MACHINE LEARNING APPROACHES TO IDENTIFYING SOCIAL DETERMINANTS OF HEALTH IN ELECTRONIC HEALTH RECORD CLINICAL NOTES

Rachel A. Stemerman

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Carolina Health Informatics Program in the Graduate School.

Chapel Hill
2020

Approved by:

Rebecca Kitzmiller

Ashok Krishnamurthy

Jaime Arguello

Lukasz Mazur

Jane Brice

ABSTRACT

Rachel Stemerman: Machine learning approaches to identifying social determinants of health in electronic health record clinical notes
(Under the direction of Rebecca Kitzmiller PhD)

Social determinants of health (SDH) represent the complex set of circumstances in which individuals are born, or with which they live, that impact health. Relatively little attention has been given to processes needed to extract SDH data from electronic health records. Despite their importance, SDH data in the EHR remains sparse, typically collected only in clinical notes and thus largely unavailable for clinical decision making. I focus on developing and validating more efficient information extraction approaches to identifying and classifying SDH in clinical notes. In this dissertation, I have three goals: First, I develop a word embedding model to expand SDH terminology in the context of identifying SDH clinical text. Second, I examine the effectiveness of different machine learning algorithms and a neural network model to classify the SDH characteristics financial resource strain and poor social support. Third, I compare the highest performing approaches to simpler text mining techniques and evaluate the models based on performance, cost, and generalizability in the task of classifying SDH in two distinct data sources.

ACKNOWLEDGEMENTS

I am greatly indebted to my adviser, Rebecca Kitzmiller, for her scientific expertise and professional mentorship. I would to thank her for their steadfast commitment to the research, and for her thoughtfulness and patience. Without her assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. Thank you for such a rewarding collaboration.

I would like to thank the members of my dissertation committee, including Dr. Ashok Krishnamurthy, Dr. Jaime Arguello, Dr. Jane Brice, and Dr. Lukasz Mazur. I would like to show gratitude to Dr. Brice, my first Medical Director at Orange County Emergency Medical Services, and constant supporter and mentor.

This study and my completion of a P.h.D. would not be possible without the funding and support of the National Library of Medicine Institutional Training Grants for Research Training in Biomedical Informatics. The support and friendship I have received from my NLM colleagues has been an opportunity of a lifetime and I hope to continue to contribute to this amazing community of health professionals and academics. No words can describe how appreciative I am for Dr. Javed Mostafa and the Carolina Health Informatics Program. Dr. Mostafa has been a guiding light for much of my doctoral journey.

I would also like to acknowledge the Department of Emergency Medicine at the University of North Carolina for being more than academic support and closer to family. Thank you for not just your open and honest feedback, but a shining example of a team collaborating to better the overall health of our patients. I need to express gratitude to Orange County Emergency Medical Services for their support

of my career and advancement of research in EMS. OCEMS is my family and given me the opportunity to serve a fantastic community.

Getting through my P.h.D. required more than academic support, and I have many, many people to thank for listening to, and at times, having to tolerate me over the past three years. Most importantly, none of this could have happened without my family. To my parents and brother thank you for trusting my path, to my aunt and uncle thank you for the dinners and conversations, and to my dog Rex who was always there to remind me to take a break for a walk. This dissertation stands as a testament to the unconditional love and encouragement from everyone in my life.

TABLE OF CONTENTS

CHAPTER 1: Introduction……1

Social determinants of health.....1

Research goals.....6

CHAPTER 2: DEVELOPING A WORD EMBEDDING MODEL.....14

Introduction.....14

Natural Language Processing: Word Embedding.....15

Methods.....17

Study design.....17

Setting and Subjects.....18

Data curation and preprocessing.....19

Manual semantic exploration.....20

Training word embedding for SDH.....21

Validate and expand notes likely to contain SDH documentation.....22

Dataset description.....23

Seed Terms.....24

Unigram Model.....30

Bigram Model.....32

Unigram vs Bigram 33

Systematic identification of SDH documentation 37

Combined unigram and bigram dataset 37

Error Analysis 42

Discussion 43

Limitations 46

Conclusion 47

CHAPTER 3: Identification of SDH using multi-label classification of EHR clinical notes 114

Introduction 114

Patient record labeling 116

Biomedical texts 117

Diagnosis code assignment 117

Methodology 118

Setting and sample population 118

Curation of social determinants of health 118

Gold-standard corpora purposeful sampling 119

Gold-standard annotation guidelines 120

Collection of clinical notes 120

Input Texts 121

Outcome variables 121

Experimental design 122

Evaluation 122

Results 128

Study population 128

Characteristics of gold-standard corpus 130

Features used for SDH label classification 132

Classifier performance 133

Error analysis 138

Discussion 140

CHAPTER 4: Identifying social determinants of health in clinical notes 161

Introduction 161

Background 163

Related work 163

Methods 165

Study corpus, annotation schema, and gold-standard development 165

Text pre-processing 166

Model development 166

SDH encoding 167

Rule-based system development 168

Machine learning approach 169

Results 171

Corpora characteristics 171

Text mining and ML model performance 173

Auto-encoding performance 174

Rule-based model performance 175

Feature selection for machine learning models 177

Machine learning model performance 177

Cost vs performance 181

Error analysis 182

Discussion 183

CHAPTER 5: Conclusion 198

References 206

LIST OF TABLES

CHAPTER 1……1

Table 1: Institute of Medicine recommendations for inclusion of SDH…….....8

CHAPTER 2……...14

Table 1: Literature review of SDH terms used for seed terms…….19

Table 2: Study population characteristics summary statistics…….23

Table 3: Description of dataset after removing copy and paste duplicates…….....25

Table 4: Manual identification of unigram word embedding expansion terms……...30

Table 5: Manual identification of bigram word embedding expansion terms…….....31

Table 6: Comparison of identified SDH terms by the unigrams and bigram model……...32

Table 7: Word embedding expansion approach……..35

CHAPTER 3……..…106

Table 1: Study population characteristics……..….120

Table 2: Inter-rater reliability for individual SDH classes…….……122

Table 3: Performance of models using five-fold cross validation…….124

Table 4: Performance of models inferring SDH labels using five-fold cross validation……...125

Table 5: Model performance (AUC) by SDH label……...…127

CHAPTER 4……..…147

Table 1: Manually annotated SDH characteristics……157

Table 2: n2c2 manually annotated SDH characteristics……157

Table 3: Overall performance of SDH models……...…158

Table 4: Auto-encoding model performance by SDH label…….….159

Table 5: Comparison of UNC data source model performance by SDH label……...…161

Table 6: Machine learning performance by SDH label…….……162

CHAPTER 5…….183

Table 1: Institute of Medicine recommendations for inclusion of SDH……....190

LIST OF FIGURES

CHAPTER 2……...14

Figure 1. Overview and workflow……...17

Figure 2. Included seed term evaluation……..24

Figure 3. Distribution of the count of SDH characteristics among race overlaid by sex…….36

Figure 4. Distribution of the count of SDH by age……..37

Figure 5a. Top 10 SDH term count……...38

Figure 5b. Top 5 term count per individual patient…….....38

Figure 6. SDH documentation over time……..39

CHAPTER 3…….106

Figure 1. Overview of methods for machine learning……...114

Figure 2. Evaluation metrics……..115

Figure 3. Bi-LSTM model layers……..….119

Figure 4. Count of SDH per sentence……....122

Figure 5. Correlation matrix of SDH labels……...……123

Figure 6. Average precision-recall score, micro-averaged over all SDH classes……..126

Figure 7. AUC-ROC for multiple classes……..128

Figure 8. Lexical diversity of SDH classes……130

CHAPTER 4……..…147

Figure 1. Encoding example……...…153

Figure 2. AUC-ROC for SDH labels……...……..164

Figure 3. Cost vs performance of UNC data source……...…166

LIST OF APPENDIXES

CHAPTER 2……...14

Appendix 1: Seed term evaluation……...50

Appendix 2: Unigram model word embedding expanded terms…….61

Appendix 3: Word embedding expansion manual observations evaluation……..…..65

Appendix 4: Bigram word embedding expanded terms…….……..74

Appendix 5: Terms found on observation after unigram word embedding expansion……78

Appendix 6: Terms found on observation after bigram word embedding expansion……..86

CHAPTER 3…….....106

Appendix 1: SDH annotation guidelines……...140

CHAPTER 4…….147

Appendix 1: SDH dictionary……..176

LIST OF ABBREVIATIONS

AUC-ROC	Area under the Receiver Operating Characteristic
AUPRC	area under the precision-recall curve
ED	emergency department
EHR	electronic health record
FN	false negative
ICD-10	International Classification of Diseases and Related Health Problems, Tenth Revision
IE	information extraction
IOM	Institute of Medicine
LOINC	Logical Observation Identifiers Names and Codes
MHSUD	mental health and substance use disorder
MIMIC	Multiparameter Intelligent Monitoring in Intensive Care
ML	machine learning
MLL	multi-label learning
n2c2	National NLP Clinical Challenges
NLP	natural languge processing
NLTK	Natural Lanaguage ToolKit
PPV	predictive postive value
SDH	social determinants of health
SME	subject matter expert
SNOMED-CT	Systematized Nomenclature of Medicine-Clinical Terms
SVM	Support Vector Machine
UMLS	Unified Medical Lanaguage System
UNC	University of North Carolina

CHAPTER 1: INTRODUCTION

Social determinants of health

Centers for Disease Control defines Social determinants of health (SDH) as "the complex, integrated, and overlapping social structures and economic systems that are responsible for most health inequities… [including] the social and physical environment, health services, and structural and societal factors"[1]. Substantial empirical evidence links specific social and behavioral factors to poor functional status and the onset and progression of disease[2–5]. Analysis conducted by McGinnis and Foege (1993) identified behaviors such as smoking, diet and activity, alcohol, and socioeconomic status as the causal contributors of premature mortality[3]. Link and Phelan (2004) argued that social conditions related to fewer socioeconomic resources such as money, social ties, and knowledge are fundamental causes of disease[5]. Social determinants, such as income and education, have wide-ranging effects across a person's life course. Higher-income is related to better health outcomes, including a lower prevalence of cardiovascular disease, diabetes, and depression as well as lower age-adjusted mortality[6]. The United States, despite ranking among the 10 richest countries in the world per capita, experiences sizable health disparities with an average 15-year difference in life expectancy between the most and least advantaged citizens[7].

The poor and disadvantaged suffer disproportionately from common mental health disorders and their adverse consequences[4]. Household income is one factor leading to common mental disorders; low educational attainment, material disadvantage, and unemployment[8]. In older adults, social isolation and limited social support are risk factors for mental health and substance use disorders (MHSUD)[4,9]. People

are made vulnerable to mental illness by deep-rooted poverty, social inequality, and discrimination[4]. According to the Substance Abuse and Mental Health Services Administration, 18% of adults in the United States struggle with any mental illness annually, this equates to approximately 46 million people[10]. Any mental illness, defined as a mental, behavioral, or emotional disorder, varies in impact ranging from no impairment to mild, moderate, or severe. Approximately 11.2 million of all U.S. adults suffer from serious mental illness, defined as a mental, behavioral, or emotional disorder that results in serious functional impairment and substantial interference with or limits to major life activities.

Patients with poor mental health, high psychological distress ratings, and depression have greater odds of frequent emergency department (ED) use[11]. Frequent ED users often seek care for a MHSUD related complaint [12–15]. The MHSUD population makes up one in eight ED visits[15]. High utilization contributes to ED overcrowding, poor quality of care, and high costs[16], yet the ED is often the last resort for vulnerable populations with poor care access. A pattern of repeated ED use for MHSUD may signal that local health and social services fail to meet a patient's needs[17]. In order to reduce negative health outcomes associated with SDH, health professionals will require information about their patients' individual SDH characteristics to better address their needs. For example, while ED physicians may attribute a patient's frequent visits for depression to poorly managed mental health issues because of a patient's unwillingness to follow up with specialists (i.e. willful noncompliance), might instead be caused by lack of transportation or financial constraints. Thus, medical treatment of a disease such as depression, without regard to the SDH, suffers the danger of being ineffective. Just as fluid volume overload cannot be treated without first understanding the physiology of the kidney, heart, lungs, and their interaction, a patient's MHSUDs treatment will be substandard without understanding associated SDH.

SDH and the electronic health record

The inclusion of SDH in the electronic health records (EHR) is vital since EHRs provide crucial information to support health professionals' treatment of individual patients, drive health system operations, provide insights about the health of the population, and guide researcher investigation. In 2014, the Institute of Medicine published two reports recommending specific social and behavioral-related measures for data collection in EHRs[2,18]. At a broad level, the SDH domains were divided into two categories: individual-level determinants specific to a patient (e.g., education level, employment status, or housing situation); and community-level determinants that measure socioeconomic, neighborhood, or environmental characteristics (e.g., air and water quality). A full description of these SDH domains appear in Table 1. The American College of Physicians[7] endorsement and federal initiatives supporting SDH data collection through EHRs, has assisted the evidenced-based creation of the Comprehensive Primary Care Plus (CPC +) model, Medicare Accountable Care Organizations (ACOs), and Accountable Health Communities (AHCs)[19].

Despite this level of interest, SDH information is neither routinely nor systematically collected in EHRs and lacks standardization[20–22]. Successfully implementing appropriate clinical decision support interventions within an EHR system and across different systems[23] depends upon standardized data and terminology. No single current biomedical standard (e.g., International Classification of Diseases and Related Health Problems, Tenth Revision [ICD-10]) captures the breadth of information necessary for documenting SDH in a format appropriate for clinical care, quality improvement, and research[24,25]. Several expert groups, including the National Academy of Medicine[18] (NAM) and the National Quality Forum[26] noted that a lack of standardized, interoperable terminology for SDH data collection and action in health care settings remains an obstacle to both scaling and studying SDH initiatives. As a result, existing clinical standards lack explicit translation algorithms to comprehensively integrate SDH as relevant clinical findings or problems into an EHR.

Recent studies suggest that if health professionals choose to document SDH data, they often do so in free text clinical notes[27–29]. Much of the EHR data is in free-text form and compared to structured data, free text is more natural and expressive method to document clinical events and facilitate communication among the care team[30]. Vest and colleagues found that SDH structured data was difficult to extract and varied across health IT systems[31]. While Feller and colleagues found higher performance among SDH models using structured and unstructured data, although performance was not statistically significantly higher than using text only[32]. A critical component to facilitate the use of EHR data for clinical decision support, quality improvement, or translational research is the information extraction (IE) task. IE, commonly recognized as a specialized area in empirical natural language processing (NLP), refers to the automatic extraction of concepts, entities, and events, as well as their relationships and associated attributes from free text[33]. Most IE systems are expert-based systems that consist of pattern identification to define lexical, syntactic, and semantic constraints[33]. Hybrid techniques that combine NLP and machine learning (ML) are the most common biomedical approach to extracting clinical text[34]. Multiple NLP methods have been successfully designed to identify various clinical diseases[28,35–37] and some SDH components such as opioid use[28], homelessness[29,38], and low socioeconomic status[31,38,39] in EHR data. However, methods for extracting a patients' complete SDH history from clinical text are less well developed and demonstrate that administrative codes are not sufficient to capture the breadth of SDH characteristics[21,29]. Additionally, IE systems' generalizability remains limited due to costs and time needed for close collaboration between a strong informatics team (including NLP experts) and clinicians with domain knowledge[33]. Most research in the informatics community focused on integrating SDH into the EHR using screening or referral tools[19,20,40]. However, infrequent documentation of SDH in the patient record and lack of national standards for collecting data related to SDH[22,32] impedes these efforts. Additionally, low adoption of EHR SDH clinical screening tools serves as an additional barrier to the collection of this information in a usable format[22,25].

Identifying SDH

Various studies effectively applied NLP information extraction techniques to different types of SDH domain classification including homelessness[29,35,48], employment status[35,49], and exposure to violence[29,50]. Bejan and colleagues constructed high performance (area under the precision-recall curve [AUPRC] of 0.94) word-embedding models for extracting homelessness-related words from clinical notes[29]. However, these models are unable to learn contextual information, such as where the homeless patient sleeps. Thus, Bejan and colleagues suggest word embedding models may best be applied in a prefiltering phase to significantly reduce the number of patients to analyze[29]. Feller and colleagues classified SDH as individual classes using only clinical notes and found that a gradient boosting tree algorithm, AdaBoost, performed the best (F1=79.2 for sexual orientation)[39]. Because deep learning models have been previously shown to successfully leverage clinical data for classification tasks[51–53], the authors attempted to train a neural network but yielded worse performance compared to the traditional machine learning approach due to the small training dataset size[39]. Previous studies have focused on classifying the patient or a specific document with the EHR as containing SDH characteristics, thus suffering information loss[1–3]. A limitation of this approach is that often SDH occur in clusters. For example, if a patients lost their job (i.e., employment insecurity), they may also lack insurance associated with employment. New approaches in ML, such as multi-label learning[1,2] (MLL) may be a viable candidate for modeling the profile of a patient affected by several SDH. To our knowledge, no study has developed an information extraction system to classify SDH on the sentence-level.

The general goal of classification problem is how to train a model based on training data such that the accuracy of predicting unseen test data is as high as possible. Prior research notes fairly low incident rates for many SDH domains[1–3], thus creating imbalance between positive and negative cases. A slight imbalance (e.g., 4:6) is often not a concern, and the problem can often be treated like a normal

classification modeling problem[1]. A severe imbalance (e.g. 1:100 or more) of the classes can be challenging to model and may require the use of specialized techniques. In many applications such as credit card fraud detection or adverse medication events, highly class-imbalanced problems are significant challenges as it is hard to detect rare but important cases successfully (e.g. 1:1000 up to 1:5000). Studies that assessed the presence of SDH characteristics in EHR unstructured data report a wide range of prevalence[29,35,40,41]. For example, similar incidence of housing instability related documentation were found by Navathe et al.[41] and Hatef et al.[40] (2% and 3%); however Hollister and colleagues[35] found much greater documentation of homelessness (13.7%). Therefore, in this study we experiment with developing our training dataset on a rich SDH data source, similar to oversampling technique.

