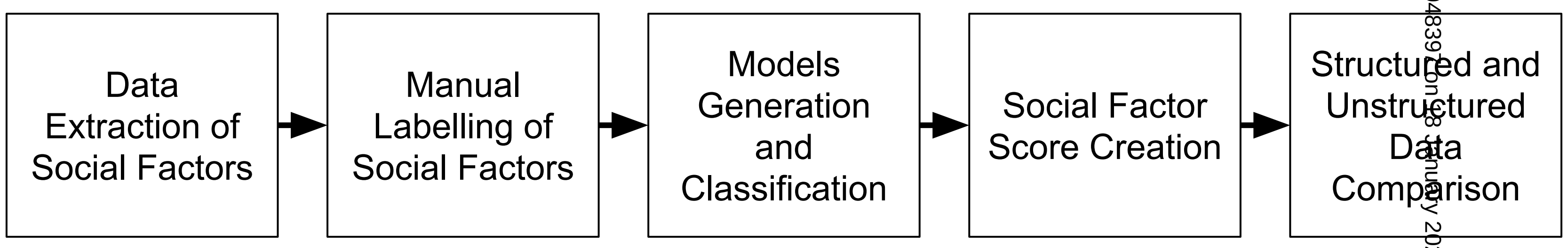
Acknowledgements:
Sally Lee
Abdelhak Abdou
Marion Granich
David Carlbom

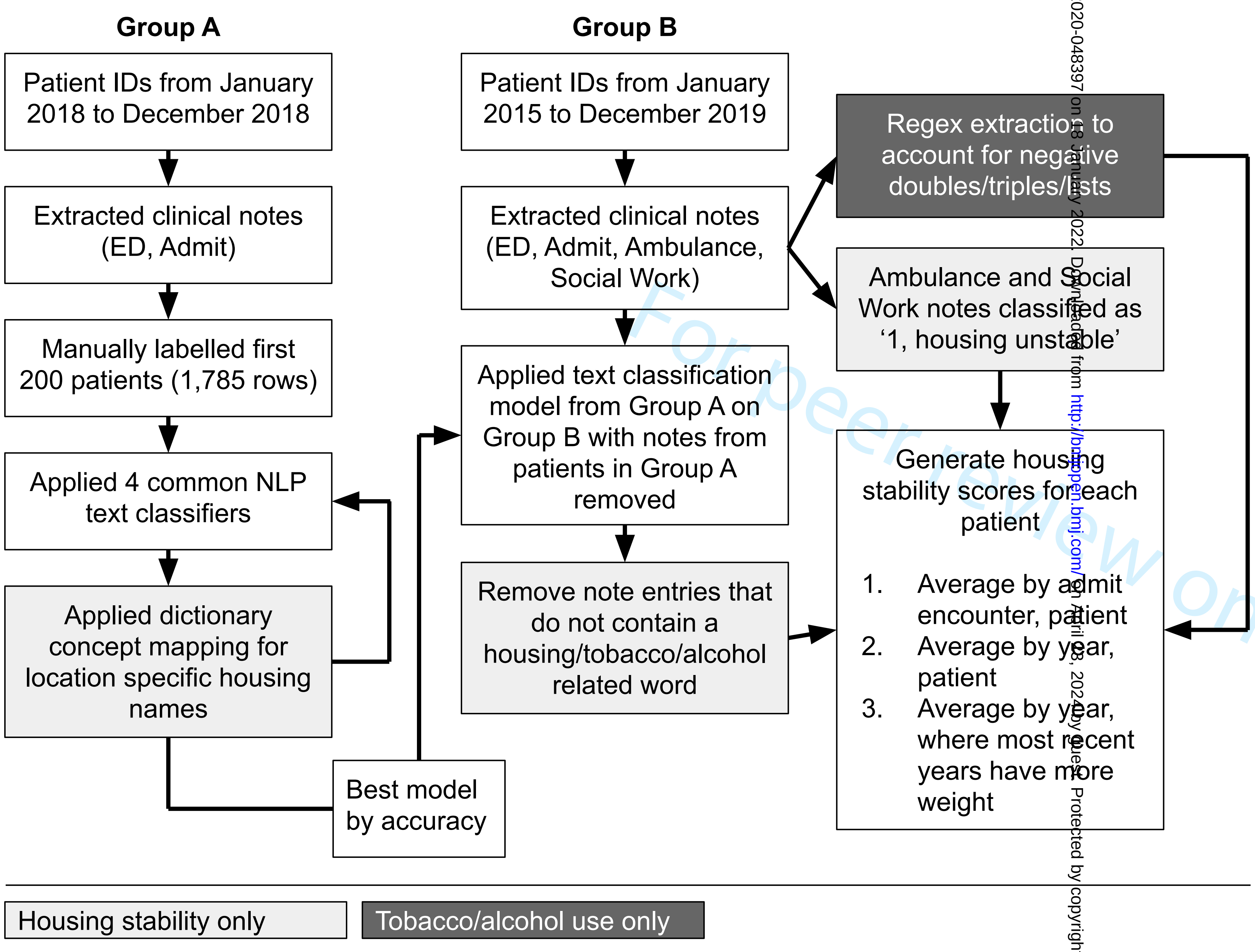
Figure legend:

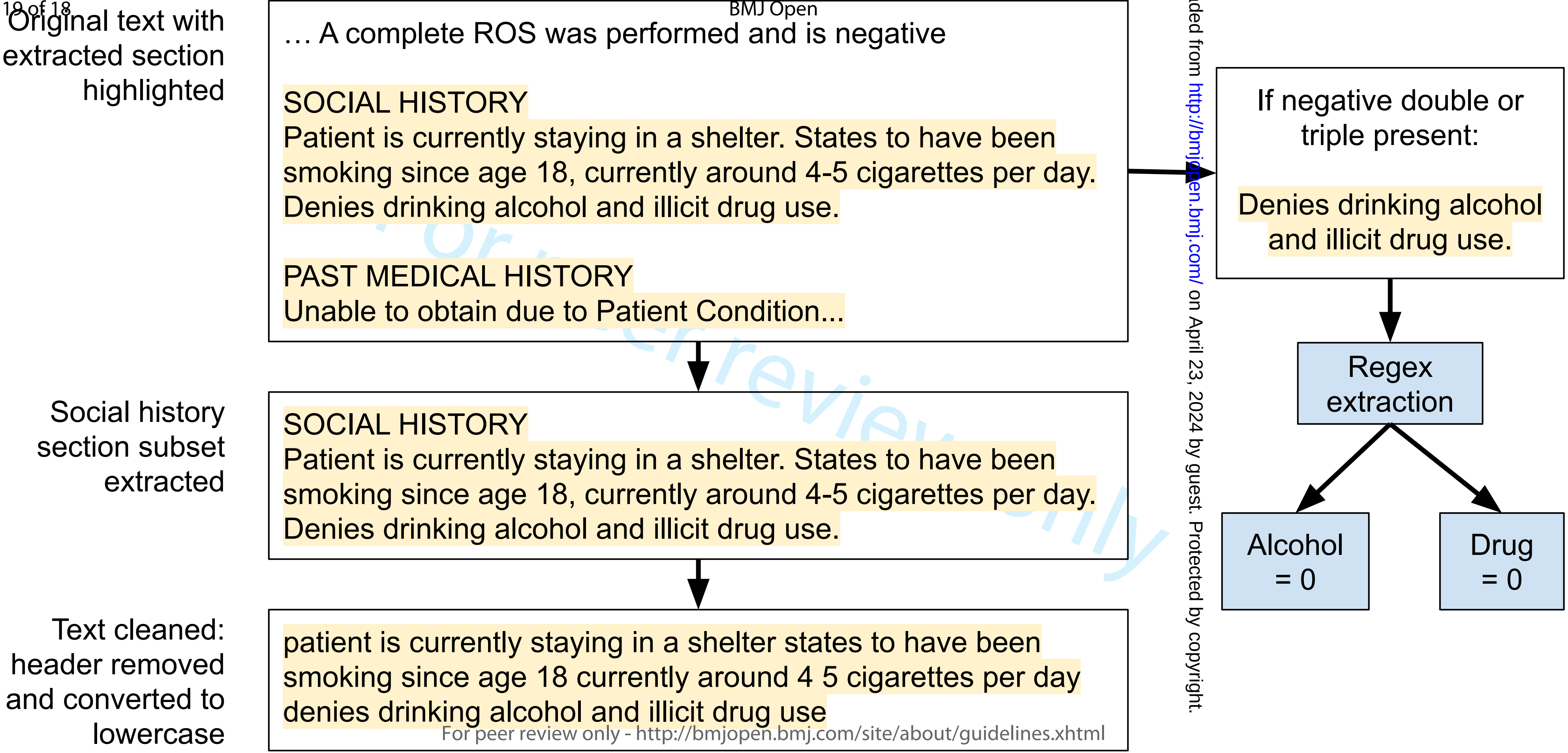
Figure 1: High-level overview of the workflow process