Imbalanced classification problems occur when the event of interest are rare or the size of the interesting minority group is small proportion in the training data set[42]. This situation often creates a model with high accuracy because the results are overwhelmed by the majority class instances[43]. Re-sampling methods during pre-processing of data is popular among techniques proposed to deal with class imbalance problems[43,44]. Previous studies exploring approaches to identifying SDH in clinical notes have a common limitation of a small training set[35,39,45]. MHSUD patients access services frequently, especially through the ED, resulting in copious amounts of EHR data[46,47]. This suggests that patients who frequent the ED with MHSUDs may be a rich information source for developing a generalizable approach to identifying and extracting SDH.

Research goals

In this research, I investigated a series of machine learning and NLP approaches to identify and classify SDH using EHR clinical note data to achieve the following research goals:

1. Expand SDH terminology by developing word embedding models built upon existing terminology found in previously validated literature. Evaluate and compare the feasibility and

performance of different models to identify SDH characteristics related to financial resource strain among MHSUD patients who frequent the ED.

2. Develop and evaluate novel applications of MLL to classify the SDH domains financial resource strain and poor social support among MHSUD patient who frequent the ED.
3. Compare the performance, cost, and generalizability of three different information extraction techniques to automatically classify SDH characteristics using two distinct datasets.

This dissertation is organized as follows: In Chapter 2, I propose a word embedding model to expand SDH terminology in the context of identifying SDH in clinical text. In Chapter 3, the different multi-label machine learning algorithms and neural network model are proposed and evaluated in the task of classifying SDH. The highest performing approaches are compared to simpler text mining techniques and evaluated based on performance, cost, and generalizability in the task of classifying SDH within two datasets (UNC and open source in Chapter 4. The conclusions and contributions to knowledge are summarized in Chapter 5.

TABLE 1: Institute of Medicine recommendations for inclusion of SDH

Individual Factors	**Sociodemographic**	**Sexual orientation**
		Racial identity
		Ethnic identity
		Country of origin/migration history
		Education
		Employment
		Financial resource strain food and housing insecurity
		Health literacy
	Psychological	Depression/anxiety
		Stress
		Optimism/Self-efficacy/Patient empowerment/activation/engagement
	Behavioral	Dietary patterns
		Physical activity
		Tobacco use and exposure
		Alcohol use
	Individual-level social relationships and living conditions	Social isolation and social connections
		Exposure to violence
Neighborhoods/ Communities	Compositional characteristics	Neighborhood and community compositional characteristics

REFERENCES

1. Social Determinants of Health | Healthy People 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health. Accessed March 13, 2020.

2. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing social and behavioral domains and measures in electronic health records: phase 2*. Washington (DC): National Academies Press (US); 2015. doi:10.17226/18951

3. McGinnis JM, Foege WH. Actual causes of death in the United States. *JAMA*. 1993;270(18):2207-2212.

4. Allen J, Balfour R, Bell R, Marmot M. Social determinants of mental health. *Int Rev Psychiatry*. 2014;26(4):392-407. doi:10.3109/09540261.2014.928270

5. Phelan JC, Link BG. Fundamental social causes of disease and mortality. In: *Encyclopedia of health and behavior*. 2455 Teller Road, Thousand Oaks California 91320 United States : SAGE Publications, Inc.; 2004. doi:10.4135/9781412952576.n102

6. Adler NE, Glymour MM, Fielding J. Addressing social determinants of health and health inequalities. *JAMA*. 2016;316(16):1641-1642. doi:10.1001/jama.2016.14058

7. Daniel H, Bornstein SS, Kane GC, Health and Public Policy Committee of the American College of Physicians. Addressing social determinants to improve patient care and promote health equity: an american college of physicians position paper. *Ann Intern Med*. 2018;168(8):577-578. doi:10.7326/M17-2441

8. Fryers T, Melzer D, Jenkins R, Brugha T. The distribution of the common mental disorders: social inequalities in Europe. *Clin Pract Epidemiol Ment Health*. 2005;1:14. doi:10.1186/1745-0179-1-14

9. Pantell M, Rehkopf D, Jutte D, Syme SL, Balmes J, Adler N. Social isolation: a predictor of mortality comparable to traditional clinical risk factors. *Am J Public Health*. 2013;103(11):2056-2062. doi:10.2105/AJPH.2013.301261

10. Mental Health in America - Adult Data | Mental Health America. http://www.mentalhealthamerica.net/issues/mental-health-america-adult-data. Accessed June 2, 2018.

11. Smith JL, De Nadai AS, Storch EA, Langland-Orban B, Pracht E, Petrila J. Correlates of length of stay and boarding in florida emergency departments for patients with psychiatric diagnoses. *Psychiatr Serv*. 2016;67(11):1169-1174. doi:10.1176/appi.ps.201500283

12. Ondler C, Hegde GG, Carlson JN. Resource utilization and health care charges associated with the most frequent ED users. *Am J Emerg Med*. 2014;32(10):1215-1219. doi:10.1016/j.ajem.2014.07.013

13. Brennan JJ, Chan TC, Hsia RY, Wilson MP, Castillo EM. Emergency department utilization among frequent users with psychiatric visits. *Acad Emerg Med*. 2014;21(9):1015-1022. doi:10.1111/acem.12453

14. Chang G, Weiss AP, Orav EJ, Rauch SL. Predictors of frequent emergency department use among patients with psychiatric illness. *Gen Hosp Psychiatry*. 2014;36(6):716-720. doi:10.1016/j.genhosppsych.2014.09.010

15. Krieg C, Hudon C, Chouinard M-C, Dufour I. Individual predictors of frequent emergency department use: a scoping review. *BMC Health Serv Res*. 2016;16(1):594. doi:10.1186/s12913-016-1852-1

16. Data on behavioral health in the United States. http://www.apa.org/helpcenter/data-behavioral-health.aspx. Accessed June 2, 2018.

17. Urbanoski K, Cheng J, Rehm J, Kurdyak P. Frequent use of emergency departments for mental and substance use disorders. *Emerg Med J*. 2018;35(4):220-225. doi:10.1136/emermed-2015-205554

18. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing social and behavioral domains in electronic health records: phase 1*. Washington (DC): National Academies Press (US); 2014. doi:10.17226/18709

19. Freij M, Dullabh P, Lewis S, Smith SR, Hovey L, Dhopeshwarkar R. Incorporating social determinants of health in electronic health records: qualitative study of current practices among top vendors. *JMIR Med Inform*. 2019;7(2):e13849. doi:10.2196/13849

20. Gold R, Cottrell E, Bunce A, et al. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *J Am Board Fam Med*. 2017;30(4):428-447. doi:10.3122/jabfm.2017.04.170046

21. Bettencourt-Silva JH, Mulligan N, Sbodio M, et al. Discovering New Social Determinants of Health Concepts from Unstructured Data: Framework and Evaluation. *Stud Health Technol Inform*. 2020;270:173-177. doi:10.3233/SHTI200145

22. Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. *Health Aff (Millwood)*. 2018;37(4):585-590. doi:10.1377/hlthaff.2017.1252

23. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Med*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

24. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. *JAMIA Open*. 2019;2(1):81-88. doi:10.1093/jamiaopen/ooy051

25. Gottlieb L, Tobey R, Cantor J, Hessler D, Adler NE. Integrating social and medical data to improve population health: opportunities and barriers. *Health Aff (Millwood)*. 2016;35(11):2116-2123. doi:10.1377/hlthaff.2016.0723

26. National Quality Forum. *Risk adjustment for socioeconomic status or other sociodemographic factors*. (National Quality Forum, ed.). washington, DC: National Quality Forum; 2014.

27. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. *J Acquir Immune Defic Syndr*. 2018;77(2):160-166. doi:10.1097/QAI.0000000000001580

28. Carrell DS, Cronkite D, Palmer RE, et al. Using natural language processing to identify problem usage of prescription opioids. *Int J Med Inform*. 2015;84(12):1057-1064. doi:10.1016/j.ijmedinf.2015.09.002

29. Bejan CA, Angiolillo J, Conway D, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc*. 2018;25(1):61-71. doi:10.1093/jamia/ocx059

30. Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform*. 2018;87:12-20. doi:10.1016/j.jbi.2018.09.008

31. Vest JR, Grannis SJ, Haut DP, Halverson PK, Menachemi N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int J Med Inform*. 2017;107:101-106. doi:10.1016/j.ijmedinf.2017.09.008

32. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data. *Appl Clin Inform*. 2020;11(1):172-181. doi:10.1055/s-0040-1702214

33. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review. *J Biomed Inform*. 2018;77:34-49. doi:10.1016/j.jbi.2017.11.011

34. Mishra R, Bian J, Fiszman M, et al. Text summarization in the biomedical domain: a systematic review of recent research. *J Biomed Inform*. 2014;52:457-467. doi:10.1016/j.jbi.2014.06.009

35. Doan S, Maehara CK, Chaparro JD, et al. Building a Natural Language Processing Tool to Identify Patients With High Clinical Suspicion for Kawasaki Disease from Emergency Department Notes. *Acad Emerg Med*. 2016;23(5):628-636. doi:10.1111/acem.12925

36. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21(2):221-230. doi:10.1136/amiajnl-2013-001935

37. Wieneke AE, Bowles EJA, Cronkite D, et al. Validation of natural language processing to extract breast cancer pathology procedures and results. *J Pathol Inform*. 2015;6:38. doi:10.4103/2153-3539.159215

38. Hollister BM, Restrepo NA, Farber-Eger E, Crawford DC, Aldrich MC, Non A. Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. *Pac Symp Biocomput*. 2017;22:230-241. doi:10.1142/9789813207813_0023

39. Casey JA, Pollak J, Glymour MM, Mayeda ER, Hirsch AG, Schwartz BS. Measures of SES for Electronic Health Record-based Research. *Am J Prev Med*. 2018;54(3):430-439. doi:10.1016/j.amepre.2017.10.004

40. Bazemore AW, Cottrell EK, Gold R, et al. Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. *J Am Med Inform Assoc*. 2016;23(2):407-412. doi:10.1093/jamia/ocv088

41. 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.

42. Lindemann EA, Chen ES, Rajamani S, Manohar N, Wang Y, Melton GB. Assessing the Representation of Occupation Information in Free-Text Clinical Documents Across Multiple Sources. *Stud Health Technol Inform*. 2017;245:486-490.

43. Blosnich JR, Marsiglio MC, Dichter ME, et al. Impact of Social Determinants of Health on Medical Conditions Among Transgender Veterans. *Am J Prev Med*. 2017;52(4):491-498. doi:10.1016/j.amepre.2016.12.019

44. Tran T, Kavuluru R. Predicting mental conditions based on “history of present illness” in psychiatric notes with deep neural networks. *J Biomed Inform*. 2017;75S:S138-S148. doi:10.1016/j.jbi.2017.06.010

45. Sulieman L, Gilmore D, French C, et al. Classifying patient portal messages using Convolutional Neural Networks. *J Biomed Inform*. 2017;74:59-70. doi:10.1016/j.jbi.2017.08.014

46. Du J, Chen Q, Peng Y, Xiang Y, Tao C, Lu Z. ML-Net: multi-label classification of biomedical texts with deep neural networks. *J Am Med Inform Assoc*. 2019;26(11):1279-1285. doi:10.1093/jamia/ocz085

47. Zufferey D, Hofer T, Hennebert J, Schumacher M, Ingold R, Bromuri S. Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Comput Biol Med*. 2015;65:34-43. doi:10.1016/j.compbiomed.2015.07.017

48. Zhang M-L, Zhou Z-H. A Review on Multi-Label Learning Algorithms. *IEEE Trans Knowl Data Eng*. 2014;26(8):1819-1837. doi:10.1109/TKDE.2013.39

49. Navathe AS, Zhong F, Lei VJ, et al. Hospital Readmission and Social Risk Factors Identified from Physician Notes. *Health Serv Res*. 2018;53(2):1110-1136. doi:10.1111/1475-6773.12670

50. Hatef E, Rouhizadeh M, Tia I, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system. *JMIR Med Inform*. 2019;7(3):e13802. doi:10.2196/13802

51. Krawczyk B. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell*. 2016;5(4):221-232. doi:10.1007/s13748-016-0094-0

52. Lin S-C, Wang C, Wu Z-Y, Chung Y-F. Detect Rare Events via MICE Algorithm with Optimal Threshold. In: *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. IEEE; 2013:70-75. doi:10.1109/IMIS.2013.21

53. Agrawal A, Viktor HL, Paquet E. SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling. In: *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. SCITEPRESS - Science and and Technology Publications; 2015:226-234. doi:10.5220/0005595502260234

54. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *jair*. 2002;16:321-357. doi:10.1613/jair.953

55. Winden TJ, Chen ES, Monsen KA, Wang Y, Melton GB. Evaluation of flowsheet documentation in the electronic health record for residence, living situation, and living conditions. *AMIA Jt Summits Transl Sci Proc*. 2018;2017:236-245.

56. Hudon C, Sanche S, Haggerty JL. Personal characteristics and experience of primary care predicting frequent use of emergency department: A prospective cohort study. *PLoS One*. 2016;11(6):e0157489. doi:10.1371/journal.pone.0157489

57. Kahan D, Leszcz M, O'Campo P, et al. Integrating care for frequent users of emergency departments: implementation evaluation of a brief multi-organizational intensive case management intervention. *BMC Health Serv Res*. 2016;16:156. doi:10.1186/s12913-016-1407-5

CHAPTER 2: DEVELOPING A WORD EMBEDDING MODEL APPROACH FOR SOCIAL DETERMINANTS OF HEALTH TERMINOLOGY EXPANSION

Introduction

Despite the significant impact of social determinants of health (SDH) on health outcomes, health care professionals rarely have standardized tools available to systematically collect and incorporate information about SDH factors into clinical decision making, program interventions, and policy initiatives[1,2]. Risk factors for many common mental disorders are significantly associated with social inequalities, whereby the greater the social inequality the higher the risk[3]. For instance, housing instability and increased homelessness occurs disproportionately among mental health and substance use disorder (MHSUD) who frequent the emergency department [4,5]. MHSUD patients who frequent the ED (≥ 4 visits in one year[6]) express a variety of other financial resource constraints including an inability to purchase medications[4,6,7], provide transportation[8], or secure food[4]. As a result, the Institute of Medicine identified SDH domains that should be included in electronic health records (EHRs)[1] and the Meaningful Use program that incentivizes a range of SDH data collection including housing stability, social isolation, and support, and employment[9,10]. This mandate prompts the need to develop standardized vocabulary and data collection tools.

The extent to which SDH are encoded in EHRs remains largely unknown[11–14]. EHRs generally lack formalized, structured SDH data collection fields, thus even if health professionals collect the data, it is impossible to systematically capture this data[15]. If health professionals choose to document SDH data, they often do so in free text clinical notes[11,13,16]. Compared to structured data, free-text is a more natural and expressive method used to document clinical events and facilitate communication among the care

team[17]. Structured data fields that specifically capture SDH may be limited because many SDH characteristics do not fit the traditional structure of EHR data collection, a structure originally designed to facilitate accurate billing and procedures[14]. When compared to free-text clinical notes, structured data such as diagnostic codes can be unreliable because they lack overall clinical context[18]. The vast amount of SDH characteristics remain buried in the clinical notes and, therefore, remain largely unusable.

Few studies have explored health professionals SDH documentation practices[2,11,12,14,19]. To facilitate the meaningful use of EHR data towards improving the overall quality of care, it is vital to understand how SDH are represented in the EHR to describe the syntactic variability of SDH in clinical notes, define SDH terminology as a precursor to SDH integration into decision support tools and programmatic interventions. Due to the nature of clinical natural language, the terms and phrases that describe SDH in clinical notes often have a wide variety of syntactic and semantic variability. For example, Hatef et al[2] found SDH documentation generated from health professionals of different clinical and non-clinical backgrounds contributed to the varied linguistic expression of SDH[2]. Existing terminologies such as Unified Medical Language System (UMLS) and Logical Observation Identifiers Names and Codes (LOINC) exhibit diminished performance when extracting SDH because the language used to express these concepts is often regional and idiosyncratic[19]. These limitations suggest that a natural language processing (NLP) may be the most effective method with which to extract SDH semantic variants in EHR clinical notes. Identification of SDH characteristics in clinical notes is required to develop downstream applications, such as information extraction, which will serve as an initial step for future SDH data collection and tools.

Natural Language Processing: Word Embedding

Clinical natural language processing (NLP) is a text analysis approach applied to documents written by health professionals in clinical settings, such as EHR notes and reports [20]. Although early clinical NLP approaches were primarily rule-based, recent statistical NLP methods (e.g. machine

learning) have demonstrated superior performance on clinical information extraction tasks such as classification[18,21,22]. Machine learning based approaches rely on high-quality, manual annotation of the clinical corpora[23,24] by clinical experts. Generating this 'gold-standard', is time consuming, expensive and labor intensive task[20,24]. Machine learning-based pre-annotation is built upon training the model on a small amount of annotated text[25]. A question remains as to whether the machine learning model is necessary in these cases, or whether a simple dictionary-based pre-annotation set is sufficient. Lindgren and colleagues concluded dictionary-based pre-annotation is a feasible and practical method to reduce the cost of annotation of clinical named entity recognition in the eligibility sections of clinical trial announcements without introducing bias in the annotation process[25].

Recent work on unsupervised distributional semantics for corpus expansion on clinical notes[2,11,19,26] suggests that word embedding analysis can capture the semantic properties and linguistic relationships between words using deep neural networks[27]. This method is based on distributional semantics, in which word similarity is estimated based on word distribution found in the entire data set, or corpus. Distributional semantics assumes that words with similar meanings tend to occur in similar contexts.[28] Compared with NLP methods that rely on gold-standard human annotated training data, word embedding models are more efficient and scalable since they can be trained on a large amount of unannotated data.[27,29] Two word embedding models, Word2vec[28] and GloVe[30], have been successfully applied on a variety of NLP tasks, such as named entity recognition[31–33] and text classification[19,34] in the healthcare domain.