Figure 2: Text extraction, classification, and scoring workflow

Figure 3: Text extraction and cleaning process. Additional steps were performed for notes when classifying text related to tobacco and alcohol use to extract negative sentiment doubles or triples.


Data Extraction of Social Factors
Manual Labelling of Social Factors
Models Generation and Classification
Social Factor Score Creation
Structured and Unstructured Data Comparison


Group A
Patient IDs from January 2018 to December 2018
Extracted clinical notes (ED, Admit)
Manually labelled first 200 patients (1,785 rows)
Applied 4 common NLP text classifiers
Applied dictionary concept mapping for location specific housing names
Best model by accuracy
Group B
Patient IDs from January 2015 to December 2019
Extracted clinical notes (ED, Admit, Ambulance, Social Work)
Applied text classification model from Group A on Group B with notes from patients in Group A removed
Remove note entries that do not contain a housing/tobacco/alcohol related word
Regex extraction to account for negative doubles/triples/lists
Ambulance and Social Work notes classified as '1, housing unstable'
Generate housing stability scores for each patient
1. Average by admit encounter, patient
2. Average by year, patient
3. Average by year, where most recent years have more weight
Housing stability only
Tobacco/alcohol use only


Original text with extracted section highlighted
… A complete ROS was performed and is negative
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
PAST MEDICAL HISTORY
Unable to obtain due to Patient Condition...
Social history section subset extracted
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
Text cleaned: header removed and converted to lowercase
patient is currently staying in a shelter states to have been smoking since age 18 currently around 4 5 cigarettes per day denies drinking alcohol and illicit drug use
If negative double or triple present:
Denies drinking alcohol and illicit drug use.
Regex extraction
Alcohol = 0
Drug = 0

BMJ Open

A Simplified Data Science Approach to Extract Social and Behavioral Determinants: A Retrospective Chart Review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-048397.R4
Article Type:	Original research
Date Submitted by the Author:	02-Nov-2021
Complete List of Authors:	Teng, Andrew; University of Washington, Biomedical Informatics and Medical Education Wilcox, Adam ; University of Washington,
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Health informatics
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, SOCIAL MEDICINE, HISTORY (see Medical History), BIOTECHNOLOGY & BIOINFORMATICS

A Simplified Data Science Approach to Extract Social and Behavioral Determinants: A Retrospective Chart Review

Andrew K. Teng[1], Adam B. Wilcox[1]

Affiliation:
[1] Biomedical Informatics and Medical Education
University of Washington
Seattle, WA, United States

Corresponding Author:
Andrew K. Teng
akteng@uw.edu
Biomedical Informatics and Medical Education, University of Washington
850 Republican Street, Box 358047
Seattle, WA 98195-0005, United States

Keywords: Text extraction, text classification, housing stability, social determinants, data science

Word Count: 4,985

Abstract

Objectives
We aim to extract a subset of social factors from clinical notes using common text classification methods.
Design
Retrospective chart review.
Setting
We collaborated with a local Level I trauma hospital located in an underserved area that has a housing unstable patient population of about 6.5% and extracted text notes related to various social determinants for acute care patients.
Participants
Notes were retrospectively extracted from 43,798 acute care patients.
Methods
We solely utilize open source Python packages to test simple text classification methods that can potentially be easily generalizable and implemented. We extracted social history text from various sources, such as admission and emergency department notes, over a five-year timeframe and performed manual chart reviews to ensure data quality. We manually labelled the sentiment of the notes, treating each text entry independently. Four different models with two different feature selection methods (bag of words (BOW) and bigrams) were used to classify and predict housing stability, tobacco use, and alcohol use status for the extracted clinical text.
Results
From our analysis, we found overall positive results and metrics in applying open-source classification techniques; the accuracy scores were 91.2%, 84.7%, 82.8% for housing stability, tobacco use, and alcohol use respectively. There were many limitations in our analysis including social factors not present due to patient condition, multiple copy-forward entries and shorthand. Additionally, it was difficult to translate usage degrees for tobacco and alcohol use. However, when compared to structured data sources, our classification approach on unstructured notes yielded more results for housing and alcohol use; tobacco use proved less fruitful for unstructured notes.

Article Summary
Strengths and limitations of this study

- From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status.
- Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use.
- Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social and behavioral determinants and can supplement current structured sources to provide a more complete social history for patients.
- However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

I. INTRODUCTION

Most data can be generally categorized as structured or unstructured, where structured data can consist of items such as vital signs and lab results and unstructured data can consist of items such as text notes or images.[1] Although structured data can generally be easier to extract and analyze, unstructured data can potentially provide an array of information not present or easily identifiable in structured data. As healthcare institutions expand data collection to include non-clinical features, more unstructured data surrounding behavioral health and social determinants of health (SDoH) information, are starting to become more readily available. Furthermore, there has a been a growing interest around Medicaid patients, as SDoH can drive up to 80% of health outcomes, especially within this patient demographic.[2] Therefore, SDoH and REAL (Race, Ethnicity and Language) data are now being used for secondary analysis as recent research has indicated that there is a correlation between SDoH and health outcomes and the increasing need to research health disparities across populations.[3]

SDoH and REAL can include housing stability, access jobs and health care services, education level, language, and socioeconomic conditions.[4] These indicators are descriptors of different societies and are useful as predictors of health outcomes and the uptake of health interventions.[5] Because they can potentially be powerful indicators of health, many institutions are now starting to analyze and intake SDoH and REAL information, whether through text notes or standardized coding, such as International Classification of Diseases (ICD).[6] Additionally, SDoH can provide health teams with a greater understanding of a patient condition holistically.[7] However, there are challenges with SDoH intake as there is no standardized SDoH screening tool in the EHR across institutions[8]; additionally, coding schemes like ICD can prove to be unreliable in secondary analysis as coding can oversimplify symptoms and diagnoses leading to coding uncertainties and the fact that coding errors may be present from unintentional mistakes or even upcoding.[9,10] Additionally certain SDoH data may be more complete than others due to reimbursement incentives or other priorities.[11] Past research has shown that hospital readmissions are highly influenced by patient health status and SDoH and suggest that clinical staff and researchers should consider SDoH when assessing readmission risk.[12]

The 2018-2019 [redacted for review] Community Health Needs Assessment (CHNA) reported the results from a health needs assessment survey given to residents to identify regional perceived healthcare issues. It was determined that housing affordability and housing stability were major challenges dominating overall health.[13] Mental health was also highlighted as a challenge for healthcare providers; mental illness can be caused by depression, schizophrenia, and alcohol and substance-related disorders.[13] The CHNA reported that adults in the lowest income tier were about 15 times more likely to experience severe psychological distress compared to their high-income counterparts. Additionally, it noted that part of the region had continued challenges with adult smoking rates.[13] Locally, it is estimated that there are at least 22,000 homeless individuals in [redacted for review] and more than 12,000 people in the [redacted for review] region, a four percent increase over the previous year.[14] Housing instability is associated with various health inequalities, such as shorter life expectancy, higher morbidity, and increased usage of acute hospital services, "as the social determinants of homelessness and health inequities are often intertwined, and long term homelessness further exacerbates poor health".[15] It is therefore important to treat housing stability and other SDoH as a combined health issue to aid in improving health outcomes in clinical settings. Although some research has shown that patients who experience housing instability are more likely to die following admission for severe sepsis than those with insurance,[16] other research indicates that the effects of health inequalities are still unclear and need further investigation.[17] Additionally, various behavioral habits, including tobacco and alcohol use, although may not directly be considered a SDoH, can impact health decisions and outcomes. For example, one study found that participants who drank alcohol and reported tobacco use consumed more foods higher in fat and sugar, low in vitamins and minerals as well as foods, considered by them to be less healthy and prepared in a less healthy way.[18]

Within our region, it has been noted in recent years that the smoking rate is around 13 percent; however, among Black/African-Americans or individuals with multiple races, is double the rate among white adults and four times higher than Asian adults. Additionally, it was reported that, when compared to high income households, low income households were three times more likely to be smokers.[13,19] Drug and alcohol use also shared similar metrics; within the region, "drug and alcohol-caused deaths was 22% higher among Blacks and four times greater among American Indian/Alaskan Native than among non-Hispanic Whites" and alcohol use represented 4.97 per 100,000 deaths locally in 2015.[20,21] Therefore it may be important to look at social determinants and health behaviors, together known as social and behavioral determinants of health (SBDH) to better understand the patient population.[18]

Recent technological advances in machine learning and artificial intelligence have shown great potential in providing a pathway for informaticians and clinicians to better understand unstructured data.

Within the clinical setting, there have been numerous approaches in adopting natural language processing (NLP) to aid with processing unstructured clinical text notes. Common uses of NLP include extracting diagnoses and chief complaints as well as grouping of information for quality improvement. There are various NLP methods that can be used in the clinical setting, such as automatic tagging of conditions or variables of interest, sentiment classification, or even text extraction. Various open source NLP and ontological tools, such as Automated Retrieval Console, Apache clinical Text Analysis and Knowledge Extraction System (Apache cTAKES), MetaMap, and HITEx, Unified Medical Language System (UMLS) Metathesaurus and BioPortal have been used to aid with text extraction or classification.[22–24] On the other hand less complex classification methods have been used as well to identify specific groups of patients, risk assessment, or aid in validating structured annotation.[25,26,27] A recent scoping review found that although practitioners collect a variety of SBDH data at point of care through EHR, the overall use of automated technology is limited to date.[28]

With the idea of implementing an easily generalizable approach to classify selected social factors, we extracted both unstructured and structured data sources related to SBDH from a local hospital to identify and generate a framework to automatically extract and classify SBDH from text notes. We focused on housing stability status, tobacco use, and alcohol use. These three social factors were chosen due to their direct impact on health outcomes and the local public health impact[14–18] and presence in the EHR. To tackle challenges associated with SBDH extraction from unstructured text notes, we aimed to create a generalizable framework using low barrier open-source tools that are commonly used in the data science field. Because notes and stylistic choices can be institution and location specific, we sought not to create a model that is generalizable but rather a simplified method that could be potentially easily implemented using common off the shelf NLP and data science tools.

II. METHODS

Study Design and Overview

A high-level overview of our workflow can be seen in Figure 1. We retrospectively extracted patient data from the acute care setting at a Level I trauma center and academic teaching hospital with the aim to create a general and easily applicable workflow to extract and classify SBDH factors from clinical notes. We applied a two-pronged approach and collected unstructured data from a subset of patients over a 1-year timespan (Group A) to create and test the text classification model and also collected structured and unstructured data from a subset of patients over a 5-year timespan (Group B) to apply the best model created from Group A and compare results between the two data types. We performed automatic classification and scoring of patients via various NLP classification methods on three social factors: (1) housing stability, (2) tobacco use, and (3) alcohol use. Our general workflow for housing stability, a similar approach was also used for tobacco and alcohol use, can be seen in Figure 2.

Study Population

Data were extracted from [redacted for review], a 413-bed academic hospital that has a patient population consisting mostly from Washington, but also from a five-state area.[29] In 2014, there were 17,121 inpatient admissions, where 19 percent of the patients belong to a racial or ethnic minority and 37 percent of patients were enrolled in Medicaid.[29,30] Additionally, in 2015, the non-US born population was estimated to be around 21 percent in [redacted for review], highlighting the potential diversity that could be found with this patient population.[30]

Data Sources, Extraction, and Validation

We extracted both structured and unstructured data sources related to housing stability, tobacco use, and alcohol use using SQL queries called directly from an integrated python-based Jupyter Notebook:

a. Structured data sources include billing and diagnostic/International Classification of Disease (ICD) 9 and 10 codes, questionnaire or Epic SmartForm responses, address fields (location), problem list (ICD 9), patient encounters, clinical events (actual encounters of care), and discharge/disposition location.
b. Unstructured data sources consisted of text notes from the emergency department (ED), admission (admit) notes, social work, and ambulance notes.

Discharge notes were not explored as they were not recorded in the same subdivided format as the admit and ED notes, making selective text extraction of SBDH difficult. From our initial list of patient identifiers over a one-year timespan from Group A, we performed manual EHR validation of a random subset of 50 patients to validate the completeness of the clinical notes and confirm the location of social history and social factors in clinical notes. Extensive research and conversations with an internal data analyst confirmed the location of these topics (housing, tobacco use, and alcohol use) within structured data sources.

Data Cleaning

After confirmation, clinical notes were extracted for both Groups A and B. The notes were cleaned (e.g. symbols removed, converted to lowercase) prior to classification and analysis in the Python Jupyter notebook via NLTK. Our general text extraction and cleaning workflow can be seen in Figure 3. However, housing stability notes and tobacco or alcohol use notes were stylistically and grammatically different, and both sets needed distinct additional cleaning steps. Housing stability notes that contained the phrase 'not homeless' were converted via regex to say 'housed' instead. Additionally, for housing stability, a concept dictionary was also created to substitute local facility names with more general concept (e.g. 'Union Gospel Mission' was converted to 'shelter'). This was done to explore how the algorithms handle formal nouns.