Word2vec employs a shallow neural network that incrementally iterates over a training corpus to develop a model. [28] Rather than evaluating a single term, or unigram, word2vec determines frequency of relevant terms within an entire patient record to identify representations of a word or term, such as homelessness. Word2vec generates word vectors using two different language models schemes: continuous bag of words (CBOW) and skip-gram[28]. The goal of the CBOW method is to predict a word

given the surrounding words, whereas in skip-gram, given a single word, a window or context of words is predicted. Bejan et al[11] employed a skip-gram word2vec model to identify and rank terms associated with homelessness among patients in an EHR repository (N=185) and achieved an overall precision of 93%. Alternatively, GloVe[26] learns word representations by examining word-word co-occurrence, to determine how frequently words occur together. GloVe adds practical meaning into word vectors by considering the relationships between bigrams rather than unigrams. However, because the model is trained on the co-occurrence matrix of words it uses significant memory, making similarity comparisons within a large corpus time-consuming.

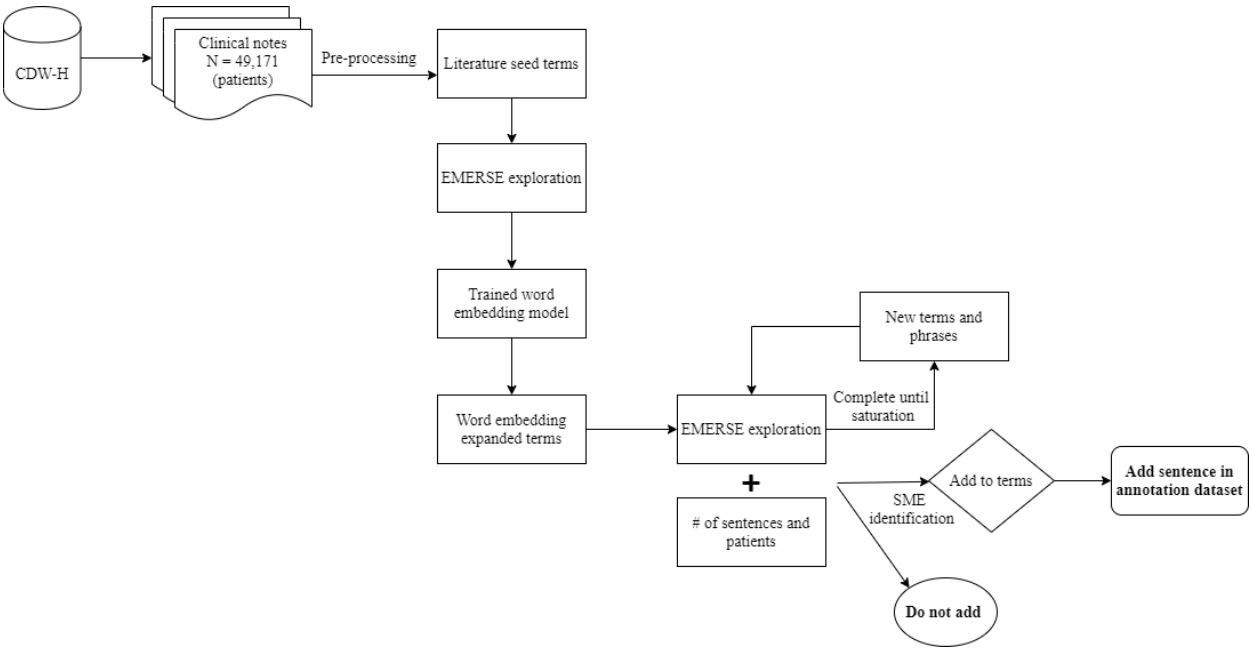
In a single study comparing the performance of word2vec and GloVE for terminology expansion, Fan et al[26] found that word2vec more effectively detected semantically similar terms than GloVe when applied to infrequently occurring terms. Muneeb et al[35] used biomedical literature to compare word2vec and GloVe and found that a word2vec skip-gram model performed best when compared to other models in a semantic similarity task. Thus, word2vec may prove more effective for detecting semantically or syntactically similar SDH terms and phrases in clinical notes. We aim to expand SDH terminology by developing word embedding models built upon existing terminology found in previously validated literature and evaluate the feasibility and performance of the model to identify SDH characteristics related to financial strain among MHSUD patients who frequent the ED.

Methods

Study design

The study was carried out in three steps outlined as follows: (1) collect and preprocess clinical notes; (2) train word vectors using a word2vec word embedding model; (3) conduct intrinsic and extrinsic evaluation. The overview and workflow of the method is shown in Figure 1.

FIGURE 1. Overview and workflow



Setting and Subjects

Clinical notes from April 2014 to December 2019 were collected from the Carolina Data Warehouse for Health (CDW-H), a centralized data repository containing clinical, research, and administrative data sourced from the University of North Carolina Health Care System. This timeframe begins with the health system's transition to a single EHR. Clinical notes that met the following inclusion criteria were retained: (1) visited University of North Carolina at Chapel Hill Emergency Department (UNC-CH ED) between 2014-2019, (2) greater than 18 years old in the CDW-H as of 2014, and (3) documented MHSUD "final primary diagnosis" as defined by their International Classification of Diseases and Related Health Problems 10th Revision (ICD-10 CM) code F00-F99 mental and behavioral disorders. Patients who had less than four ED visits in the year 2017 or 2018 within a rolling 365-day period were excluded because they do not meet the criteria for a frequent user of the ED[6]. Institutional review board (IRB) approval was obtained for accessing the clinical notes.

Data curation and preprocessing

This query yielded 49,171 patients and contained over 14 million clinical notes. Preprocessing of the collected corpus included lowercasing, removal of metadata, punctuation (except periods), numbers, single letter words, and stop words. Stop words are extremely common words (i.e. "but", "or", "what", etc.) that appears to have little value in helping select sentences matching our SDH task were excluded from the vocabulary completely[36]. The National Language Toolkit (NLTK) library, one of the oldest and most commonly used Python libraries for NLP, English stop word list was used for this study[37]. The corpus was tokenized and empty arrays removed. Tokenization is the process of splitting longer string of text into tokens or single words (unigram). All notes were separated into sentences, identified by the presence of a period and placed in a Pandas DataFrame indexed by the patient's medical record number (MRN) and clinical note array. Pandas is a fast and efficient DataFrame object for data manipulation with integrated indexing for doing practical, real world data analysis in Python.

Seed term curation

In this step, I performed a literature review of studies of SDH among MHSUD patients frequently treated in emergency departments that use informatics techniques to identify SDH terms or phrases in clinical notes. The results of this review generated a list of 34 terms and/or phrases describing the SDH characteristics financial resource strain and poor social support, used as seed terms, to initially explore the dataset (Table 1).

TABLE 1: Literature review of SDH terms used for seed terms

SDH domain	SDH component	Unigram	Reference
Financial resource strain	Employment insecurity	incarceration	Bejan et al (2018)
Financial resource strain	Employment insecurity	jail	Bejan et al (2018)
Financial resource strain	Employment insecurity	jobless	Bejan et al (2018)
Financial resource strain	Employment insecurity	police	Bejan et al (2018)
Financial resource strain	Employment insecurity	prison	Bejan et al (2018)
Financial resource strain	Employment insecurity	prisoner	Bejan et al (2018)
Financial resource strain	Employment insecurity	prostitute	Bejan et al (2018)
Financial resource strain	Employment insecurity	prostitution	Bejan et al (2018)
Financial resource strain	Employment insecurity	retired	Hollister et al (2017)

Financial resource strain	Employment insecurity	trespassing	Bejan et al (2018)
Financial resource strain	Employment insecurity	unemployed	Bejan et al (2018), Hollister et al (2017)
Financial resource strain	Employment insecurity	unemployment	Bejan et al (2018), Hollister et al (2017)
Financial resource strain	General	afford	UMLS
Financial resource strain	General	lack	UMLS
Financial resource strain	General	welfare	Bejan et al (2018)
Financial resource strain	Housing	homelessness	Hatef et al (2019) Hollister et al (2017)
Financial resource strain	Housing	inadequate housing	Hatef et al(2019) UMLS
Financial resource strain	Housing insecurity	cluttered	Winden et al (2018)
Financial resource strain	Housing insecurity	evicted	Bejan et al (2018)
Financial resource strain	Housing insecurity	eviction	UMLS
Financial resource strain	Housing insecurity	excessive	Winden et al (2018)
Financial resource strain	Housing insecurity	homeless	UMLS, Bejan et al (2018), Hollister et al (2017)
Financial resource strain	Housing insecurity	houseless	Hatef et al (2019)
Financial resource strain	Housing insecurity	housing	UMLS
Financial resource strain	Housing insecurity	inadequate	Winden et al (2018)
Financial resource strain	Housing insecurity	motel	Bejan et al (2018)
Financial resource strain	Housing insecurity	shelter	Bejan et al (2018)
Financial resource strain	Housing insecurity	streets	Bejan et al (2018)
Financial resource strain	Housing insecurity	transitional	UMLS
Financial resource strain	Insurance	medicaid	Hollister et al (2017)
Financial resource strain	Insurance	uninsured	Hollister et al (2017)
Financial resource strain	Social support	lack of assistance	Hatef et al (2019)
Financial resource strain	Social support	lives alone	Hatef et al (2019)
Financial resource strain	Social support	no family support	Hatef et al (2019)
Financial resource strain	Social support	no social support	Hatef et al (2019)

Manual semantic exploration

We then used the terms identified above to explore clinical notes within our dataset using Electronic Medical Record Search Engine (EMERSE). EMERSE, developed at the University of Michigan in 2005 and launched at UNC-Chapel Hill in 2017, allows researchers to search unstructured EHR clinical notes. A 4% (N=2000) randomized sample of the study population's clinical notes were manually reviewed using EMERSE until all seed terms were analyzed. We used frequency of occurrence, context, and observational true positives to determine which terms to retain. For each clinical note

selected through EMERSE, a SME (Author RS) read the entire clinical note to identify new unigrams or phrases that described SDH characteristics among MHSUD who frequent the ED. Once a patients' clinical note was reviewed the patient's entire record was removed from the corpus to reduce bias and oversampling[12, 13].

Training word embedding for SDH

Finally, we used unsupervised distributional semantics for corpus expansion on clinical notes[2,11,19,26].We used the skip-gram model of word2vec[28]. For each word *(w)* in a training corpus *(T)* that belong to a vocabulary *(V)* over the text collection, this model learns a word embedding *(v)* and a context embedding *(c)* that are useful for predicting the surrounding words in the context window of *w*. For example, for the word at position *j*, w_j and the word at position *k*, w_k (in the context of w_j), the task is to compute $p(w_k|w_j)$. The skip-gram model computes this probability by passing the dot product between the target vector of w_j, v_j and the context of w_k,C_k through a softmax function[28]:

$$p(wk|wj) = \frac{\exp(c\frac{T}{k} - Vj)}{\sum_{i=1}^{|V|}\exp(c\frac{T}{i} - Vj)}$$

The objective was to find the word embedding that maximized the sum over log probabilities of all context words given each current target word from the collection.

We trained word2vec's skip-gram model on our entire corpus. For model configuration, we used hierarchical softmax to approximate the full softmax function, a vector dimension of 300, and context window sizes of 5. A minimum count of 5 defined the threshold frequency value that was needed to be reached for the word to be included in the model. Models were trained for unigrams and bigrams. Once the word embeddings were learned, we used the cosine similarity metric to measure the similarity between vectors corresponding to all words in *V* and vectors associated with the seeds in *S*. The model parameters were chosen based on similar studies expanding terminology in the biomedical domain[11,26,35].

Validate and expand notes likely to contain SDH documentation

Next, we invested significant effort into a better understanding of how SDH are documented in clinical text and improving the process for retrieving SDH from the EHR. One primary assessor (Author RS) in consultation with SMEs (i.e., ED physicians, clinical social workers, psychiatrists) identified SDH semantically related terms through manual examination of the retrieved patient notes. For each SDH seed term input, the trained Word2vec model returned a list of the top 5 semantically related terms with their cosine similarity scores. Expert judgment was used to evaluate these semantically related terms and for each clinical note selected for review, the assessor analyzed the entire clinical note. Terms that appeared to be true positive on a sentence-level were collected for manual annotation (Chapter 3). The patient's record was used only once. During this evaluation the assessor added additional terms and phrases that described SDH and were extrinsically evaluated. This process continued until saturation, no new terms were discovered[38]. Terms and phrases were categorized following IOM definitions[1]. Extrinsic evaluation included counting the number of patients records that contained the SDH term or phrase and counting the number of sentences within the collection that contained the SDH term or phrase.

To derive an unbiased estimate of likely SDH documentation, we addressed auto-generated and copy and paste entries that appeared to duplicate sentences. The frequency of redundant text in clinical notes presents a number of challenges for training and evaluating text mining models[39]. This task included the removal of any sentence that was the exact same as another sentence within an individual patient's clinical record regardless of the time period between the occurrences of these entries. This process was completed for both unigram and bigram models to minimize over-representation of SDH.

In addition, we conducted an error analysis to gain insight into dictionary performance for SDH financial resource strain characteristics. This was performed by reviewing a 4% random sampling of the patients who did not have an identified SDH sentence in their clinical record notes. The patient's entire

clinical record notes were reviewed for the following characteristics: (1) idiosyncratic language used to express SDH, (2) unrecognized negation (3) syntactic dependencies and (4) misspellings.

Results

Dataset description

This dataset included 49,171 patients, a total of 14,948,813 clinical notes, averaging 304.03 (±1439.4) notes per patient and 3,706.14 (±6212.6) sentences per patient. Each clinical note was associated with one clinical documentation by a health professional. Clinical documentation may be best described as an event with multiple clinical notes generated daily or multiple times a day by different health professionals over a period of time. The corpus included a total of 169,177,769 sentences, 1,096,166 contained instances of SDH. Of the 49,171 patients, 38,971 (79.3%) had at least one sentence that likely represented an SDH characteristic, 25,290 (51.4%) patients had at least 5 sentences.

The study population was 50.8 (±20.8) years of age, primarily White or Caucasian 73.1% (N= 35,938), and female 52.7% (N=25,888). A complete description of the study population can be found in Table 2.

TABLE 2: Study population characteristics summary statistics

Characteristic	n (%)	Male	Female
All	49171 (100%)	23729 (47.3%)	25888 (52.7%)
Race			
American Indian or Alaska Native	180 (0.4%)	75 (0.2%)	105 (0.2%)
Asian	333 (0.7%)	132 (0.3%)	201 (0.4%)
Black or African American	9559 (19.4%)	4867 (10.0%)	4692 (9.5%)
NAIS Race	4 (0.0%)	3 (0.0%)	1 (0.0%)
Native Hawaiian or other Pacific Islander	16 (0.0%)	9 (0.0%)	7 (0.0%)
Other race	2087 (4.2%)	1139 (2.3%)	948 (2.0%)
Other/hispanic	1 (0.0%)	0 (0.0%)	1 (0.0%)
Patient refused	53 (0.1%)	25 (0.1%)	28 (0.1%)
Unknown	999 (2.0%)	415 (0.8%)	583 (1.2%)
White or Caucasian	35938 (73.1%)	16614 (33.8%)	19322 (39.3%)
Age			
18-29	9449 (19.2%)	4584 (9.3%)	4865 (9.9%)

30-39	7989 (16.2%)	4109 (8.4%)	3880 (7.9%)
40-49	7114 (14.5%)	3535 (7.2%)	3579 (7.3%)
50-59	8197 (16.7%)	4147 (8.4%)	4050 (8.2%)
60-69	6464 (13.1%)	3281 (6.7%)	3183 (6.5%)
70-79	4409 (9.0%)	1798 (3.7%)	2611 (5.3%)
>80	5545 (11.3%)	1825 (3.7%)	3720 (7.6%)
Clinical notes per patient	14948813 (100%)	6842021 (45.8%)	8106052 (54.2%)
SDH documentation per patient	548138 (100%)	304067 (55.5%)	243788 (44.5%)

Seed Terms

We performed literature review of studies that identified social determinants of health in clinical notes using information extraction techniques to yield a total of 35 terms characterizing financial resource strain and poor social support[2,11–13,40]. These terms were categorized into housing insecurity (14), insurance insecurity (2), general financial insecurity (3), employment/income insecurity (12), and poor social support (4).

In the manual review of N=2000 records using EMERSE to calculate term frequency, we found that SDH documentation was most likely generated by a Social Worker (21.1%, N=21) or occurred in an "unnamed note" (20.2%, N=20) within the EMERSE system. Certain unigram terms were poorly related to SDH (e.g. transitional, lack), but were more appropriate as a bigram or phrase such as "transitional housing" and "lack of transportation" (Figure 2). Appendix 1 contains the evaluation of all seed terms.

Manual review of sentences in the dataframe identified a significant number of auto-generated sentences and copy and paste documentation. To reduce over representation, the final dataset for every model included only unique instances of a sentence.

FIGURE 2. Included seed term evaluation

SDH Domain	Unigram/Phrase	Reference	Example from EMERSE	Note Type	Corpus Frequency	Number of Patients
Housing insecurity	streets	Bejan et al. (2017)	"Exposure/Witness to Violence: Yes; patient reports living on streets in DC and saw "a lot of dead bodies" and saw his "best friend shot dead.""	Psychiatry emergency service initial consult	5743	1890
			"Type of Residence: Homeless: on the streets."	Social work psychosocial assessment		
			"Patient reports"I hope next time I see you it'll be on the streets.""	Unnamed note		
nsurance insecurity	uninsured	Hollister et al. (2017)	"Pt appears uninsured"	CM screening assessment	11161	3348
			"uninsured"	Unnamed note		
			"Pt is uninsured and patient's fiance has reported that the mechanism of burn was likely due to methamphetamine use."	Progress note		
Poor social support	lives alone	Hatef et al. (2019)	"Clinical Risk Factors: Multiplie Diagnoses (Chronic), Lives Alone or Absence of Caregiver to Assist with Discharge and Home Care, Functional Limitations"	Care management initial transition planning assessment	44411	8025
			"Pt lives alone."	Crisis and assessment psychologist assessment		
			"She reports that he lives alone with his dog, states that he is in a wheelchair, lost his legs and fingers, TBI from Army deployment to Afghanistan 6 years ago. "	Family therapy progress note		
General financial insecurity	afford	UMLS	"Patient is requesting one of our doctors to re-write the RX so the medication can be covered under her Charity Care as she is unable to afford it."	Patient request note	39769	10298
			"Reports that he was doing well until he could not afford medication refills (including buprenorphine for relapse prevention)."	Psych emergency service initial consult		
			"Was prescribed Ranexa but has been unable to afford it."	Follow-up visit		

After manual inspection by experts, 23 unigrams and 2 bigrams were selected as the foundation for our word embedding expansion models. Trigrams or larger were not explored in this study because a previous study found that trigram and four-gram approaches performed worse than the bigram model when identifying relevant new information in clinical notes[41]. The seed term modeling alone identified 34,442 patients (66.0%) and 412,592 sentences likely containing SDH characteristics. The top terms found by the seed term model were *Medicaid* (N=15,376, 31.3%) and *afford* (N=10,298, 20.9%) (Table 3). The terms with largest percentage of change after removing copy and paste documentation were *trespassing* (-38.2%), *unemployed* (-36.4%), and *prostitution* (-34.4%). The terms with the lowest percentage of change were *uninsured* (-12.5%), *Medicaid* (-13.5%), and *shelter* (-13.9%). Table 4 shows

a complete description of terms found by all of the models and the counts before and after removal of duplicate sentences created by copy and paste documentation practices.