For text notes in Group B, we performed an additional concept extraction step. Tobacco use and alcohol use notes often contained incomplete (lacking the subject, predicate, object format) triples or doubles (e.g. 'Denies smoking, drinking, drugs'). Due to their incomplete sentence structures, common NLP tools to parse, extract, and classify triples, such as Stanford CoreNLP, were not suitable as these tools rely on having all three parts of the triple present. These notes related to tobacco and alcohol use therefore underwent an additional step that performed a separate relation extraction that would first identify a negative sentiment word (e.g. denies), then individually extract the following SBDH related objects in the list by commas or conjunctions (e.g. and, or), and then label, or reclassify if necessary, the negative sentiment to all components of the list. Our process can be seen in the left side of Figure 3. If the regex extraction of negative lists resulted in a different result from the text classification prediction, the regex extraction would overwrite the end result prior to scoring. Once these steps were performed, the data were considered clean and suitable for classification.

Model building

Cleaned text from Group A were used to generate and test the classification models. These notes were split in 70/30 validation and testing sets. We applied four different common NLP text classification models to the testing sets (via SciKit Learn): multinomial naïve Bayes, support vector machine, logistic regression, and random forest. Default parameters and a bag-of-words approach were used. The best performing model by accuracy was then chosen and applied to the larger corpus, Group B, with notes from patients in Group A removed, to avoid overfitting and classification bias. This process was performed for housing, tobacco use, and alcohol use.

Scoring generation

In order to create a simple method of identifying patients who are experiencing social instability, we created a scoring metric based on the classified notes. After applying the optimum model by accuracy to the entire corpus of extracted text notes, housing stability, tobacco use, and alcohol use scores were generated. Patient identifiers were mapped by patient location and those who were not in the acute care setting during this timeframe were removed. Three different scoring approaches were used to describe these social factors: (1) predictions were averaged by patient encounter, then averaged by patient identifier, (2) predictions were averaged by year, then by patient identifier, and (3) predictions were averaged by year, where each year then had a weight where the most recent year had the highest weight and the furthest year had the lowest weight (e.g. predictions from 2019 were weighted by a factor of 5 and predictions from 2015 were weighted by a factor of 1). This scoring generation process was then repeated on our structured data for all three social factors and the results were compared and analyzed. Structured data was also extracted for our list of patients in Group B.

Patient and Public Involvement

No patients were involved. The retrospective exploration is a part of a larger study and was approved by the [redacted for review] Institutional Review Board #STUDY00006723. Patient data elements, including encounter identifiers, race, age, and notes with SBDH, were extracted directly from the data warehouse and stored on encrypted computers and were not distributed or shared outside of the secured and closed environment. No patient identifiers or names were stored in this analysis.

III. RESULTS

Characteristics of study subjects

Clinical notes (ED, admit, social work, and ambulance) between 2015 and 2019 were extracted and included, forming Group B. Notes from the first 200 patients were included in Group A and notes from 147,457 patients were included in Group B. During the same timeframe, 61,767 patients were in acute care. After extraction and model prediction, the patient notes were cross referenced with inpatient location and only notes from those who were in acute care were retained, for a total of 43,798 patients from 2015 to 2019. The patient demographics of this final subset were 63% (*n*=27,575) male, 37% (*n*=16,223) female, 88.2% (*n*=38,634) not Hispanic or Latino, and 10.5% (*n*=4,609) Hispanic or Latino, and 1.3% (*n*=555) unknown or not answered. Further descriptive statistics can be found in Table 1.

Table 1: Population demographics

Race (*n*=43,798)	*n* (%)
White or Caucasian	31,575 (72.1%)
Black or African American	4,812 (11.0%)
Asian	3,174 (7.2%)
American Indian or Alaska Native	1,165 (2.7%)
Native Hawaiian or other Pacific Islander	524 (1.2%)
Multiple races	3 (0%)
Unavailable, unknown, or missing	2,545 (5.8%)
Age range (*n*=43,798)	***n* (%)**
0-18	1,856 (4.2%)
19-44	12,437 (28.4%)
45-64	14,863 (33.9%)
65-84	11,902 (27.2%)
85 and over	2,740 (6.3%)

Data attributes

Table 2 illustrates the amount of data for each corresponding extraction level, specifically for housing status. We first started with extracting text from the ED and admit notes, forming Group A, which consisted of 50,000 rows or text entries and covered 3,200 unique patients, over a one-year timeframe. From there, we manually labelled housing stability concepts in a binary fashion, where 0 would indicate housing stability and 1 would indicate any level of housing instability, regardless of severity. As manual labelling can be a labor-intensive process, only the first 6,000 text rows were labelled, covering 218 unique patients. However, within these first 6,000 rows, numerous notes did not contain text that alluded to housing status or were empty due to patient condition. Therefore, only 1,785 out of the 6,000 rows were labelled, covering 200 unique patients, where 995 (55.7%) were labelled as housing stable and 790 (44.3%) were labelled as housing unstable. We also found that 5.7% of the entries within this subset were duplicates or copy-forward entries. The same workflow was performed for labelling tobacco and alcohol use. However, only 1,108 rows were labelled for tobacco use and 1,220 rows for alcohol use, where in both cases 0 indicated no use, 1 indicated rare/previous/occasional use, and 2 indicated current use, regardless of degree. Tobacco use resulted in 446 (40.3%) labels for no use, 129 (11.6%) labels for rare/previous/occasional use, and 533 (48.1%) labels for current use. Similarly, alcohol use resulted in

595 (48.8%) labels for no use, 185 (15.2%) labels for rare/previous/occasional use, and 440 (36%) labels for current use.

Table 2: Extracted data amounts for housing status

Level of extraction	Rows (*n*)	Unique patients (*n*)	Unique encounters (*n*)	Social history entries (*n*/unique)
ED and Admit notes	49,955	3,233	15,664	21,876/21,334
Housing, Tobacco, Alcohol Information	6,000	218	1,995	2,408/2,211
Remove nulls/missing data	Housing: 1,785 Tobacco: 1,108 Alcohol: 1,220	Housing: 200 Tobacco: 179 Alcohol: 181	1,361	1,785/1,684

Model performance

Four different common text classifiers, mentioned in the Methods section, were applied to the manually labelled Group A data. The statistical metrics, including accuracy, precision, and recall, can be seen in Table 3 and 4. The accuracies between the classifiers and each classification technique for housing stability were overall fairly high ranging from 84.36-92.18%. The accuracies for tobacco and alcohol use were lower, ranging from 70.87-84.68% for tobacco use and 69.95-82.79% for alcohol use. Additionally, for each top performing model, the most influential words for text classification, for each social factor, can be seen in Table 5. The best performing classification models were selected for each social factor and were used to apply the model to our entire corpus in Group B.

Table 3: Accuracies amongst text classifiers

	n=1	n=1-2
Multinomial naïve Bayes	Housing: 91.62% Tobacco: 70.87% Alcohol: 70.77%	Housing: 91.43% Tobacco: 77.18% Alcohol: 69.95%
Support vector machine	Housing: **92.18%** Tobacco: 81.08% Alcohol: 76.50%	Housing: 91.99% Tobacco: 82.88% Alcohol: 81.97%
Logistic regression	Housing: 84.36% Tobacco: 75.38% Alcohol: 77.60%	Housing: 90.13% Tobacco: **84.68%** Alcohol: **82.79%**
Random forest	Housing: 90.50% Tobacco: 76.28% Alcohol: 71.31%	Housing: 91.25% Tobacco: 78.98% Alcohol: 75.68%

Table 4: Best performing classifier detailed metrics

	Classifier	Accuracy	Recall	Precision	F1
Housing status*	Support vector machine (n=1)	0.92	0.93/0.91 (0/1)	0.94/0.90	0.93/0.91
Tobacco use**	Logistic Regression (n=1-2)	0.85	0.82/0.95/0.86 (0,1,2)	0.96/0.43/0.87 (0,1,2)	0.88/0.60/0.87 (0,1,2)
Alcohol use**	Logistic Regression (n=1-2)	0.83	0.86/0.73/0.81 (0,1,2)	0.93/0.44/0.88 (0,1,2)	0.89/0.55/0.84 (0,1,2)

* 0: no use, 1: current use
** 0: no use, 1: rare/occasional/history, 2: current use

Table 5: Word or phrase importance ranking

Social factor (Classifier)	Top 20 weighted words

Housing stability (support vector machine, n=1)	['friends' 'motel' 'stay' 'cigs' 'found' 'street' 'stays' 'streets' 'van' 'incarcerated' 'desc' 'currently' 'undomiciled' 'friend' 'respite' 'kcj' 'shelters' 'homelessness' 'shelter' 'homeless']
No tobacco use (logistic regression, n=1,2)	['use denies' 'deneis' 'lives' 'tobacco drug' 'seattle denies' 'use results' 'lives seattle' 'alcohol tobacco' 'tobacco drugs' 'never smoker' 'etoh tobacco' 'drinking' 'seattle tobacco' 'denies cigarettes' 'drugs tobacco' 'denies alcohol' 'tobacco alcohol' 'denies smoking' 'denies' 'denies tobacco']
No alcohol use (logistic regression, n=1,2)	['care' 'ppd' 'tobacco' 'smoking' 'etoh tobacco' 'history cocaine' 'tobacco alcohol' 'etoh illicit' 'alcohol tobacco' 'etoh drug' 'drugs etoh' 'alcohol drug' 'use none' 'alcohol drugs' 'drug etoh' 'denies alcohol' 'lives' 'denies drug' 'denies etoh' 'denies']

Scoring results and comparison

After classifying text for housing stability, tobacco use, and alcohol use for patients in Group B, we applied a scoring metric scheme, described in the Methods section. We generated three different scores that were calculated and weighted differently based on time. Our final score weighs more recent note entries and their resulting classification score higher than notes from previous years as social factors and their influence can change over time. Using the same process, we extracted and scored housing stability, tobacco use, and alcohol use with structured data sources and compared the results with the unstructured process.

I. Housing stability

Using notes, we classified 839 patients as housing unstable, a score above 0.5, and 21,370 patients as housing stable, a score of 0.5 and below. In total, we classified 22,209 patients with this text classification workflow, which covered 50.71% of the acute care patients within the same timeframe. When compared with structured data sources, only 791 (1.81%) additional patients were found.

II. Tobacco use

We classified 4,911 patients as currently using tobacco, regardless of amount or degree (1.5-2) using text notes. We classified 1,480 patients as having rare/occasional/past use of tobacco (0.5-1.5), and 7,139 patients as not using tobacco (0-0.5). In total, we classified 13,530 patients with this text classification workflow, which covered 30.9% of the acute care patients within the same timeframe. When compared with structured data sources, 17,9351 (40.9%) additional patients were captured.

III. Alcohol use

We classified 2,738 patients as currently using alcohol, regardless of amount or degree (1.5-2) using text notes. We classified 4,050 patients as having rare/occasional/past use of alcohol (0.5-1.5), and 13,885 patients as not drinking alcohol (0-0.5). In total, we classified 20,673 patients with this text classification workflow, which covered 37% of the acute care patients within the same timeframe. When compared with structured data sources, no additional patients were found.

IV. DISCUSSION

Our approach to a simple text classification method for various social determinants of health have shown positive results. The selected classification models were chosen as they were the most commonly used classification models when researching text classification techniques. Furthermore, these models were robust enough to curtail the need for more complex machine learning based text classification methods, which may be harder to interpret in the clinical space as the weights and decisions can be confiscated due to the black box nature of these more complex classification methods. Generally, linear models are fast to train, can work well with sparse data, and offer interpretability.[31] Additionally, recent research has also suggested that more complex machine learning approaches may not yield statistically significant improvements in predictive power to justify the time and effort necessary to implement and test these more complex methods. Although promising, more advanced methods of NLP, such as convoluted neural networks, may not provide a significant tradeoff in improvement or accuracy versus transparent understanding of rule-based approaches. In fact, Yao et al. found that the F1 scores for CNN via TensorFlow did not improve significantly for interested features when compared to logistic regression and support vector machine implementations.[32] Finally, generalizable methods to create institution-specific models can be better for the healthcare system as a whole as each institution records clinical information with variances.

Although SBDH information and other social factors can be indicative of overall health, collection of SBDH heavily relies on clinical staff to screen and document SBDH. Furthermore, it also assumes that patients will respond accurately and truthfully. Various financial incentives from the federal level have propelled collection of social factors, such as tobacco use and tobacco cessation. However, other social factors, which can be equally as important, such as alcohol use are not incentivized to be captured; rather only more severe instances are incentivized, such as alcohol dependence or alcohol addiction or disorder.[33,11] Due to this discrepancy, we found that structured data sources were less reliable, and that text classification aided in detailing a patient more holistically.

Our text classification of unstructured data relied solely on ED, admit, social work, and ambulatory notes as our parsing and extraction method could only work with notes in a certain format with the social history heading. Social factors and other social history could also be recorded in other locations, but were not compatible with our approach. Furthermore, social work and ambulatory notes used for housing status only and were only extracted if the notes contained a word or phrase related to housing instability. This approach was used as the notes were typically stored in a more unstructured format compared to the ED and admit notes; there were no section headers. The lack of section headers increased the difficulty to extract the notes and the notes would often verbiage that would interfere with the simple text classification approach that we used. Therefore, we decided to extract notes that contained words relating to housing instability. Additionally, tobacco and alcohol use notes had stylistic and grammatical challenges. These social factors were often grouped together in incomplete triples (e.g. "denies drinking, smoking, illicit drug use"). The classification algorithms often had trouble reciprocating the negative connotation to all components of the triple. Therefore, we used regex to specifically extract these triples and classify the note based on the presence of words related to tobacco or alcohol. Without this additional data cleaning or manipulation step, the negative sentiment in a list would not have been applied to all elements within the list, but rather only the first element. In our example of 'denies smoking, drinking, drugs', the negative sentiment of 'denies' would have only been applied to smoking as smoking immediately follows 'denies'. However, with our additional concept extraction step, the negative sentiment of 'denies' is now also applied to 'drinking' and 'drugs'. These results would then override the text classification algorithm, if there was a discrepancy. Therefore, the scoring metrics for these cases would not necessarily reflect the accuracy or performance of our scoring method.

It was interesting to find that tobacco use was recorded significantly more often in structured data sources compared to alcohol use and housing stability. However, because tobacco use is a (Centers for Medicare and Medicare Services) CMS core quality measure, it can be expected that this feature is more available in structured form as it is often directly asked to the patient on intake forms, screeners, or during cessation treatment.[11] Furthermore, the Joint Commission created the Tobacco Performance Measure Set, which are three standardized performance measures addressing tobacco screening and cessation counseling: (1) Tobacco use screening of patients 18 years and over, (2) Tobacco use treatment, including counseling and medication during hospitalization, and (3) Tobacco use treatment management plan at discharge. CMS began using these performance measures in 2016.[34] Because alcohol consumption is not a recommended CMS core quality measure for adults, the amount of data regarding alcohol use is not complete in structured form as it may not be consistently collected during intake procedures.