TABLE 3: Description of dataset after removing copy and paste duplicates

Type of financial insecurity	Unigram or Phrase	Original Count	Original Num. Pts.	Final Count	Percent of Change	Percent of Population
Employment	receives ssdi	716	263	390	-45.50%	0.50%
Employment	trespassing	3369	540	2081	-38.20%	1.10%
Employment	shoplifting	2316	334	1470	-36.50%	0.70%
Employment	prostitution	1026	193	673	-34.40%	0.40%
Employment	unemployment	9575	1723	6510	-32.00%	3.50%
Employment	job loss	1412	427	975	-30.90%	0.90%
Employment	stole	12625	3214	8834	-30.00%	6.50%
Employment	receives ssi	1143	532	827	-27.60%	1.10%
Employment	lost job	1514	576	1103	-27.10%	1.20%
Employment	jail	31883	5343	23864	-25.20%	10.90%
Employment	forging	229	30	178	-22.30%	0.10%
Employment	probation	19941	5788	15519	-22.20%	11.80%
Employment	receives disability	2428	1186	1921	-20.90%	2.40%
Employment	prison	29336	4359	23339	-20.40%	8.90%
Employment	prostitute	659	238	535	-18.80%	0.50%
Employment	parole	3603	1382	2932	-18.60%	2.80%
Employment	disability income	1682	849	1373	-18.40%	1.70%
Employment	court date	19430	6535	16104	-17.10%	13.30%

Employment	veteran	5649	1845	4955	-12.30%	3.80%
Employment	taken into custody	203	128	181	-10.80%	0.30%
Food	food stamps	3923	1693	3404	-13.20%	3.40%
Food	food insecure	27	17	24	-11.10%	0.00%
Food	food insecurity	3193	1598	2976	-6.80%	3.20%
Food	food pantries	337	212	321	-4.70%	0.40%
General	subsidized	947	323	599	-36.70%	0.70%
General	financial stressors	4657	1266	3259	-30.00%	2.60%
General	income	86800	18033	61004	-29.70%	36.70%
General	charges	48355	10776	34956	-27.70%	21.90%
General	financial constraints	1739	734	1296	-25.50%	1.50%
General	disabled	17864	4744	13610	-23.80%	9.60%
General	does not drive	7975	3497	6323	-20.70%	7.10%
General	bankruptcy	1197	599	997	-16.70%	1.20%
General	unable to make payments	6	6	5	-16.70%	0.00%
General	financially	5184	2644	4347	-16.10%	5.40%
General	afford	39769	10298	33983	-14.50%	20.90%
General	lack of transportation	3264	1605	2814	-13.80%	3.30%
General	financial issues	12893	6251	11149	-13.50%	12.70%
General	affordable	4029	2051	3516	-12.70%	4.20%
General	transportation problems	309	202	272	-12.00%	0.40%
General	financial concerns	6929	3949	6514	-6.00%	8.00%
General	lack of resources	1698	1163	1614	-4.90%	2.40%
Housing	cluttered	1174	356	680	-42.10%	0.70%
Housing	inadequate housing	221	33	143	-35.30%	0.10%

Housing	hoarder	410	132	274	-33.20%	0.30%
Housing	lack of stable housing	238	85	159	-33.20%	0.20%
Housing	lost her home	217	99	149	-31.30%	0.20%
Housing	unstable housing	2763	714	1919	-30.50%	1.50%
Housing	banned	796	282	578	-27.40%	0.60%
Housing	evicted	3741	958	2729	-27.10%	1.90%
Housing	eviction	2340	617	1706	-27.10%	1.30%
Housing	payments	11987	4525	8792	-26.70%	9.20%
Housing	evict	6836	1489	5122	-25.10%	3.00%
Housing	landlord	4766	1081	3574	-25.00%	2.20%
Housing	streets	5743	1890	4431	-22.80%	3.80%
Housing	homelessness	28462	3339	22505	-20.90%	6.80%
Housing	infested	520	243	414	-20.40%	0.50%
Housing	homeless	101832	6749	81221	-20.20%	13.70%
Housing	motel	5583	1933	4455	-20.20%	3.90%
Housing	flooded	636	299	512	-19.50%	0.60%
Housing	pay rent	593	337	485	-18.20%	0.70%
Housing	transitional housing	856	302	710	-17.10%	0.60%
Housing	hotel	10655	2465	8849	-16.90%	5.00%
Housing	oxford house	22248	1504	18491	-16.90%	3.10%
Housing	boarding house	1877	364	1571	-16.30%	0.70%
Housing	public housing	370	162	310	-16.20%	0.30%
Housing	housing issue	547	283	465	-15.00%	0.60%
Housing	lost his home	192	123	164	-14.60%	0.30%

Housing	shelter	47546	4680	40955	-13.90%	9.50%
Housing	foreclosed	167	86	147	-12.00%	0.20%
Housing	durham rescue	3052	548	2703	-11.40%	1.10%
Housing	pay mortgage	9	6	8	-11.10%	0.00%
Housing	rescue mission	8594	1481	7663	-10.80%	3.00%
Housing	housing crisis	37	14	34	-8.10%	0.00%
Housing	housing insecure	1011	752	990	-2.10%	1.50%
Housing	lack of satisfaction with housing	1	1	1	0.00%	0.00%
Housing	mortgage assistance	8	8	8	0.00%	0.00%
Housing	staying ifc	0	0	0	0.00%	0.00%
Insurance	lost insurance	342	184	252	-26.30%	0.40%
Insurance	copay	19291	5219	14827	-23.10%	10.60%
Insurance	no insurance	8831	4343	7183	-18.70%	8.80%
Insurance	medicaid	107273	15376	92841	-13.50%	31.30%
Insurance	uninsured	11161	3348	9766	-12.50%	6.80%
Insurance	cheaper	1772	1182	1611	-9.10%	2.40%
Insurance	unc charity	3940	1802	3616	-8.20%	3.70%
Insurance	charity care	20634	4907	18954	-8.10%	10.00%
Insurance	self pay	2211	1218	2104	-4.80%	2.50%
Insurance	pays out of pocket	2216	1504	2173	-1.90%	3.10%
Insurance	selfpay	17	16	17	0.00%	0.00%
Social Support	lack of assistance	90	43	60	-33.30%	0.10%
Social Support	limited social support	10983	1919	7350	-33.10%	3.90%
Social Support	no social support	1060	349	757	-28.60%	0.70%
Social Support	no social supports	355	146	267	-24.80%	0.30%
Social Support	lives alone	44411	8025	35421	-20.20%	16.30%
Social Support	no family support	456	252	365	-20.00%	0.50%
Social Support	lack of support	13952	3893	12514	-10.30%	7.90%

Social Support	lack of caregivers	1057	755	1038	-1.80%	1.50%

Unigram Model

The unigram model found a total of 47 SDH terms or phrases within 35,749 patient records and 722,231 sentences. The copy and paste reduction task removed 146,372 sentences leaving 575,859 sentences in the unigram derived corpus. The unigram model led to 15 terms that were then manually inspected and determined to be true positive mentions. Appendix 2 lists the top 5 semantically similar terms identified using this method. During expert manual inspection of each clinical note, an additional 32 terms or phrases were identified (Table 4). Appendix 3 contains the complete list of terms that were manually inspected with highlighted terms indicating they were added to the SDH dictionary. The word2vec model identified various forms of misspelling for specific financial resource strain terms such as *Medicaid*.

TABLE 4: Manual identification of unigram word embedding expansion terms

SDH Domain	Unigram/Phrase	Reference	Example from EMERSE	Note Type	Corpus Frequency	Number of Patients
general	disabled	expansion	"He feels that he cannot work a regular job due to his level of physical incapacitation and needs father to understand that he is disabled and needs time and patience to get better.",	Psychiatry follow-up consult	13445	4744
			"Pt is disabled and does not drive."	Social work psychosocial assessment		
housing	landlord	expansion	"increasing stressors related to daughter now moved back with pt who " already have a lot" medical issues related to eye problems, also financial stressors , has to move out sec to issues with landlord",	Psychiatric note	4766	1081
			"Pt. stated that she rescued two baby goats when their landlord, who lives on the adjoining property, went on a trip and left them unsheltered and unfed to die in the winter.",	Abuse consult		
			"States he was in process of moving when had accident and has been in hospital and unable to get rent and deposit money to new landlord"	Unnamed note		
			"States he was in process of moving when had accident and has been in hospital and unable to get rent and deposit money to new landlord"	Unnamed note		
employment/income	stole	expansion	"Patient also reported another patient stole her gold rings."	Unnamed note	12625	3214
			"She also becomes a "kleptomaniac" and this last time stole something from her aunt and was caught."	Psychiatry initial consult		
			"Today law enforement was dispatched to residence for a larceny report, he had stole a tv and pawed it for crack cocaine."	Unnamed note		

The unigram model found the terms *cluttered* (-42.1%), *trespassing* (-38.2%), and *subsidized* (-36.7%) had the highest percentage of change after removing duplicate sentences. The unigrams with the lowest percentage of change were *cheaper* (-9.1%), *foreclosed* (-12.0%), and *veteran* (-12.3%). However, in the overall dictionary created by the unigram model process the term or phrase with the lowest percentage of change were *lack of caregivers* (-1.8%), *pays out of pocket* (-1.9%), and *housing insecure* (-2.1%). Of the 68 SDH phrases (unigram and seed terms), 5 of them were found in over 15% of the study population: *income* (36.7%, N=18,033), *medicaid* (31.3%, N=15,376), *charges* (21.9%, N=10,776),

afford (20.9%, N=10,298), and *lives alone* (16.3%, N=8,025). Sentence-level evaluation of the corpus revealed that all unigram terms appeared with at least one other identified SDH seed or expansion term.

Bigram Model

The bigram model led to an additional 34 terms that were manually inspected and determined to be true positives. Appendix 4 lists the top 5 semantically similar terms identified by this method. During expert manual inspection, an additional 31 terms or phrases were identified (Table 5). The bigram model process found an additional 65 terms or phrases (derived from the 23 seed terms) that were likely indicators of SDH characteristics that identified 38,748 patients and 873,935 sentences. The copy and paste reduction task removed 173,819 sentences leaving 718,379 sentences in the bigram derived corpus.

TABLE 5: Manual identification of bigram word embedding expansion terms

Bigram	Sample notes	Clinical Note Type
currently_homeless	"he is currently homeless"	Clinical Note Type
	"Patient is currently homeless"	Emergency department provider note
	"Living situation: the patient is currently homeless, recently lost his home of 20 years due to be unable to make payments"	Crisis and assessment services initial assessment
homeless_shelter	"The pt was recently discharged from 1North on 11/08 to a homeless shelter"	Behavioral health assessment team
	"Post Acute Facility: Homeless shelter, Substance Abuse Treatment Facility, Other (Oxford House.)"	Care management: continued transition planning assessment
	"She reports when she was unable to get into the homeless shelter."	UNC Wakebrook primary care initial consult note
lives_alone	"Pt states that she lives alone, but daughter and granddaughter live nearby (5-10 minutes) and could come by if she needed them to."	Physical therapy
	"Pt lives alone"	N/A
	"Living situation: the patient lives alone."	Palliative care consult
Oxford_house	"Question/Concern for provider (Specific): Patient is in an Oxford House facility and his mother wanted to let Dr. X know that 3 of the guys there with him have been diagnosed with MRSA and she would like a call for advice. "	Patient advice request message
	"Other prior treatment(s): AA, referral to Oxford House after WakeBrook discharge"	Rex psychiatry initial consult
	"The patient was accepted for enrollment into an Oxford House (a substance abuse recovery house) on 04/17/2017,"	UNC health care addictions detoxification unit at wakebrook psychiatric discharge note

The top terms found exclusively by the bigram model were *oxford house* (3.1%, N= 22,248) and *financial issues* (12.7%, N=12,893). The bigrams *receives ssdi* (-45.5%) and *unstable housing* (-30.5%) had the highest percentage of change after removing duplicate sentences. The bigrams with the lowest percentage of change were *mortgage assistance* (0.0%, N=8) and *unc charity* (-8.2%, N= 3,940). Additional information on the descriptive statistics for the bigram model is provided in Table 3.

Unigram vs Bigram

When comparing the SDH characteristics identified by the unigram and bigram model, the models shared 377,787 sentences. The bigram model uncovered 12 of the same terms that were identified

by the unigram word embedding semantic expansion (Table 6). An additional 15 terms found during bigram expert manual inspection were also identified by the unigram model process. This led to the bigram model identifying 2,999 more patients than the unigram model (Table 7).

TABLE 6: Comparison of identified SDH terms by the unigram and bigram model

SDH domain	Term	Seed Term	Unigram	U_id	Bigram	B_id
Housing	Homeless	X				
Housing	Shelter	X				
Housing	Streets	X				
Housing	Motel	X				
Housing	Evicted	X				
Housing	homelessness	X				
Housing	inadequate housing	X				
Housing	transitional housing	X				
Housing	cluttered	X				
General	afford	X				
Insurance	uninsured	X				
Insurance	medicaid	X				
Employment/income	prison	X				
Employment/income	jail	X				
Employment/income	prostitute	X				
Employment/income	prostitution	X				
Employment/income	trespassing	X				
Employment/income	unemployed	X				
Social Support	lives alone	X				
Social Support	no family support	X				
Social Support	no social support	X				
Social Support	lack of assistance	X				
Social Support	no social supports	X				
Housing	eviction			X		X
Housing	foreclosed		X			X
Housing	landlord		X		X	
Housing	banned		X			
Housing	lack of satisfaction with housing			X		
Housing	housing insecure			X		
Housing	hoarder			X		
Housing	housing crisis			X		

Housing	housing issue		X		
Housing	hotel		X	X	
Housing	evict			X	
Housing	oxford house			X	
Housing	durham rescue			X	
Housing	infested			X	
Housing	pay rent			X	
Housing	staying ifc			X	
Housing	unstable housing			X	
Housing	pay mortgage			X	
Housing	payments			X	
Housing	flooded			X	
Housing	public housing			X	
Housing	boarding house			X	
Housing	rescue mission			X	
Housing	lost her home				X
Housing	lost his home				X
Housing	lack of stable housing				X
Housing	mortgage assistance				X
General	affordable		X		X
General	disabled	X		X	
General	lack of resources		X		
General	financial stressors		X		
General	financial concerns		X		
General	financially		X		X
General	subsidized		X	X	
General	income	X		X	
General	charges	X		X	
General	lack of transportation		X		X
General	finances	X		X	
General	financial strain			X	
General	medication affordability			X	
General	unable to make payments				X
General	bankruptcy			X	
General	transportation problems				X
General	financial constraints				X
General	mortgage assistance				X
General	financial issues				X
General	does not drive				X

Insurance	charity care		X	X	
Insurance	pays out of pocket		X		X
Insurance	self pay		X		X
Insurance	no insurance		X		X
Insurance	lost insurance		X		X
Insurance	copay	X			
Insurance	cheaper	X			
Insurance	unc charity			X	
Insurance	selfpay			X	
Employment/income	stole	X			
Employment/income	disability income		X	X	
Employment/income	unemployment	X			
Employment/income	veteran	X		X	
Employment/income	forging	X			
Employment/income	shoplifting	X			
Employment/income	probation	X		X	
Employment/income	parole		X	X	
Employment/income	taken into custody		X		
Employment/income	court date		X		
Employment/income	DUI		X		
Employment/income	job loss		X		X
Employment/income	lost job			X	
Employment/income	receives disability			X	
Employment/income	receives ssi			X	
Employment/income	receives ssdi			X	
Employment/income	losing job				X
Employment/income	loss of job				X
Employment/income	not employed				X
Employment/income	difficulty maintaining employment				X
Employment/income	employment difficulties				X
Food	food pantries		X		X
Food	food stamps		X		X
Food	food insecure		X		X
Food	food insecurity		X		X
Social support	limited social support		X		
Social support	lack of caregivers		X		
Social support	lack of support		X		
Social support	lack of social support				X
Social support	poor social support				X

**U_id and B_id refer to unigram and bigram manual identification by the SME*

Systematic identification of SDH documentation

A total of 14,948,813 clinical notes containing 43,487,049, 154 tokens were used to train the word embedding models in this study. The unigram and bigram word embedding expanded terms (semantic variants) for financial resource strain including housing insecurity, employment/income insecurity, and general financial insecurity are shown in Appendix 2 and Appendix 4. The best results of our SME driven word embedding expansion approach were achieved with the combined terms identified by the unigram and bigram model (Table 7). In total, this process yielded 111 terms or phrases that characterized financial resource strain and poor social support. This count did not include terms that would be found if stemming or lemmatization was completed such as shelter and shelters.

TABLE 7: Word embedding expansion approach

Approach	# of patients	% of population	# of sentences
seed terms	32,442	66.0%	412,592
seed terms + unigram	35,749	72.7%	722,231
seed terms + bigram	38,748	78.8%	873,935
seed terms + unigram + bigram	38,971	79.3%	1,096,166

Combined unigram and bigram dataset

SDH documentation across patients were varied with a mean of 11.2 (±27.6) per patient; the 25% quartile contained 0 instances of SDH documentation. Figure 3 depicts the count distribution of SDH among race, overlaid by sex using an outlier box plot. Females are slightly more representative than males in this study population (52.7% vs 47.3%). Overall, more females (79.1%) the males (77.9%) have at least one SDH documentation instance. However, males have more total SDH sentences (382,406, 53.2%) than females (336,017, 46.8%). Males have a higher rate of SDH documentation regardless of race ((Caucasian male (35.1%) vs female (32.6%); African American male (15%) vs female (11.2%). While African American's make up 19.4% of the total study population, they account for 26.2% of SDH documentation as opposed to White or Caucasian (73.1% of study population) who, account for 67.8% of SDH

documentation. The vast majority of patients, regardless of age had < 100 instances of SDH documentation. Those with the greatest amount of SDH documentation occurred in the 30 – 65 age range as shown in Figure 4.

FIGURE 3: Distribution of the count of SDH characteristics among race overlaid by sex

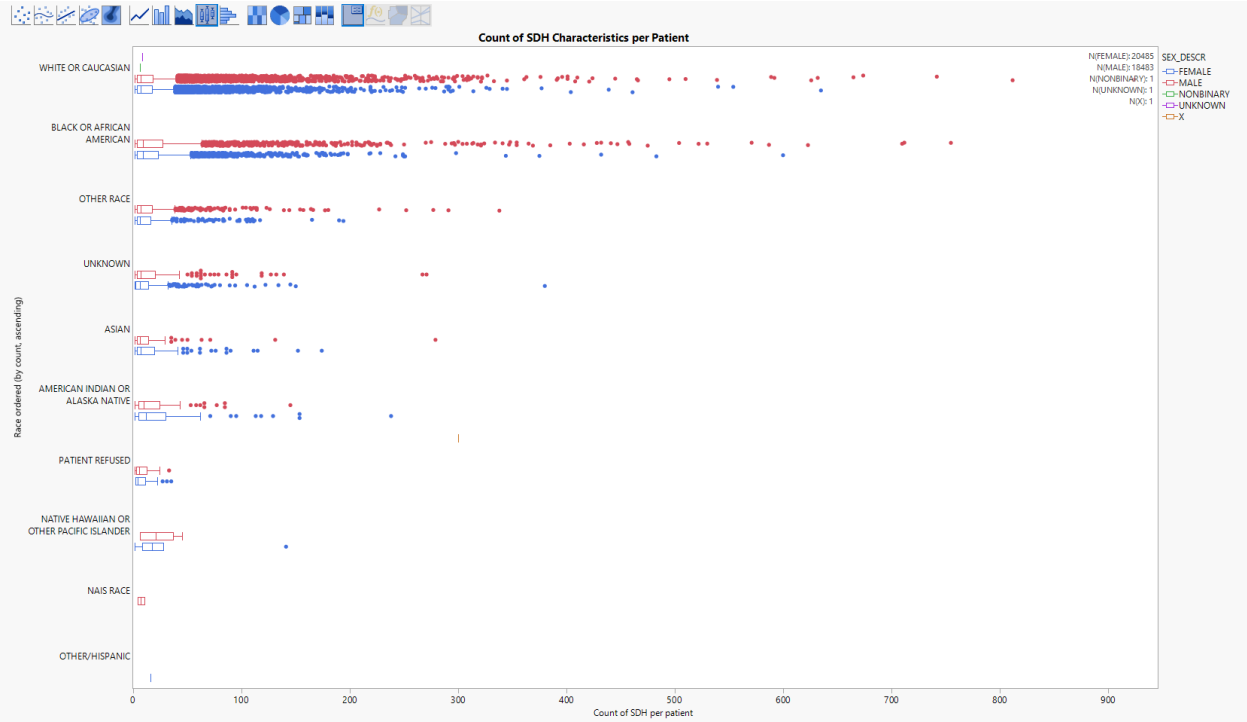
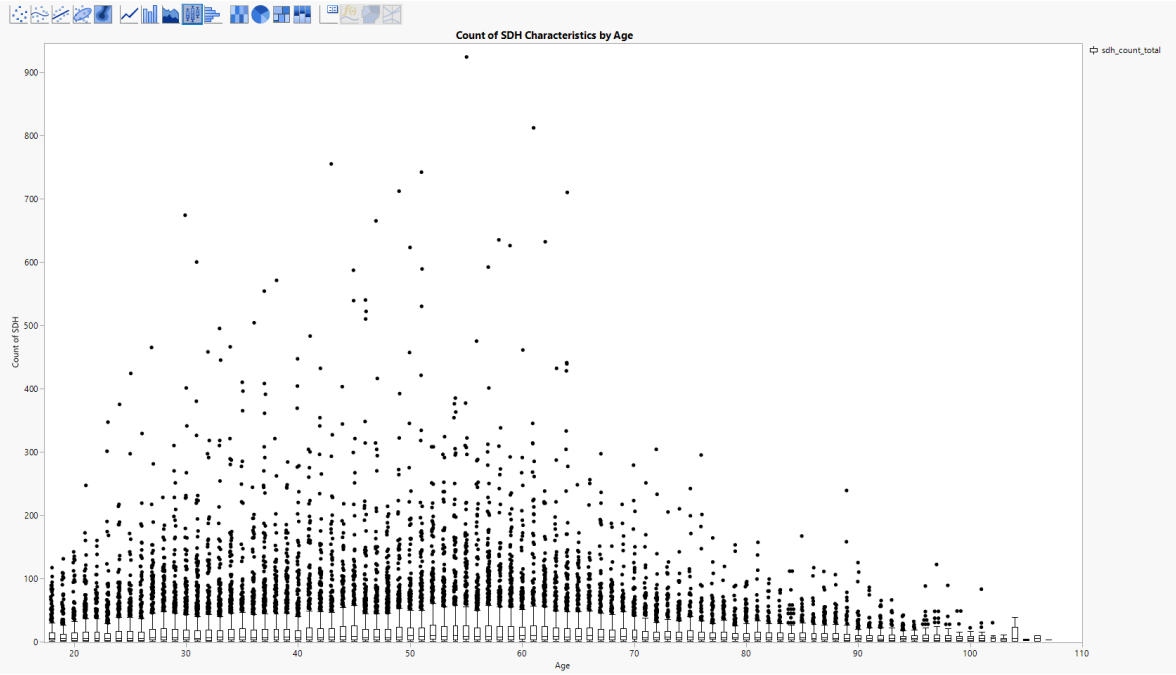
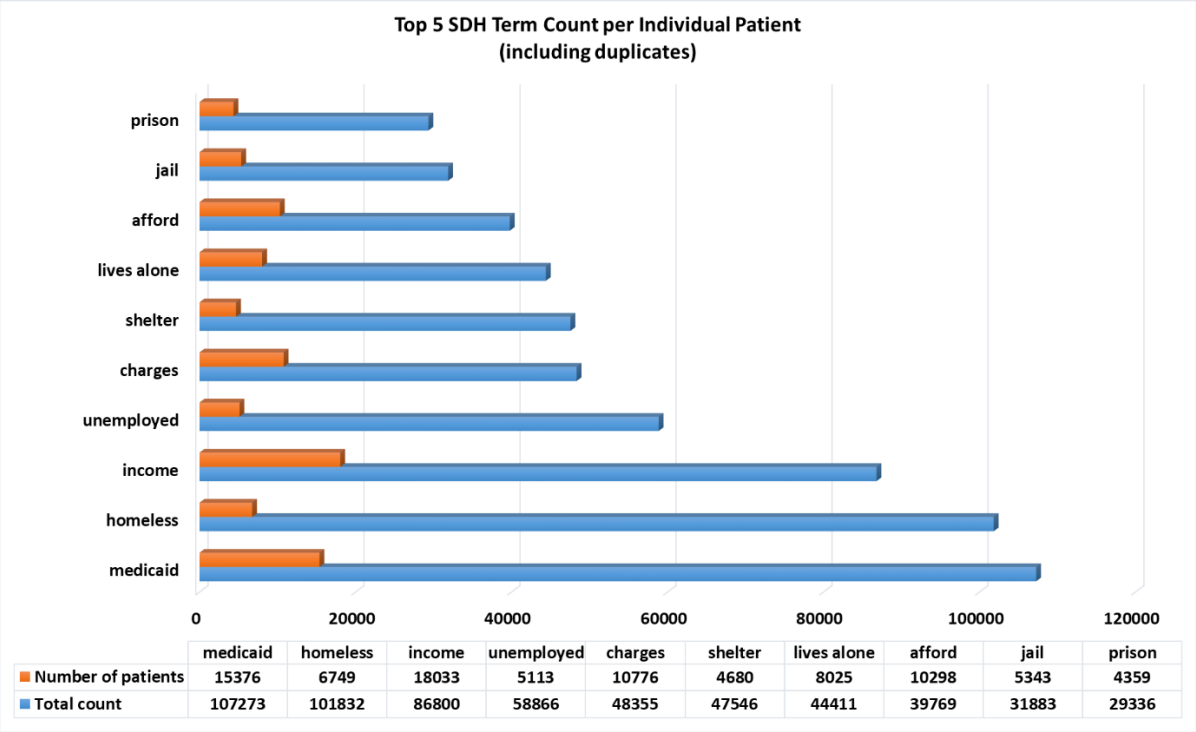


FIGURE 4: Distribution of the count of SDH by age



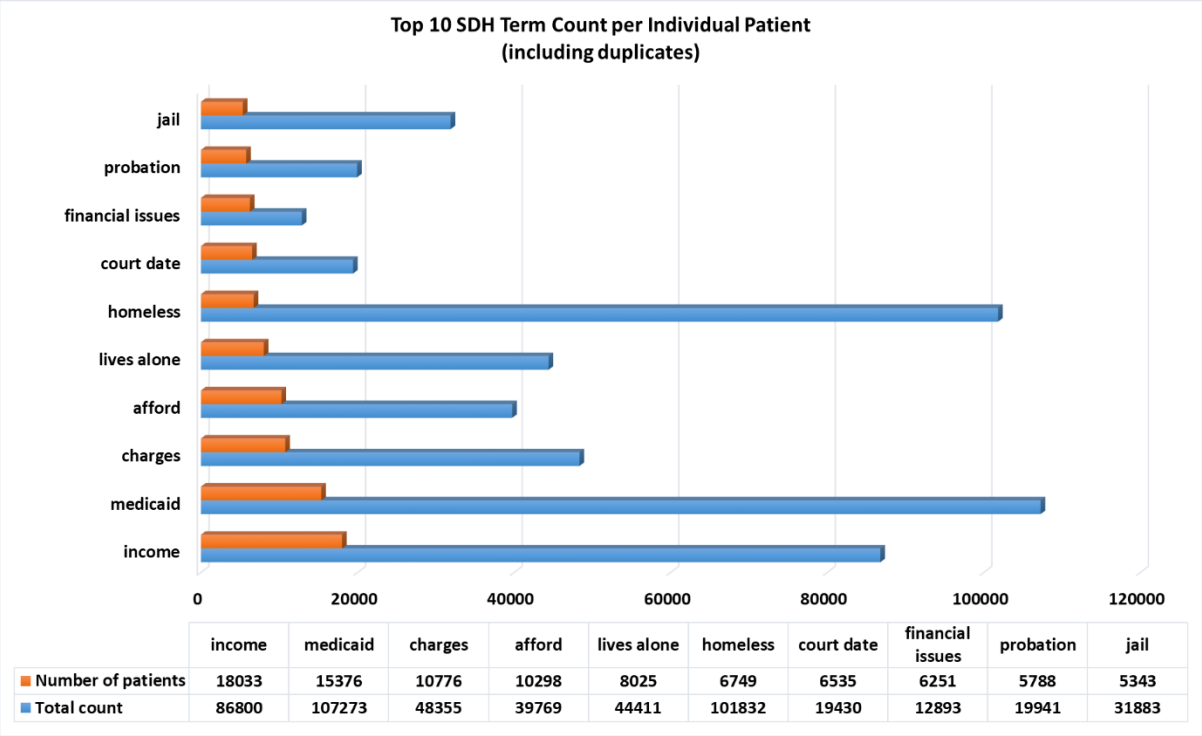
The terms that occurred most frequently in the total corpus were *Medicaid* (107,273) and *homeless* (101,832) (Figure 5a). Additionally, *Medicaid (*15,376) was second in term count per individual patients with *income* (18,033) as the top individual patient count term (Figure 5b). The terms that appeared most frequently in the same sentence were *homeless* and *shelter* (10,240) then *probation* and *court date* (3,885).

FIGURE 5a: Top 10 SDH term count



	medicaid	homeless	income	unemployed	charges	shelter	lives alone	afford	jail	prison
Number of patients	15376	6749	18033	5113	10776	4680	8025	10298	5343	4359
Total count	107273	101832	86800	58866	48355	47546	44411	39769	31883	29336

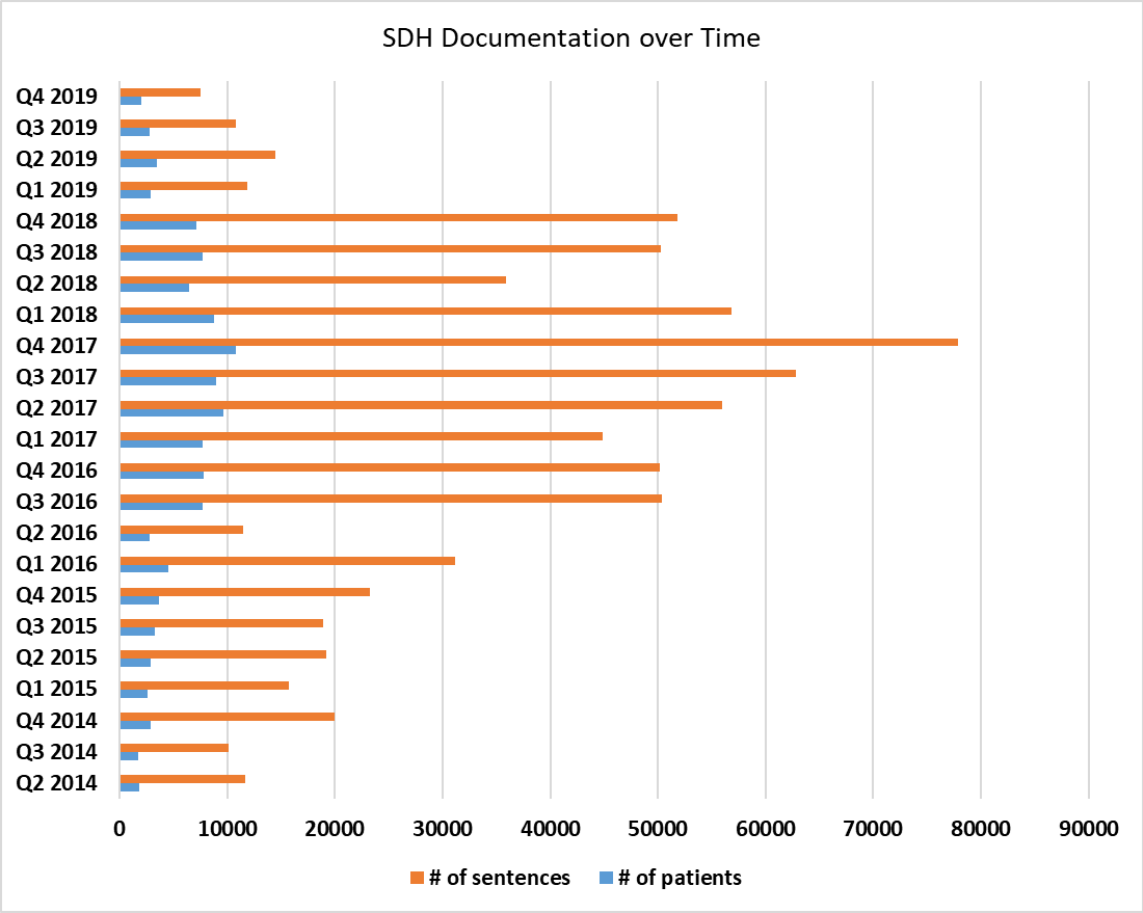
FIGURE 5b: Top 5 SDH term count per individual patient



	income	medicaid	charges	afford	lives alone	homeless	court date	financial issues	probation	jail
Number of patients	18033	15376	10776	10298	8025	6749	6535	6251	5788	5343
Total count	86800	107273	48355	39769	44411	101832	19430	12893	19941	31883

Figure 6 displays SDH documentation over time. The years 2017 and 2018 accounted for 57.9% of the total SDH documentation in the dataset. The number of patients with SDH documentation peaked in Quarter 4 of 2017 (N=10,795) with consistent SDH documentation between Quarter 3 of 2016 and Quarter 4 of 2018.

FIGURE 6: SDH documentation over time



**Dates range from 04/04/2014-12/09/2019*

Error Analysis

Among the 10,200 patients who did not have an identified SDH sentence, a randomized 4% were sampled and reviewed by a SME (Author RS) for missed SDH characteristics. No consistent characteristics of the SDH domain financial resource strain was found during this error analysis sampling. However, unique phrases that were associated with an individual patient did emerge such as, *"patient has no long term housing options"* and *"the patient is geographically isolated contributing to poor nutrition habits."* Misspelling of homeless and various unapproved abbreviations for "Medicaid" were identified

during this evaluation. During this negative sampling evaluation we found terms and phrases representing other SDH characteristic domains such as intimate partner violence, education status, and health literacy.

Discussion

In this study, we analyzed the rate of SDH documentation data among MHSUD patients who frequented the ED by using word embedding expansion of seed terms. This study demonstrates the feasibility of systematically identifying poorly documented SDH from millions of unstructured clinical notes. Identifying SDH in clinical notes may improve health systems' understanding of the risk factors that are significantly associated with social inequalities among patients with MHSUD. An assessment of availability and characteristics of SDH data in EHRs of health care systems, such as the one presented in this study, can be the first step for developing SDH data extraction tools and clinical decision support systems.