Past research has consistently pointed towards SBDH impacting patient health and outcomes. However, collection of SBDH can be a major limiting factor in the ability to model and integrate these data. There has not been a standardized collection process for SBDH data across the institution, whether it is recorded through notes or electronic forms. Additionally, many times, SBDH data may not be asked due to patient condition or it might not be updated regularly. Providers and healthcare institutions should strive to collect SBDH data more regularly even if the data fields are not empty as SBDH status can change. These intake procedures should be present and not optional; currently, only language preference must be completed due to translation laws in place. Additionally, educating patients to utilize patient portals and update information via these portals can provide more current SBDH information. However, we should note that vulnerable populations would most likely not be the primary audience to utilize this feature, and this is the subpopulation that arguably needs more attention.

Limitations

Our study has numerous limitations. There were two distinct areas in our workflow that required manual attention: (1) EHR review and (2) labelling of features. Manual EHR review was performed to ensure that the notes contained social history information in a consistent location prior to widespread text extraction. We initially validated this with a random set of 10 patients, but later expanded our validation to 25 patients. We felt that having consistent results with the 25 patients indicated a high level of confidence. Manual labelling of features was time consuming and taxing. Although only one author performed the feature labelling, having multiple team members would provide better and possibly more consistent classification.

This approach, although we aim to create a generalizable workflow, is still stunted by local customizations due to unique nuances in note taking language. Patients can withhold information about their social challenges, making text classification harder to perform due to incorrect incoming data streams. Our approach relies on the fact that the patient has been seen within the healthcare system at some point in the past five years. This approach would not be applicable to those who are new to the institution or those who are not immediately identifiable. Classification levels for unstructured notes are not concrete as descriptive wording is also not concrete and can vary (e.g. "patient was a former smoker", "patient quit last week", "patient is an occasional smoker", etc.). Structured data sources can add a more concrete sense to the classification. There were 5.7% copy-forward entries present as data collection of social factors may not always be appropriate (e.g. patient is inebriated, in an altered mental state, etc.). We did not incorporate outside ontologies, such as UMLS or MetaMap, as we were interested in creating a simple text classification approach that did not need to rely on outside entities. Furthermore, we believe that these ontologies would not have added a significant improvement in our approach due to the social factors (housing, alcohol, tobacco) that were investigated. Although minimized, applying NLP to clinical notes will always present limitations and risks with biased models, biased data, and data privacy.[35]

Community needs are constantly changing as the health of the community is not static. Currently, the King County CHNA has identified obesity, healthcare access, insurance status and drug use as other potential SBDH information to explore. These data types would be stored in different areas of the EHR and within different notes. It would be interesting to see if our designed workflow presented could be applicable and generalized to meet the needs of other SBDH data. Although we aimed to create a simplified framework to extract SBDH data from clinical notes, more complex methods such as convoluted neural networks and more advanced NLP part of speech tagging may be worth exploring as they may help improve accuracy and precision of the classification. As more notes become available for patients, it will also be important to keep in mind the potential bias of having more notes present from sicker patients and evaluating ways to reduce this bias.

We sourced data from solely one medical center. Patients might have had encounters or other visit types in neighboring hospitals and healthcare systems in the region. The lack of data sharing between institutions prevents holistic collection of SBDH data. Data completeness is vitally important to the quality and accuracy of models that are dependent on big data. Poor data quality and completeness lead to lower utilization and the lack of data can potentially lead to mistakes in the decision-making process; additionally, since there is no single or standardized source for SBDH data, the diversity of data and complexity of the associated data structures increase the difficulty and bottlenecks for data integration.[36] The lack of a standardized methodology to collect and store all SBDH data will limit the potential of this research field. Additionally, SBDH factors are constantly changing for patients as their behaviors can change depending on their circumstance. Being able to aggregate these data and create adaptable models is crucial as these features are never static. Furthermore, public health and outreach services fluctuate over time. Creating a method or utilizing an API to update the list of community shelters and other places for homeless services would be necessary to maintain an accurate understanding of a patients housing status.

V. CONCLUSION

From our analysis, we can first see that text classifiers are promising when applied to extracted clinical notes for housing stability, tobacco use, and alcohol use status. Additionally, we found that structured data sources, such as diagnosis codes and intake surveys, vary and may not be the most holistic approach to understanding housing stability, tobacco use, and alcohol use. Our simplified approach has shown that open source simple text classifiers can be used to predict text sentiment for social determinants and can supplement current structured sources to provide a more complete social history for patients. However, even with a few limitations with our approach, we believe that this workflow can help inform clinicians and provide an easily implementable snapshot on patient social history.

Contributor statement:
AT performed the data extraction, tool building, and analysis. AW provided guidance and verification when needed.

Competing interests:
There are no competing interests.

Funding:
This work was supported by the U.S. Department of Health and Human Services, National Library of Medicine Training Grant T15LM007442.

Data sharing statement
The data used are unable to be shared due to patient privacy, confidentiality, and United States healthcare laws.

Ethics statement
This study does not involve human participants. Informed consent prior to participating in the study was not applicable as there were no human participants included. This research is a part of a larger study that has been approved by the University of Washington IRB #STUDY00006723.

VI. REFERENCES

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*. 2014;2(1). doi:10.1186/2047-2501-2-3

2. Hood CM, Gennuso KP, Swain GR, Catlin BB. County Health Rankings. *American Journal of Preventive Medicine*. 2016;50(2):129-135. doi:10.1016/j.amepre.2015.08.024

3. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving Electronic Medical Records Upstream. *American Journal of Preventive Medicine*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

4. Social Determinants of Health. HealthyPeople.gov. Accessed February 1, 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

5. Social Determinants. Institute for Health Metrics and Evaluation. Accessed February 1, 2020. http://www.healthdata.org/social-determinants

6. Nerenz DR. Health Care Organizations' Use Of Race/Ethnicity Data To Address Quality Disparities. *Health Affairs*. 2005;24(2):409-416. doi:10.1377/hlthaff.24.2.409

7. Andermann A. Taking action on the social determinants of health in clinical practice: a framework for health professionals. *Canadian Medical Association Journal*. 2016;188(17-18):E474-E483. doi:10.1503/cmaj.160177

8. Olson DP, Oldfield BJ, Navarro SM. Standardizing Social Determinants Of Health Assessments. Published March 18, 2019. https://www.healthaffairs.org/do/10.1377/hblog20190311.823116/full/

9. Wockenfuss R, Frese T, Herrmann K, Claussnitzer M, Sandholzer H. Three- and four-digit ICD-10 is not a reliable classification system in primary care. *Scand J Prim Health Care*. 2009;27(3):131-136. doi:10.1080/02813430903072215

10. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40(5 Pt 2):1620-1639. doi:10.1111/j.1475-6773.2005.00444.x

11. Eligible Professional Meaningful Use Core Measures Measure 9 of 13. Published online May 2014. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/9_Record_Smoking_Status.pdf

12. Lax Y, Martinez M, Brown NM. Social Determinants of Health and Hospital Readmission. *Pediatrics*. 2017;140(5):e20171427. doi:10.1542/peds.2017-1427

13. King County Community Health Needs Assessment 2018/2019. Presented at the: https://www.kingcounty.gov/depts/health/data/community-health-indicators/~/media/depts/health/data/documents/2018-2019-Joint-CHNA-Report.ashx