We trained two-word embedding models (unigram and bigram) using seed terms abstracted from published research studies. These models were designed to detect and identify semantically similar terms that characterize financial resource strain and poor social support, yielding 109 terms or phrases. This process can serve as a common foundation on which to build high-quality datasets for machine learning modeling and analysis of SDH representation in clinical note documentation. Few studies explored expanding vocabularies automatically through word embedding models[42,43], but rather rely on time and resource intensive gold-standard manual annotation as the first step in developing[19,23]. Our approach more effectively included subject matter experts (SME) to improve dataset quality for classification tasks at a sentence level to overcome multi-label classification and combat low precision and recall. The terms identified by our word embedding models and SMEs will be applied in a sentence classification task in

future work. Fan and colleagues achieved similar success using a word embedding (word2vec) expansion assisted by SME annotation to expand dietary supplement terminology from 14 to 35 terms.[26]

This is the first study, to our knowledge, to utilize a bigram word embedding model to expand terminology using EHR clinical notes. Although Zang and colleagues did not begin model development using existing terminology, their bigram models successfully identified redundant versus relevant new information in clinical notes and demonstrated that bigram models outperformed other n-gram models in prior studies.[44] In our study, the introduction of the bigram model significantly reduced the topic drift we observed in the unigram model. For example, the two models produced very different semantically similar words for the seed term *afford.* The unigram model returned the terms: *articulate, vocalize, copays, converse and produce* as opposed to the bigram model that returned: *afford pay, expensive, pay bills, can't afford, and afford copay.* This may be one reason the bigram model resulted in more terms and phrases that indicate a high likelihood of SDH characteristics. The bigram model also uncovered several regionally specific terms related to local shelters and resources such as "Oxford House" and "Healing Transitions". These two phrases represent the formal names of organizations that provide emergency shelter and services for low-income populations geographically close to this study's ED. The ability to identify regionally sensitive SDH terms is critical to high capture performance, especially if the SME annotator is not local or aware of the regional linguistics that often describes these services. Capturing regional specifics may have greater importance in SDH and other non-clinical identification tasks as disease processes largely have the same presentation despite location. Additionally, regionally specific terms may represent SDH characteristics among a population. For example, relying on public transportation in a metropolis city may not be a barrier to wellness, however, in rural or suburban areas where public transportation is scarce, this may be a significant barrier. This finding may limit the potential generalizability of SDH models if model development is dependent on local knowledge. Our use of locally knowledgeable SMEs minimized the impact of this issue on our methodology.

Studies that assessed the presence of SDH characteristics in EHR unstructured data report a wide range of findings[2,11,12,21,45]. Overall we found 38,971 (79.3%) patients with at least one documentation of financially related SDH, a finding similar to Hollister et al who extracted at least one type of socioeconomic status information from 8,282 (83.0%) individuals.[12] Our incidence of poor social support (23.2%) was higher than results published by Hatef et al[2] who found 16% of their patient EHR records mentioned social support[2]. However, Chen et al, found much higher rates of poor social support among their small sample size of 85 geriatric patients (62%)[21]. In contrast, we found a greater incidence of housing instability related documentation (23.2%), as opposed to 2% and 3% found by Navathe et al[45] and Hatef et al[2] respectively. When compared to Hollister et al[12], who aimed to determine homelessness status, we found much greater documentation of homelessness (13.7%). The term unemployed was found in 10.4% of our population, which was lower than Hollister et al (19.8%), however, our study expanded upon existing terms to capture employment and income insecurity finding documentation among 31.5% of our population. One reason for our data source having a higher incidence of SDH terminology may be the rich SDH data source consisting of MHSUD patients who frequent the ED. Additionally, our word embedding models uncovered several terms and phrases that were indicative of other SDH domains beyond financial resource strain and poor social support. The majority of these terms described exposure to violence including intimate partner violence such as *abusive relationship* and *sexual assaults*.

Major efforts are underway to increase the standardized vocabulary and content of EHR data across the United States[46–48], efforts that may eventually impact the quality and thoroughness of SDH documentation. Although several tools and methods have been developed to screen and address SDH in EHRs, these lack consistent application across EHRs. A recent study that evaluated the ability of six EHR vendor products that support health care professionals' identification of and response to patients' SDH found a number of challenges[10]. The results indicate barriers related to analyzing SDH data and sharing these data between health systems due to the absence of standardization of SDH screening instruments,

measurement, and codification[10,15]. We found similar evidence in our study, to include what appeared to be unstructured free-text fields, as well as auto-generated narratives from drop-downs and data entry fields. Because providers did not consistently enter data into these fields, these auto-generated prompts complicated our analysis. Without intervention, our model would have counted the structured data prompt as true SDH documentation. For example, a screen for financial resource strain appeared in a clinical note in the EHR: *"financial information: patient source of income: need for financial assistance?:"* The data structure of auto-generated prompts may create problems for various machine learning tasks because the data appear as lengthy paragraphs with no punctuation separating questions and thus resulting in excessive noise. Data warehouses storing this information should be aware of the downstream effects and would benefit from architectures that assist researchers in analyzing textual information.

The varied data structures (Figure 8) may be related to the massive organizational and technological growth of the UNC Healthcare System during the study years (2014-2019) (Figure 9). Over this time period over 10 hospitals across North Carolina, including WakeBrook (a large behavioral health facility), were acquired and integrated into UNC Healthcare Systems' EHR. Further, the variation in the number of patients with documented SDH over this time period may be largely due to organizational changes that included increased screening for SDH within primary care practices and care management teams. Finally, SDH documentation may be over-represented in this population due to prolonged boarding in EDs among MHSUD patients[49,50], thus leading to increased clinical note documentation during a continuous period of time. Other contributions to variation may be due to a series of natural disasters in the area resulting in an increase in displaced populations within the UNC Healthcare System.

Limitations

There are several limitations in this study. First, we only tested unigram and bigram SDH terms in this study. In the future, we would apply this method on multiword SDH terms for further investigation

and evaluation. Second, we did not experiment on hyper-parameters or different word embedding models on the task of SDH terminology development. Hyper-parameters such as window size that controls the number of contextual words surrounding a target word can range resulting in varying vector dimension sizes. This may have diminished the number of different semantic terms found for each SDH category. Third, we were not able to fully omit all auto-generated prompts or copy and paste documentation. We found frequent instances of copy and paste where additional words were added to the entry, even though the entry appeared to speak to the same issue. For example, *"cm* [case manager] *consulted by bedside rn to assist with homeless shelter resources as patient is reporting he has no where to go"* and *"cm paged d/t pt stating he is homeless and has no where to go."* The frequency of redundant text in clinical notes may create over representation of the construct of interest, thus challenging effective training and text mining model evaluation[39]. Future work should explore best practices for data management of free-text and the integration of auto-generated narratives. Fourth, it is unclear how the use of EHR clinical note templates may impact NLP methods. We encountered more than 100 different clinical note templates enumerated with various HTML strings such as *"10{{lack of transportation wildcard_metadata_end placeholder_metadata_begin 0{{trn mod high risk:304195102}."* These HTML strings may inflate term count, and require careful evaluation by SMEs. Fifth, we did not study SDH documentation or extraction in the general population, this may limit generalizability. Therefore, future studies must validate that SDH terms and phrases identified among a specific population (i.e., MHSUD) are applicable to the general population. Finally, because we examined clinical notes from a single health system emergency department, we may have captured linguistic patterns dependent on regional dialects and education thus limiting application to other geographic regions.

Conclusion

This was the first study, to our knowledge, that built upon existing biomedical research findings to develop an SDH seed term list and expand that list using a word embedding model. Our SME aided

NLP approach produced a high-quality training data set for future machine learning classification tasks and may represent a better use of costly, time-consuming gold-standard SME annotated datasets. The reported NLP findings in our study were based on the occurrences of specific terms and their semantic relatedness (e.g. terms, such as homelessness) within clinical notes. Documentation patterns related to SDH may help us develop an efficient NLP pipeline in future work; however, advanced study (e.g. manual annotation of SDH in a large body of text) is needed to evaluate the rate of false negative cases. In addition, deterministic information found in the structured fields (i.e., embedded questionnaires) can be used to create valuable training and validation datasets for ML experiments however the questionnaire prompts must more easily be identified and removed. Advanced NLP techniques may automatically identify highly associated patterns from the notes of specific cohorts and utilize those patterns to improve SDH extraction. We hope that our results will inform clinicians, researchers, and healthcare systems to understand the potential of EHRs to effectively capture SDH data. This study also provides support to informaticians to advance the standardization of EHRs data collection tools and terminologies for SDH and help inform public health policy.

REFERENCES

1. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington (DC): National Academies Press (US); 2015. doi:10.17226/18951

2. Hatef E, Rouhizadeh M, Tia I, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system. *JMIR Med Inform*. 2019;7(3):e13802. doi:10.2196/13802

3. Allen J, Balfour R, Bell R, Marmot M. Social determinants of mental health. *Int Rev Psychiatry*. 2014;26(4):392-407. doi:10.3109/09540261.2014.928270

4. Doran KM, Kunzler NM, Mijanovich T, et al. Homelessness and other social determinants of health among emergency department patients. *J Soc Distress Homeless*. 2016;25(2):71-77. doi:10.1080/10530789.2016.1237699

5. Ku BS, Fields JM, Santana A, Wasserman D, Borman L, Scott KC. The urban homeless: super-users of the emergency department. *Popul Health Manag*. 2014;17(6):366-371. doi:10.1089/pop.2013.0118

6. Krieg C, Hudon C, Chouinard M-C, Dufour I. Individual predictors of frequent emergency department use: a scoping review. *BMC Health Serv Res*. 2016;16(1):594. doi:10.1186/s12913-016-1852-1

7. Hudon C, Sanche S, Haggerty JL. Personal characteristics and experience of primary care predicting frequent use of emergency department: A prospective cohort study. *PLoS One*. 2016;11(6):e0157489. doi:10.1371/journal.pone.0157489

8. Birmingham LE, Cochran T, Frey JA, Stiffler KA, Wilber ST. Emergency department use and barriers to wellness: a survey of emergency department frequent users. *BMC Emerg Med*. 2017;17(1):16. doi:10.1186/s12873-017-0126-5

9. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. Washington (DC): National Academies Press (US); 2014. doi:10.17226/18709

10. Freij M, Dullabh P, Lewis S, Smith SR, Hovey L, Dhopeshwarkar R. Incorporating social determinants of health in electronic health records: qualitative study of current practices among top vendors. *JMIR Med Inform*. 2019;7(2):e13849. doi:10.2196/13849

11. Bejan CA, Angiolillo J, Conway D, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc*. 2018;25(1):61-71. doi:10.1093/jamia/ocx059

12. Hollister BM, Restrepo NA, Farber-Eger E, Crawford DC, Aldrich MC, Non A. Development and performance of text-mining algorithms to extract socioeconomic status from de-identified

electronic health records. *Pac Symp Biocomput*. 2017;22:230-241. doi:10.1142/9789813207813_0023

13. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. *J Acquir Immune Defic Syndr*. 2018;77(2):160-166. doi:10.1097/QAI.0000000000001580

14. Vest JR, Grannis SJ, Haut DP, Halverson PK, Menachemi N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int J Med Inform*. 2017;107:101-106. doi:10.1016/j.ijmedinf.2017.09.008

15. Gold R, Cottrell E, Bunce A, et al. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *J Am Board Fam Med*. 2017;30(4):428-447. doi:10.3122/jabfm.2017.04.170046

16. Carrell DS, Cronkite D, Palmer RE, et al. Using natural language processing to identify problem usage of prescription opioids. *Int J Med Inform*. 2015;84(12):1057-1064. doi:10.1016/j.ijmedinf.2015.09.002

17. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review. *J Biomed Inform*. 2018;77:34-49. doi:10.1016/j.jbi.2017.11.011

18. Sarmiento RF, Dernoncourt F. Improving patient cohort identification using natural language processing. In: *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing; 2016:405-417. doi:10.1007/978-3-319-43742-2_28

19. Feller DJ, Zucker J, Don't Walk OB, et al. Towards the Inference of Social and Behavioral Determinants of Sexual Health: Development of a Gold-Standard Corpus with Semi-Supervised Learning. *AMIA Annu Symp Proc*. 2018;2018:422-429.

20. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearb Med Inform*. 2015;10(1):183-193. doi:10.15265/IY-2015-009

21. Chen T, Dredze M, Weiner JP, Hernandez L, Kimura J, Kharrazi H. Extraction of geriatric syndromes from electronic health record clinical notes: assessment of statistical natural language processing methods. *JMIR Med Inform*. 2019;7(1):e13039. doi:10.2196/13039

22. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*. 2004;11(5):392-402. doi:10.1197/jamia.M1552

23. Juckett D. A method for determining the number of documents needed for a gold standard corpus. *J Biomed Inform*. 2012;45(3):460-470. doi:10.1016/j.jbi.2011.12.010

24. Wei Q, Franklin A, Cohen T, Xu H. Clinical text annotation - what factors are associated with the cost of time? *AMIA Annu Symp Proc*. 2018;2018:1552-1560.

25. Lingren T, Deleger L, Molnar K, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J Am Med Inform Assoc*. 2014;21(3):406-413. doi:10.1136/amiajnl-2013-001837

26. Fan Y, Pakhomov S, McEwan R, Zhao W, Lindemann E, Zhang R. Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA Open*. 2019;2(2):246-253. doi:10.1093/jamiaopen/ooz007

27. Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform*. 2018;87:12-20. doi:10.1016/j.jbi.2018.09.008

28. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. 2013.

29. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *J Biomed Inform*. 2017;76:102-109. doi:10.1016/j.jbi.2017.11.007

30. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014:1532-1543. doi:10.3115/v1/D14-1162

31. Ruch P, Baud R, Geissbühler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artif Intell Med*. 2003;29(1-2):169-184. doi:10.1016/S0933-3657(03)00052-6

32. Quimbaya AP, Múnera AS, Rivera RAG, et al. Named Entity Recognition Over Electronic Health Records Through a Combined Dictionary-based Approach. *Procedia Computer Science*. 2016;100:55-61. doi:10.1016/j.procs.2016.09.123

33. Bai T, Chanda AK, Egleston BL, Vucetic S. Joint Learning of Representations of Medical Concepts and Words from EHR Data. *Proceedings (IEEE Int Conf Bioinformatics Biomed)*. 2017;2017:764-769. doi:10.1109/BIBM.2017.8217752

34. Sulieman L, Gilmore D, French C, et al. Classifying patient portal messages using Convolutional Neural Networks. *J Biomed Inform*. 2017;74:59-70. doi:10.1016/j.jbi.2017.08.014

35. Muneeb TH, Sahu S, Anand A. Evaluating distributed word representations for capturing semantics of biomedical concepts. In: *Proceedings of BioNLP 15*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2015:158-163. doi:10.18653/v1/W15-3820

36. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544-551. doi:10.1136/amiajnl-2011-000464

37. Natural Language Toolkit — NLTK 3.5 documentation. https://www.nltk.org/. Accessed May 12, 2020.

38. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. *Qual Quant*. 2018;52(4):1893-1907. doi:10.1007/s11135-017-0574-8

39. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*. 2013;14:10. doi:10.1186/1471-2105-14-10

40. Winden TJ, Chen ES, Wang Y, Lindemann E, Melton GB. Residence, living situation, and living conditions information documentation in clinical practice. *AMIA Annu Symp Proc*. 2017;2017:1783-1792.

41. Zhang R, Pakhomov S, Melton GB. Automated identification of relevant new information in clinical narrative. In: *Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics - IHI' ' '12*. New York, New York, USA: ACM Press; 2012:837. doi:10.1145/2110363.2110467

42. Leroy G, Gu Y, Pettygrove S, Kurzius-Spencer M. Automated lexicon and feature construction using word embedding and clustering for classification of ASD diagnoses using EHR. In: Frasincar F, Ittoo A, Nguyen LM, Métais E, eds. *Natural Language Processing and Information Systems*. Vol 10260. Lecture notes in computer science. Cham: Springer International Publishing; 2017:34-37. doi:10.1007/978-3-319-59569-6_4

43. Zhang Y, Li H-J, Wang J, Cohen T, Roberts K, Xu H. Adapting Word Embeddings from Multiple Domains to Symptom Recognition from Psychiatric Notes. *AMIA Jt Summits Transl Sci Proc*. 2018;2017:281-289.

44. Zhang R, Pakhomov SV, Lee JT, Melton GB. Using language models to identify relevant new information in inpatient clinical notes. *AMIA Annu Symp Proc*. 2014;2014:1268-1276.

45. Navathe AS, Zhong F, Lei VJ, et al. Hospital Readmission and Social Risk Factors Identified from Physician Notes. *Health Serv Res*. 2018;53(2):1110-1136. doi:10.1111/1475-6773.12670

46. Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. *Health Aff (Millwood)*. 2018;37(4):585-590. doi:10.1377/hlthaff.2017.1252

47. Gottlieb L, Tobey R, Cantor J, Hessler D, Adler NE. Integrating social and medical data to improve population health: opportunities and barriers. *Health Aff (Millwood)*. 2016;35(11):2116-2123. doi:10.1377/hlthaff.2016.0723

48. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Med*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

49. Nolan JM, Fee C, Cooper BA, Rankin SH, Blegen MA. Psychiatric boarding incidence, duration, and associated factors in United States emergency departments. *J Emerg Nurs*. 2015;41(1):57-64. doi:10.1016/j.jen.2014.05.004

50. Pearlmutter MD, Dwyer KH, Burke LG, Rathlev N, Maranda L, Volturo G. Analysis of emergency department length of stay for mental health patients at ten massachusetts emergency departments. *Ann Emerg Med*. 2017;70(2):193-202.e16. doi:10.1016/j.annemergmed.2016.10.005

APPENDIX 1: Seed term evaluation

Highlighted terms mean they were included

Unigram/Phrase	Reference	Example from EMERSE	Note Type	Corpus Frequency	Number of Patients
homeless	UMLS	"Patient is currently homeless, after having moved back to North Carolina from Florida where he had been living with his sister until her son and family moved in with her which caused problems."	Psychiatry initial consult	101832	6749
	Bejan et al. (2017)	"Patient reports being homeless"	Case management progress note		
	Hollister et al. (2017)	"Chief ComplaintPatient presents with◊Homeless"	Emergency department encounter		
shelter	Bejan et al. (2017)	"The patient weighed the risks and benefits of an inpatient admission and voiced understanding of the risks involved in returning to homeless shelter, including the Interactive Recourse Center. ",	Psychiatry initial consult	47546	4680
		"He is encouraged to use outpatient resources, meetings, shelter, church and outpatient MH providers to continue to get better.",	Psychiatry follow-up		

		"Patient describes a detailed and circumstantial story basically starting with him being a homeless vet living at the Wilmington Street shelter in Raleigh, moving into transitional housing in Durham from November 2016 through March 2017, at which time his name came up for housing spot located in Clayton."	Initial psychiatric evaluation		
housing	UMLS	"Patient does already have a therapist (Natalie with Cary Behavioral Health), a psychiatrist (Dr. Celso Labao at Holly Hill) and safe housing in place."	Assessment/Plan	104791	7001
		"CM concerned continued path is on will result in numerous housing issues and possible readmissions."	NCHV Established patient visit		
		"Worked for NYC housing authority."	Unnamed note		
afford	UMLS	"Patient is requesting one of our doctors to re-write the RX so the medication can be covered under her Charity Care as she is unable to afford it."	Patient request note	39769	10298
		"Reports that he was doing well until he could not afford medication refills (including buprenorphine for relapse prevention)."	Psych emergency service initial consult		
		"Was prescribed Ranexa but has been unable to afford it."	Follow-up visit		
transitional	UMLS	"Pt reports 6/10 pain in low back with transitional movements secondary to compression fracture."	Inpatient occupational therapy	14229	4372

		"The metabolic activity within the prostate involves the transitional zone in left peripheral zone and is similar to that of mediastinal activity. "	Primary care physician		
		"The needle was exchanged for a 5-F transitional catheter which was used to place an 0.035 inch guidewire."	Procedure note		
uninsured	Hollister et al. (2017)	"Pt appears uninsured"	CM screening assessment	11161	3348
		"uninsured"	Unnamed note		
		"Pt is uninsured and patient's fiance has reported that the mechanism of burn was likely due to methamphetamine use."	Progress note		
medicaid	Hollister et al. (2017)	"CC explained that the bed would be paid for by Medicaid and there are no Medicaid beds available"	Unnamed note	107273	15376
		"Family is in the process of applying for Medicaid"	Unnamed note		
		"Secondary insurance MEDICAID NC"	Care management initial transition planning assessment		
streets	Bejan et al. (2017)	"Exposure/Witness to Violence: Yes; patient reports living on streets in DC and saw "a lot of dead bodies" and saw his "best friend shot dead."" ,	Psychiatry emergency service initial consult,	5743	1890
		"Type of Residence: Homeless: on the streets.",	Social work psychosocial assessment,		

		"Patient reports"I hope next time I see you it'll be on the streets.""	Unnamed note		
welfare	Bejan et al. (2017)	"pt is worried about his cat's welfare--no food or water in the apartment;"	Unnamed note	930	304
		"…due to the safety and welfare of the patient."	Unnamed note		
		"Child Welfare Referral"	Child welfare referral		
motel	Bejan et al. (2017)	"Patient lives with her granddaughter, granddaughters boyfriend, and great grandchild in a motel room"	Unnamed note	5583	1933
		"Patient recently left motel in Greensboro"	Clinical social worker progress note		
		"Pt was found on bench outside of Motel 6, unresponsive."	Emergency department provider note		
evicted	Bejan et al. (2017)	"Patient is now banned from the property and evicted after threatening Mr. X over the phone."	Unnamed note	3741	958
		"suspect underlying personality disorder: no need for acute psychiatric evaluation and no admission to suicidal ideation or intent, likely some form of secondary gain as this also coincides with her being evicted today",	Admission history and physical		
		"She was evicted from her apt in March 2016 due to non-payment."	Crisis and assessment services		
inadequate	Winden et al. (2017)	"breathe support: inadequate for speech"	Progress note	90428	17512
		"fluid intake inadequate"	Progress note		
cluttered	Winden et al. (2017)	"the home appeared cluttered in bedrooms"	Unnamed note	1360	96

		"His work has reported increase in abnormal behavior: not responding to questions, pacing around the office, cluttered work spaces and increased expressed anxiety about a work project.",	Emergency department provider note		
		" Pt has limited help and resources, her home is cluttered and she has limited room for mobility in home."	Unnamed note		
excessive	Winden et al. (2017)	Did not relate to SDH, only to burns and other medical conditions	N/A	132305	23881
lack	UMLS	"lack of support"	Clinical social worker progress note	4380407	30128
		"lack of active SI/HI"	Emergency department triage note		
		"lack of transportation"	Clinical social worker progress note		
lives alone	Hatef et al. (2019)	"Clinical Risk Factors: Multiplie Diagnoses (Chronic), Lives Alone or Absence of Caregiver to Assist with Discharge and Home Care, Functional Limitations",	Care management initial transition planning assessment,	44411	8025
		"Pt lives alone.",	Crisis and assessment psychologist assessment,		
		"She reports that he lives alone with his dog, states that he is in a wheelchair, lost his legs and fingers, TBI from Army deployment to Afghanistan 6 years ago. "	Family therapy progress note		

homelessness	Hollister et al. (2017)	"Discharge Needs AssessmentConcerns To Be Addressed-- compliance issue concerns;decision making concerns;homelessness;home safety concerns;basic needs concerns-" ,	Plan of care review	28462	3339
	Hatef et al. (2019)	"Unfortunate male with very poor social situation including homelessness living in a shelter currently with nonishcemic cardiomyopathy EF of 10% ,pulmonary hypertension, history of DVT, COPD, chronic tobacco, prior history of alcohol and marijuana use who presented with bright red blood per rectum, dyspnea on exertion, lower extremity weakness, orthopnea, and dyspnea.",	Discharge summary		
		"Discharge Needs AssessmentConcerns To Be Addressedmental health concerns;substance/tobacco abuse/use concerns;homelessness"	Plan of care review		
no home	Hatef et al. (2019)	"Post Acute Facility needed at discharge?: No Home Care/ Home Medical Equipment needed at discharge?: No"	Care Management Final Transition Planning Assessment	6767	3843
		"Post Acute Facility needed at discharge?: No Home Care/ Home Medical Equipment needed at discharge?: No"	Care Management Final Transition Planning Assessment		

		"Post Acute Facility needed at discharge?: No Home Care/ Home Medical Equipment needed at discharge?: No"	Care Management Final Transition Planning Assessment		
inadequate housing	UMLS	"Clinician met with client to discuss anxiousness and feelings of hopelessness attributed to limited income, unemployment and inadequate housing."	Transitional Care Clinic Behavioral Health Progress Note,	221	33
	Hatef et al. (2019)	"Client attributes these symptoms to his inadequate housing conditions (ie living in a boarding house with bed bugs), health concerns and lack of employment. Problem-solving(ie brainstorming) used to identify alternative locations for client to reside to gain relief from inadequate housing conditions"	Transitional Care Clinic Behavioral Health Progress Note,		
		"V60.1(Z59.1) Inadequate Housing"	Transitional Care Clinic Behavioral Health Progress Note,		
no family support	Hatef et al. (2019)	" In addition, he goes out of town every weekend for his business and she has no family support in the state to help her with the baby."	Emergency department progress note	456	252
		"He say he has no family support, lives with roommates Stated that he just wants to sleep."	Unnamed note		

		"Expected Family involvement (Who is able and willing to provide support?) : Other (Comment) (Pt reports no family support.)"	Social work psychosocial assessment		
no social support	Hatef et al. (2019)	"Patient lives alone with no social support and per physical therapy is not safe to be discharged home in this situation."	Daily progress note,	355	146
		"The patient is not safe to be discharged home as she is nonambulatory, has no social support and no available care at home."	Recreational therapy evaluation,		
		"Auditory or visual hallucinations, paranoia, delusions : no Social support : yes"	ED Clinical impression		
lack of assistance	Hatef et al. (2019)	"Pt's father expressed concerns regarding pt's return home and lack of assistance at home."	Social work psychosocial assessment	90	43
		"Patient reported to admitting physician that her husband does not give her medication "too many times"and she cannot get out of bed to chair due to lack of assistance."	Unnamed note		
		"Pt is unsafe to return home with current functional abilities and lack of assistance at home, and will require d/c to rehab at this time."	Physical therapy note		
no social supports	Hatef et al. (2019)	"Reports having no social supports."	Psychosocial addendum	1060	349
		"Current Level of Social Involvement and Support: Reports/displays no social involvement (Pt reports no social supports)",	Recreational therapy evaluation		

		"Current Level of Social Involvement and Support: Reports/displays no social involvement (pt reports no social supports)"	Recreational therapy evaluation		
unemployment	Bejan et al. (2017)	"Stressors: son with special needs, multiple medical problems, unemployment, divorce"	Psychiatry daily progress note,	9575	1723
	Hollister et al. (2017)	"Patient's current stressors include: unemployment, homelessness and recent relapse on illicit substances."	Crisis and assessment services initial assessment		
		Consequences of Use: several legal charges, unemployment	Psychiatry daily progress note,		
retired	Hollister et al. (2017)	"Pt is a retired DEA agent, was very talkative and active at Brookdale per friend."	Occupational therapy note,	88339	6783
		"Occupational History◊Retired"	Assessment and plan note,		
		"Retired"	GI hospitalist consult note		
prison	Bejan et al. (2017)	"Worked in a prison as a counsellor, correction officer."	Assessment	29336	4359
		"He was in prison couple of years ago."	Emergency department care encounter		
		"Her uncle was a drug dealer, her father was in and out of prison and had drug problems, her brother has problems with drugs."	Psychiatry initial consult		

jail	Bejan et al. (2017)	"Writer wonders if mother and patient may be "Doctor shopping" for personal gain such as with legal system (patient was in jail 1/14/2018)."	Crisis and assessment services	31883	5343
		"LEGAL HISTORY: Ms.XX denies all past and current history of legal issues, including jail/probation time, litigation, bankruptcy, gambling, or any other legal problems."	Social work assessment		
		"LEGAL HISTORY: Mr. XX denies current or historical legal issues, including jail/probation time, litigation, bankruptcy, gambling, or any other legal problems."	Social work assessment		
prositute	Bejan et al. (2017)	"He states that he had gone out to party on "the bad side of town," consumed both EtOH and Crack cocaine, solicited a prostitute, and was subsequently robbed."	Psychiatry daily progress note,	659	238
		"Pt left him and she went to her dad's until he wanted his prostitute wife back than Was homeless and than came her. "	Adult psychosocial assessment,		
		"She is repeating that her boyfriend raped her and wants her to prostitute."	Treatment team review		
prostitution	Bejan et al. (2017)	"States that daughters had given her advice of using prostitution to make money, which patient uses as an example of oddness of their relationship."	Psychiatry progress note	1026	193

		"History of violence: To recent sexual assaults /reporting in high-risk situations due to prostitution"	Psychiatry history and physical note		
		"Pt reports in 2015 she was involved in a prostitution ring in Atlanta after her aunt took her to stay with a friend."	Social work psychosocial assessment		
police	Bejan et al. (2017)	"Discharge from ED may need security and/or police assistance"	Unnamed note	108784	7331
		"Called hospital police"	Unnamed note		
		"Hospital police picked her up around 1045 and she was calm and cooperative with transport."	Unnamed note		
trespassing	Bejan et al. (2017)	"Counselor learned that client refused to leave upon discharge and was arrested for trespassing. "	Unnamed note	3369	540
		"Counselor was informed that client wants to be cared for and if a hospital will not admit him, he would prefer going to jail for trespassing or other charges to keep "a roof over his head"."	Unnamed note		
		"Pt is under arrest for trespassing and disruptive behavior per HPPD officer."	Unnamed note		
prisoner	Bejan et al. (2017)	"XX is a 28 y.o. female prisoner seen in clinic today for a palpable breast mass. "	Attending note	1251	753
		"Patient put shoes on and walked out the room stating " I came in voluntarily. I can leave. I'm not a prisoner"	Unnamed note		

		"Pt initially calm and cooperative after sitter d/c'd, pt then became restless, trying to get OOB and yelling that we are 'holding him prisoner'."	Patient care overview note		
incarcerations	Bejan et al. (2017)	"Chronically ill-appearing 53-year-old male with no prior psychiatric diagnoses, multiple incarcerations and polysubstance abuse presents with homicidal ideation towards 3 specific people that state he states robbed him last night."	ED provider note	1246	832
		"X had no meaningful contact with patient due to recurrent incarcerations and/or hospitalizations.",	Interdisciplinary huddle note		
		"She has detoxed periodically during her multiple incarcerations (x5, all for drug possession), and detoxed successfully during her last two pregnancies using methadone."	ED provider note		
veteran	Hollister et al. (2017)	"Pt is retired RN and is not a veteran."	Social worker psychosocial assessment	5649	1845
		"Camilla stated that they didn't write anything on the FL-2 about the patient being a veteran with PTSD/and or other behaviors, that they just knew patient had dementia."	Social worker note		
		"Legal: (Pt is a veteran and declined VA transfer.)"	Social work psychosocial assessment		

APPENDIX 2: Unigram model word embedding expanded terms

	word (seed term)	w1	w2	w3	w4	w5
Housing insecurity	**homeless (N=121740)**	shelter (N=55952)	mens (N=258697)	resides (N=26238)	disabled (N=20276)	shelters (N= 4887)
		0.5438	0.4881	0.4394	0.403	0.4004
	shelter (N=55952)	shelters (N= 4887)	homeless (N=121740)	hotel (N= 12139)	streets (N=6353)	motel (N= 6272)
		0.6737	0.5438	0.5252	0.5218	0.5073
	housing (N= 104791)	explore (N= 44194)	rental (N=7281)	lodging (N=1653)	apartment (N= 63602)	shelters (N= 4887)
		0.4595	0.4557	0.4487	0.4479	0.4474
	transitional (N= 14229)	exploring (N= 7927)	homehealth (N= 451)	continuity (N= 20199)	coordinating (N=23917)	charity (N= 30504)
		0.5784	0.5003	0.4949	0.4946	0.4336
	streets (N=6353)	residing (N= 6568)	motel (N= 6272)	sober (N= 63868)	roommates (N= 7553)	hotel (N= 12139)
		0.5357	0.4488	0.4459	0.4166	0.3978

	motel (N= 6272)	trailer (N= 10369)	basement (N= 3259)	roommates (N= 7553)	fiance (N= 31600)	apartment (N= 63602)
	0.6052	0.5762	0.5701	0.5272	0.5133
	evicted (N= 4087)	homeless (N=121740)	stole (N= 36389)	foreclosed (N= 205)	landlord (N= 5477)	rent (N= 6658316)
	0.5406	0.5204	0.5174	0.5135	0.5053
	inadequate (N= 90428)	imbalanced (N= 15753)	sporadic (N= 7011)	inadequatemobility (N = 101)	inadequateinfectionmobility (N= 101)	impaired (N= 778564)
	0.4735	0.4023	0.3878	0.3745	0.3579
	cluttered (N= 1360)	smells (N= 9810)	rug (N=1467149)	messy (N= 970)	filthy (N= 458)	garbage (N= 2209)
	0.5168	0.5124	0.4723	0.4717	0.4389
	excessive (N= 132305)	burned (N= 14047)	blistering (N= 15588)	tanning (N= 9934)	sun (N= 237356)	sunscreen (N= 23731)
	0.5081	0.457	0.4219	0.3958	0.3721
General	**afford (N= 45435)**	articulate (N= 10555)	vocalize (N= 2287)	copays (N=1323)	converse (N= 6973)	produce (N= 26154)
	0.5994	0.5766	0.5089	0.4957	0.4919

	affordable (N = 4554)	subsidized (N= 1011)	roommates (N= 7553)	rental (N=7281)	purchased (N= 4715)	cheaper (N=1987)
		0.508	0.5059	0.5017	0.4139	0.3901
	uninsured (N= 13108)	enrolled (N=18428)	accepts (N= 10345)	federal (N= 8099)	discount (N= 5116)	eligible (N= 16319)
		0.5443	0.5376	0.5362	0.5354	0.5199
	medicaid (N= 153937)	bcbs (N= 22883)	mcd (N= 43785)	mcaid (N=181)	uhc (N=3656)	tricare (N= 3516)
		0.6706	0.662	0.6273	0.5977	0.5963
	welfare (N= 930)	mens (N=258697)	cedar (N= 5418)	child (N= 18288)	horizon (N= 12932)	check (N = 104567)
		0.6187	0.5068	0.503	0.4722	0.4609
Employment insecurity	**unemployment (N= 10592)**	finances (N= 37855)	unemployed (N= 58866)	stressors (N= 387511)	discord (N= 17171)	homeless (N=121740)
		0.557	0.5264	0.5211	0.5208	0.5064
	retired (N= 88339)	cpm (N= 46130)	row (N= 2183297)	sgrp (N= 36755)	inv (N= 1134426)	ghr (N= 5196)
		0.9032	0.7524	0.453	0.4044	0.3654

prison (N= 34521)	forging (N= 239)	shoplifting (N= 2470)	charges (N= 56826)	probation (N= 22999)	jail (N= 37092)
	0.6721	0.6403	0.5573	0.5402	0.5217
jail (N= 37092)	prison (N= 34521)	trespassing (N= 3829)	arrested (N= 9670)	drunk (N= 11647)	violation (N= 5474)
	0.6159	0.492	0.4835	0.4784	0.4643
prostitute (N= 748)	reserves (N= 608)	exgirlfriend (N= 78)	boyfriends (N= 1959)	bern (N= 35838)	clts (N= 220)
	0.4899	0.4741	0.4456	0.4021	0.4002
prostitution (N= 1130)	marrying (N= 336)	sold (N= 4000)	abusive (N= 28137)	johnny (N= 10653)	girlfriend (N= 72068)
	0.8261	0.8254	0.543	0.5312	0.4961
police (N= 108784)	officer (N= 42624)	banned (N= 854)	trespassing (N= 3829)	desk (N= 51554)	escort (N= 33573)
	0.5677	0.5519	0.5106	0.5106	0.4836
trespassing (N= 3829)	misdemeanor (N= 4155)	banned (N= 854)	police (N= 108784)	court (N= 95504)	trespass (N= 4666)
	0.5146	0.5125	0.5106	0.4955	0.4086

prisoner (N= 1251)	isolates (N= 3549)	subsidized (N= 1011)	upstairs (N= 9268)	apartment (N= 63602)	garden (N= 19126)
	0.493	0.4807	0.4699	0.4683	0.4625
incarcerations (N= 1246)	centimeters (N= 4378)	maculopapular (N= 2316)	deroofed (N= 106)	egds (N= 1116)	basing (N= 146)
	0.52	0.5032	0.4979	0.4893	0.4635
veteran (N= 6671)	military (N= 79873)	criminal (N= 4065)	incomeemploymentdisability (N= 1203)	justice (N= 7837)	income (N= 106064)
	0.4743	0.4592	0.4452	0.445	0.3971