14. Henry M, Mahathey A, Morrill T, et al. *The 2018 Annual Homeless Assessment Report (AHAR) to Congress*. The U.S. Department of Housing and Urban Development OFFICE OF COMMUNITY PLANNING AND DEVELOPMENT; 2018. https://files.hudexchange.info/resources/documents/2018-AHAR-Part-1.pdf

15. Stafford A, Wood L. Tackling Health Disparities for People Who Are Homeless? Start with Social Determinants. *International Journal of Environmental Research and Public Health*. 2017;14(12):1535. doi:10.3390/ijerph14121535

16. Ahmad S, Baig S, Taneja A, Nanchal R, Kumar G. The Outcomes of Severe Sepsis in Homeless. *Chest*. 2014;146(4):230A. doi:10.1378/chest.1995140

17. Bambra C, Gibson M, Sowden A, Wright K, Whitehead M, Petticrew M. Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. *Journal of Epidemiology & Community Health*. 2010;64(4):284-291. doi:10.1136/jech.2008.082743

18. Papadopoulou S, Hassapidou M, Katsiki N, et al. Relationships Between Alcohol Consumption, Smoking Status and Food Habits in Greek Adolescents. Vascular Implications for the Future. *Current Vascular Pharmacology*. 2017;15(2):167-173. doi:10.2174/1570161114666161024123357

19. Wong E. *Tobacco Use in King County*. Public Health Seattle & King County; 2012. https://www.kingcounty.gov/depts/health/data/~/media/depts/health/data/documents/tobacco-use-in-king-county-may-2012.ashx

20. Bogan S, Donohue B. King County drug and alcohol deaths rose 9.5% in 2018. https://newsroom.uw.edu/news/king-county-drug-and-alcohol-deaths-rose-95-2018

21. Drug-caused deaths in King County. Published online February 21, 2017. https://adai.washington.edu/WAdata/KingCountyDrugDeaths.htm

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. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513. doi:10.1136/jamia.2009.001560

24. Gundlapalli AV, Carter ME, Divita G, et al. Extracting Concepts Related to Homelessness from the Free Text of VA Electronic Medical Records. *AMIA Annu Symp Proc*. 2014;2014:589-598.

25. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE*. Published online 2017. doi:10.1371/journal.pone.0174708

26. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment: *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2018;77(2):160-166. doi:10.1097/QAI.0000000000001580

27. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying Patients with Significant Problems Related to Social Determinants of Health with Natural Language Processing. *Stud Health Technol Inform*. 2019;264:1456-1457. doi:10.3233/SHTI190482

28. Berg K, Doktorchik C, Quan H, Saini V. *Meaningful Information in the Age of Big Data: A Scoping Review on Social Determinants of Health Data Collection for Electronic Health Records*. In Review; 2019. doi:10.21203/rs.2.16433/v1

29. 2015 CDC HA-VTE PREVENTION CHALLENGE CHAMPION. https://www.cdc.gov/ncbddd/dvt/documents/champ-fact-sheet-harborview.pdf

30. Bulger EM, Kastl JG, Maier RV. The history of Harborview Medical Center and the Washington State Trauma System. *Trauma Surgery & Acute Care Open*. 2017;2(1):e000091. doi:10.1136/tsaco-2017-000091

31. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*. 2017;105:110-120. doi:10.1016/j.ijmedinf.2017.06.004

32. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*. 2019;19(S3). doi:10.1186/s12911-019-0781-4

33. Medicare & Medicaid EHR Incentive Program: Meaningful Use Stage 1 Requirements Overview. Presented at the: 2010. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/MU_Stage1_ReqOverview.pdf

34. Quality Measures and Tobacco Cessation. https://www.bhthechange.org/wp-content/uploads/2017/12/Quality-Measures-and-Tobacco-Cessation.pdf

35. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H. Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*. Published online June 4, 2020:161-168. doi:10.14745/ccdr.v46i06a02

36. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *CODATA*. 2015;14(0):2. doi:10.5334/dsj-2015-002

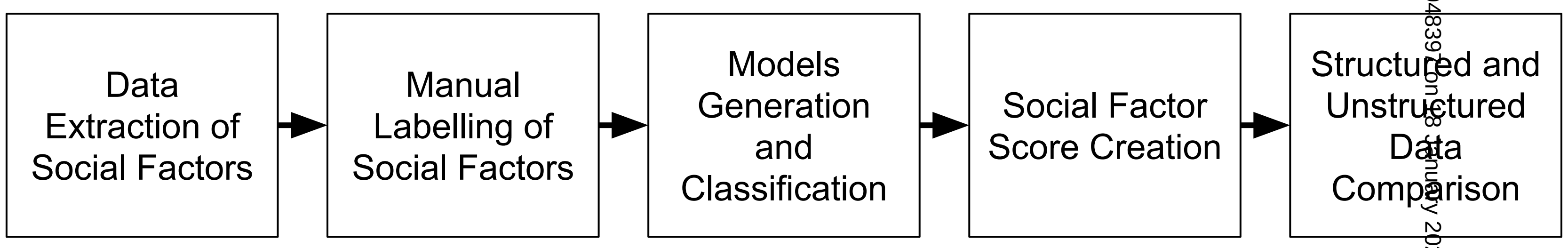
Acknowledgements:
Sally Lee
Abdelhak Abdou
Marion Granich
David Carlbom

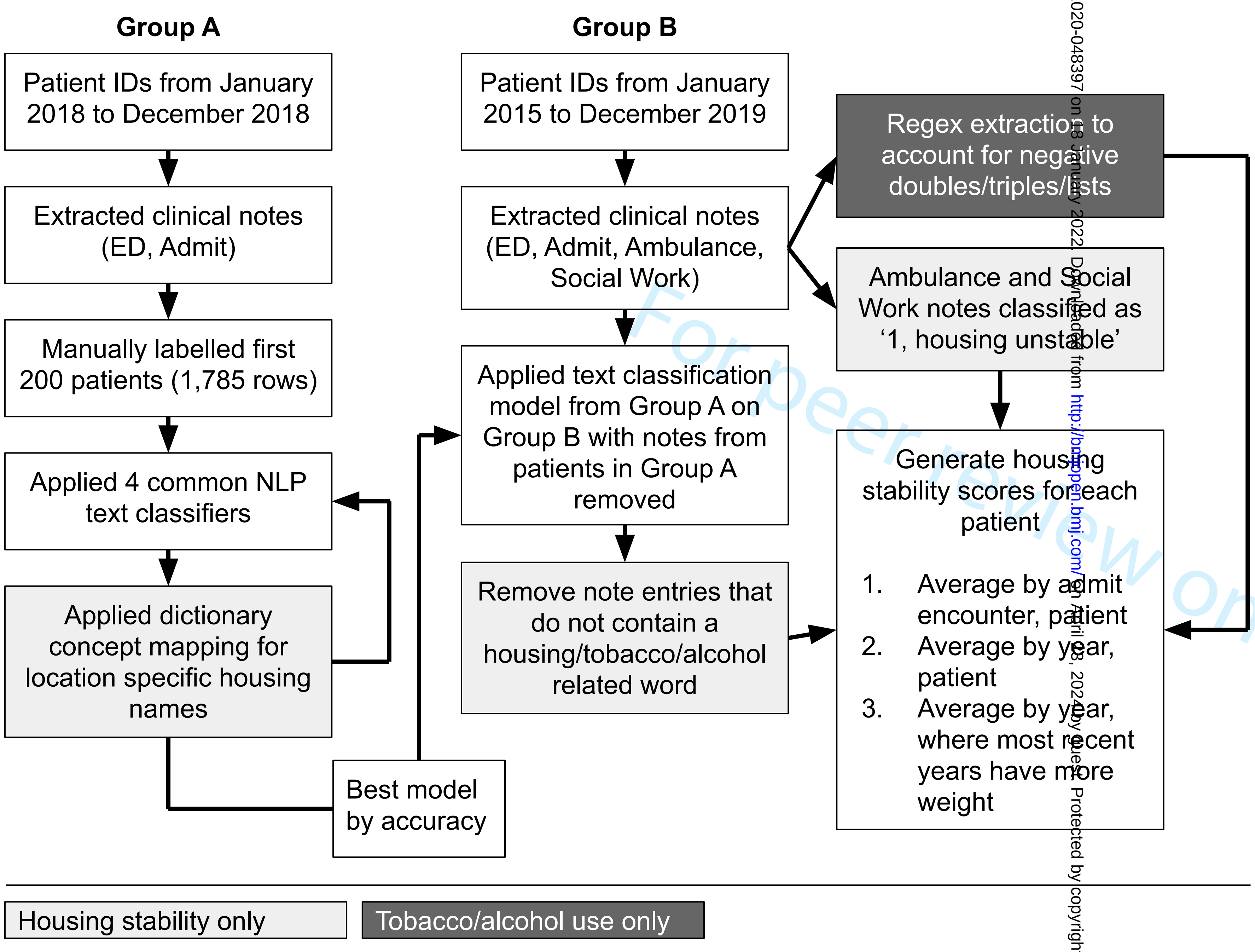
Figure legend:

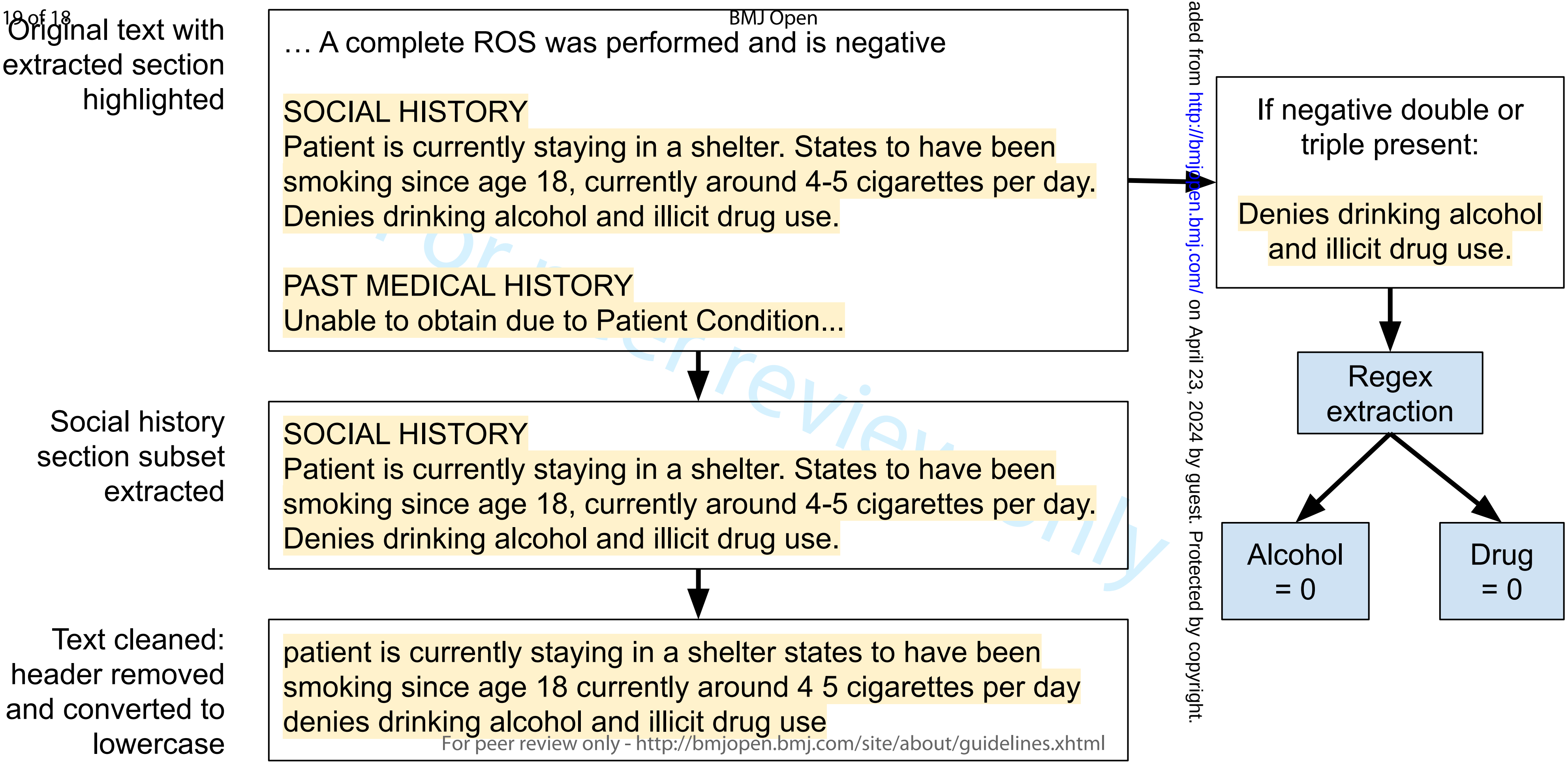
Figure 1: High-level overview of the workflow process

Figure 2: Text extraction, classification, and scoring workflow

Figure 3: Text extraction and cleaning process. Additional steps were performed for notes when classifying text related to tobacco and alcohol use to extract negative sentiment doubles or triples.


Data Extraction of Social Factors
Manual Labelling of Social Factors
Models Generation and Classification
Social Factor Score Creation
Structured and Unstructured Data Comparison


Group A
Group B
Patient IDs from January 2018 to December 2018
Extracted clinical notes (ED, Admit)
Manually labelled first 200 patients (1,785 rows)
Applied 4 common NLP text classifiers
Applied dictionary concept mapping for location specific housing names
Best model by accuracy
Patient IDs from January 2015 to December 2019
Extracted clinical notes (ED, Admit, Ambulance, Social Work)
Applied text classification model from Group A on Group B with notes from patients in Group A removed
Remove note entries that do not contain a housing/tobacco/alcohol related word
Regex extraction to account for negative doubles/triples/lists
Ambulance and Social Work notes classified as '1, housing unstable'
Generate housing stability scores for each patient
1. Average by admit encounter, patient
2. Average by year, patient
3. Average by year, where most recent years have more weight
Housing stability only
Tobacco/alcohol use only


Original text with extracted section highlighted
… A complete ROS was performed and is negative
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
PAST MEDICAL HISTORY
Unable to obtain due to Patient Condition...
Social history section subset extracted
SOCIAL HISTORY
Patient is currently staying in a shelter. States to have been smoking since age 18, currently around 4-5 cigarettes per day. Denies drinking alcohol and illicit drug use.
Text cleaned: header removed and converted to lowercase
patient is currently staying in a shelter states to have been smoking since age 18 currently around 4 5 cigarettes per day denies drinking alcohol and illicit drug use
If negative double or triple present:
Denies drinking alcohol and illicit drug use.
Regex extraction
Alcohol = 0
Drug = 0