APPENDIX 3: Word embedding expansion manual observations evaluation

Highlighted terms mean they were included

Unigram/Phrase	Reference	Example from EMERSE	Note Type	Corpus Frequency	Number of Patients
disabled	expansion	"He feels that he cannot work a regular job due to his level of physical incapacitation and needs father to understand that he is disabled and needs time and patience to get better."	Psychiatry follow-up consult	13445	4744
		"Pt is disabled and does not drive."	Social work psychosocial assessment		
		"She has been extremely treatment resistant and severely disabled."	Psychiatry daily progress note		

hotel	expansion	"Expected Discharge Date: (Once bed available at CRH vs d/c to hotel until permanent housing found)",	Care management	10655	2465
		"She is currently in a hotel.",	Psychiatry initial consult		
		"Pt states he walked here from a hotel in Durham because he "didn't feel right", but says he can't explain what he means by that, and then says he was actually brought by ambulance."	Psychiatry initial consult		
copay	expansion	"Patient's copays should not be expensive due to being insured by medicaid.",	Unnamed note	19291	5219
		"The patient has social or financial issues that need to be addressed at this time: Yes (details: pt currently uninsured, does not have a PCP as he cannot afford any sliding scale copays period. " ,	Case management initial assessment		
		"Patient had PET scan done in 2016 and is concerned regarding the cost of the copay. Patient has copays of $45 to several specialty physicians."	Unnamed note		
subsidized	expansion	"CC provided 2 new Rx through Guild for 23.95 and explained that pt did not not qualify for senior housing and would need to go through Housing Authority for subsidized housing.",	Case management discharge closing note	947	323
		"Waiting Lists for Subsidized Housing: n/a",	New patient psychiatric evaluation note		

		"Mr. X lives in HUD subsidized housing, at Ashley Forrest where he shares a common space with other residents. "	Unnamed note		
cheaper	expansion	"I can complete a prior auth for Eliquis but the pharmacist stated even with auth, Xarelto will be cheaper.",	Unnamed note	1772	1182
		"Patient would like to know if you could call her in something cheaper."	Unnamed note		
		"She is self pay and she is looking for cheaper alternatives to her medications. "	Unnamed note		
foreclosed	expansion	"Son, reported that parent are broke now, "loss everything" and had house foreclosed in FL before moving here a few months ago.",	Unnamed note	167	86
		"Per X: Pt's home is being foreclosed on him but pt is unaware of this.",	Unnamed note		
		"Pt noted he is in the process of being evicted from his home because it has been foreclosed on due to his wife not paying the mortgage bills."	Social work Psychosocial assessment		
filthy	expansion	"Pt reports she lives in a filthy house."	Unnamed note,	458	127
		"The bottom of her feet are black and filthy so she is clearly been full weightbearing in the room."	Orthopedic progress note,		
		"The pt stated that he wanted to shower "because I feel filthy being here""	Unnamed note		

garbage	expansion	" She reported that someone had put garbage all over her room, stones in her food, and a toenail in her glasses case. ",	Unnamed note	2209	834
		"X pulled a picture off of our waiting room wall, crumpled it and threw it and the frame in the garbage",	Psychiatry note		
		" Patient did not complete task and tray and garbage remain in room."	Unnamed note		
landlord	expansion	"increasing stressors related to daughter now moved back with pt who " already have a lot" medical issues related to eye problems, also financial stressors , has to move out sec to issues with landlord",	Psychiatric note	4766	1081
		"Pt. stated that she rescued two baby goats when their landlord, who lives on the adjoining property, went on a trip and left them unsheltered and unfed to die in the winter.",	Abuse consult		
		"States he was in process of moving when had accident and has been in hospital and unable to get rent and deposit money to new landlord"	Unnamed note		
stole	expansion	"Patient also reported another patient stole her gold rings."	Unnamed note	12625	3214
		"She also becomes a "kleptomaniac" and this last time stole something from her aunt and was caught."	Psychiatry initial consult		

		"Today law enforement was dispatched to residence for a larceny report, he had stole a tv and pawed it for crack cocaine."	Unnamed note		
economic	expansion	"Stressors: Relational Problems (Recent separation from husband), Abuse and Neglect (Husband is verbally and physically abusive; they are currently separated), Economic Problems (Patient's separation from husband is causing financial stress)",	Psychiatry psychotherapy note	12382	3610
		"Risk factors for harm to others: high emotional distress, violence towards others in the last 6 months, history of violence towards others, exposure to violence, diminished economic activities, childhood abuse, past substance abuse and lack of insight ",	Psychiatry follow-up consult note		

		"08/21/17 0844SOCIO-ECONOMIC/LIFESTYLE DATASupport System to Help with DiabetesFamilyCultural/ Religious Influences on Medical CareNoCultural/Religious Influences on DietNoFinancial Issues Buying FoodNoExercise ActivitiesYesExercise FrequencyOccasionallyExercise Duration10 minutes of walking; recently startedDaily Stress LevelModerateStress Management TechniquesReadsSleeping Habits< 8 hoursReadiness to ChangeMaybeMotivation to Change Scale7"	Unnamed note		
produce	expansion	"Treatment will often involve judicious use of pain medications and interventional procedures to decrease the pain, allowing the patient to participate in the physical activity that will ultimately produce long-lasting pain reductions. ",	Anesthesiology note	26154	7458
		"He reported improvement in his depressive symptoms course told stay cannabis free and was discharged on Wellbutrin, BuSpar and cerefolin which contains an amino acid known to produce cannabis cravings",	Unnamed note		
		"The photic stimulation does produce a symmetric driving response. "	Procedure note		

forging	expansion	"Legal History shoplifting 2004, spent 22 months in prison for forging xanax prescription (2009-2011)",	Psychiatry progress note	229	30
		"After having a slip up at that time and being arrested for forging prescriptions, he got back on Suboxone.",	Psychiatry history and physical note		
		"Mr. X reported that he had been in prison for two years due to forging prescriptions, and since getting out had been on and off substances."	Social work psychosocial assessment		
shoplifting	expansion	"Type of Offense-shoplifting"	Initial Psychiatry Assessment Evaluation	2316	334
		"He had an episode of random shoplifting that he does not recall and ultimately has been sent to a neuro-cognitive specialist for workup.",	Psychiatry initial consult		
		"Pt says today that this dx was given about 30-40 years ago and was due to 'impulsive behavior: hitchhiking, shoplifting, promiscuous-- but I don't know how they could tell when I was drinking'. "	Psychiatry initial consult		
charges	expansion	"At this time patient denies feeling unsafe at home, feeling threatened at home, wanting to press charges, that he would feel uncomfortable going back home." ,	Emergency department encounter	48355	10776
		"They will not be interpreted and no charges will apply.",	Unnamed note		

		"Today's Charges (noted here with $$):"	Outpatient physical therapy		
probation	expansion	"He is currently not on probation/parole. "	Psychiatric admission assessment note	19941	5788
		"Current/Prior Legal: Yes; 1 count of B&E and currently on probation. "	Crisis and assessment services initial assessment		
		"Records were faxed by Referral Coordinator to Pain Clinic in Wilson (pt stated he was referred there by probation officer)"	Unnamed notes		
banned	expansion	"Current living arrangements, marital status, children:Pt reports he has been homeless since 2003 when out of prison; reports he stays around the Raleigh area reporting Healing Transitions, Wilmington St shelter to which he is now banned but cannot explain when or why, and Wake co jail or a tent in the woods.",	Psychosocial assessment	796	282
		" I URGED her to seek therapy--pt reports verbal abuse from her husband who has a long hx of anger control problems (has been banned from several doctors offices due to verbal outbursts).",	Assessment & Plan		
		"He reports that recently he was banned for 30 days after getting into some sort of conflict with a peer which he is vague about."	Psychiatry discharge summary		

violation	expansion	"I warned her about the violation of her contract by obtaining a percocet prescription from Dr. Shields.",	Discharge summary	5474	1081
		"He had an extensive violation."	Assessment/ Plan		
		"Due to her history of coronary artery disease patient was noted to Hospital for further violation treatment. "	Discharge summary		
income	expansion	" He has no income currently and states he's living off of his mother's life insurance"	Unnamed note	86800	18033
		"Patient source of income: employed and medicaid"	Care management initial transition planning assessment		
		"Source of Income: Student/Parents"	Social work psychosocial assessment		
charity	expansion	"Signed: Charity XX, SLP"	Speech language pathology clinical swallow assessment	30504	12304
		"Subsequent discussion with Hartland hospice nurse Charity at 919-268-9664 she advised the patient was sent to the emergency department by staff due to the above symptoms and they were not notified.",	History of Present Illness		
		"Type of financial assistance required: Medicaid application, Charity care, Medication assistance"	Care management initial transition planning assessment		

finances	expansion	"1) Does patient need assistance with ADLs, managing finances, remembering to take medications, using the phone, shopping for food, making meals, and performing housework? Yes",	Assessment/ Plan	37855	13453
		"1) Does patient need assistance with ADLs, managing finances, remembering to take medications, using the phone, shopping for food, making meals, and performing housework? No",	Assessment/ Plan		
		"1) Does patient need assistance with ADLs, managing finances, remembering to take medications, using the phone, shopping for food, making meals, and performing housework? Yes, has caregiver 3x/week."	Assessment/ Plan		
rent	expansion	"Pt was living in Lumberton in Section 8 Housing and per mother he was evicted because he was not paying rent",	Psychiatry discharge summary	6658316	19521
		"Client is concerned about the increased rent but stated that they are splitting the cost.",	Established patient psychotherapy		
		"LCSWA X discussed financial opportunities for pt, including possible assistance with rent/ mortgage, utility bills, food pantries, and discretionary funds. "	Advocate intensive case management brief note		

APPENDIX 4: Bigram word embedding expanded terms

	word (seed term)	w1	w2	w3	w4	w5
Housing insecurity	homeless	currently_homeless	resides	unemployed	homeless_shelter	lives_alone
		0.6373	0.5598	0.5485	0.5388	0.5281
	shelter	oxford_house	homeless_shelter	ifc_shelter	hotel	durham_rescue
		0.7303	0.6776	0.6662	0.6371	0.608
	housing	employment	finances	unemployment	income	subsidized_housing
		0.5514	0.5459	0.5334	0.5157	0.5129
	transitional	unc_charity	committee	continuity	transfer_critical	helping_hands
		0.5521	0.5502	0.54799	0.532	0.5072
	streets	street	buying	selling	drunk	stealing
		0.5339	0.5284	0.5127	0.5086	0.4898
	motel	homeless_shelter	shelter	hotel	apartment_complex	subsidized_housing
		0.5774	0.5608	0.5287	0.5267	0.5216
	evicted	evict	landlord	infested	firearm	police_called
		0.5968	0.565	0.5498	0.5463	0.5373
	inadequate	intervention_evaluatemaintain	poor	imbalance_indequate	lack	insufficient
		0.4806	0.4586	0.4527	0.4461	0.4369
	cluttered	filthy	exposed_rug	edges_furniture	messy	tidy
		0.5745	0.5466	0.5262	0.5205	0.4963
	excessive	heavy	excess	heavier	increased	substantial
		0.5077	0.4535	0.4415	0.4341	0.4149
	banned	robbed	pay_rent	staying_ifc	ovedosed_heroin	lost_job
		0.4263	0.425	0.4227	0.4183	0.4077
	homelessness	unemployment	financial_strain	medication_nonadherence	legal_charges	unstable_housing
		0.6636	0.6516	0.5899	0.5882	0.5789
	hoarder	narcissistic_traits	aspergers	conditions_squalid	isnt_worth	smell_smoke
		0.6502	0.5282	0.4992	0.4862	0.4841
	foreclosed	condemned	pay_mortgage	payments	flooded	staying_secu
		0.5359	0.4959	0.4775	0.5965	0.5909

foreclosure	foreclosed	secu_hospice	staying_secu	rarely_leaves	especially_stairways
	0.5071	0.5069	0.5008	0.4951	0.4922
landlord	rent	suing	evicted	apartment	rommate
	0.5894	0.5726	0.565	0.55	0.5422
eviction	landlord	rent	dispute	evicted	recent_breakup
	0.4428	0.4326	0.4249	0.4171	0.4056
currently_homeless	lives_parents	homeless	shelter	unemployed	employed
	0.6583	0.6373	0.5793	0.5716	0.5488
homeless_shelter	shelter	oxford_house	hotel	motel	boarding_house
	0.6776	0.6544	0.5828	0.5774	0.5764
oxford_house	shelter	healing_transition	homeless_shelter	residential_treament	halfway_house
	0.7303	0.7197	0.6544	0.6198	0.6139
rescue_mission	salvation_army	ministries	public_housing	durham_rescue	oxford_house
	0.666	0.5977	0.592	0.5828	0.5767
transitional_housing	accepts_clients	oxford_house	subsidized_housing	public_housing	halfway_house
	0.49612	0.4829	0.4829	0.4783	0.4752
unstable_housing	homelessness_lack	separation_husbandchildren	family_discord	stressors_homelessness	financial_strain
	0.6141	0.5919	0.5781	0.5519	0.5472
ifc_shelter	shelter	oxford_house	stratford	homeless_shelter	hotel
	0.6203	0.6506	0.5986	0.5913	0.5907
subsidized_housing	public_housing	apartments	housing_authority	motel	ssdi
	0.5938	0.5838	0.5736	0.5364	0.5357
public_housing	apartments	housing_authority	womens_shelter	condo	0.5357
	0.6823	0.6228	0.6065	0.6043	0.5938
halfway_house	oxford_house	residential	freedom_house	residential_program	homeless_shelter
	0.5965	0.5504	0.504	0.4965	0.4903
stressors_homelessness	stressors	stressors_homeless	unstable_housing	homelessness	family_discord
	0.6148	0.5623	0.5519	0.5227	0.5216
stressors_homeless	stressors_unemployment	family_discord	stressors_homelessness	stressors	martial_discord

		0.581	0.5755	0.5623	0.5526	0.5487
General	afford	afford_pay	expensive	pay_bills	cant_afford	afford_copay
		0.5481	0.5117	0.4994	0.4931	0.4928
	affordable	able_afford	cost	discounted	afford	unable_afford
		0.5171	0.5108	0.507	0.4863	0.4803
	welfare	mold_bathroom	stat_plasma	prolapse_pessary	try_inform	criminal_bg
		0.62702	0.6107	0.573	0.5651	0.5281
	income	food_stamps	housing	sliding_fee	insurance	receives_disability
		0.5808	0.5157	0.5105	0.5088	0.507
	financially	housing	rent	income	currently_homeless	receives_ssi
		0.5457	0.532	0.5163	0.4988	0.4969
	subsidized	mlk_blvd	income	targeted_housing	public_housing	housing_authority
		0.6813	0.6633	0.6553	0.6403	0.6411
Employment/income insecurity	unemployment	homelessness	financial_strain	stressors_unstable	stressors	legal_charges
		0.6636	0.6321	0.5845	0.583	0.574
	unemployed	currently_unemployed	employed	disabled	retired	employment
		0.7087	0.6821	0.6446	0.627	0.616
	retired	currently_unemployed	currently_employed	unemployed	married	employed
		0.6804	0.6287	0.627	0.6257	0.6178
	prison	jail	charges	incarcerated	restraining_order	arrested
		0.7046	0.5444	0.5419	0.5405	0.5383
	jail	prison	arrested	restraining_order	probation	charges
		0.7046	0.692	0.6721	0.6016	0.5901
	prostitution	possession_charge	theft	murdered	kidnapped	protective_services
		0.4498	0.4379	0.4276	0.4265	0.4194
	prostitute	drug_dealer	cnc_wordlist	another_man	conspiracy	trajan
		0.4544	0.4268	0.4262	0.4217	0.4193
	police	law_enforcement	sheriff	police_officer	rpd	leo
		0.7091	0.7063	0.6949	0.6652	0.6614
	incarcerations	hes_jail	abusive_relationship	sexual_assaults	addicted_opiods	living_locations

		0.5896	0.486	0.4769	0.4741	0.471
	incarcerated	lived	separated_husband	meetings_clubs	jail	jailed
		0.5293	0.4866	0.4845	0.4777	0.4657
	tresspassing	obtaining_property	misdemeanor_larceny	false_pretenses	panhandling	breaking_entering
		0.6832	0.6741	0.6579	0.6166	0.6159
	innmate	rubbing_feces	joco	sherriff_deputy	brought_cas	orange_county
		0.582	0.4929	0.4907	0.4832	0.4738
	prisoner	loaded_gun	access_gun	shotgun	robbed	trashed
		0.5571	0.5565	0.5308	0.53	0.5246
	veteran	military_service	receives_disability	pension	food_stamps	currently_employed
		0.5534	0.5228	0.5196	0.5183	0.5175
	probation	currentprior_legal	charges	felony	dwi	parole
		0.6628	0.6609	0.6557	0.6489	0.6295
	parole	probation	jail	arrested	restraining_order	charges
		0.6295	0.5467	0.5167	0.5062	0.506
	disability	unemployed	medicaid	food_stamps	high_school	disability_assessment
		0.5668	0.4939	0.4924	0.4745	0.4716
	disabled	unemployed	divorced	retired	lost_job	married
		0.6059	0.6009	0.5708	0.5697	0.5565
	jobless	unemployed	exfiance	substance_useabusesupport	relaspe	determine_eligibility
		0.5614	0.5232	0.5169	0.5149	0.5102
	disability_income	receives_disability	ssi	ssdi	pension	food_stamps
		0.4859	0.4784	0.4769	0.4499	0.4277
	lost_job	quit_job	kicked	relapsed	losing_job	abusive_relationship
		0.6979	0.6316	0.6208	0.6009	0.5954
	receives_disability	receives_ssi	receives_ssdi	ssdi	employed	pension
		0.6491	0.63	0.5835	0.5601	0.5463
Food insecurity	food_pantries	food_stamps	homeless_shelters	meals_wheels	food_pantry	ifc
		0.527	0.5252	0.5187	0.5185	0.5129

	food_stamps	ssdi	pension	charity_care	medicaid	funds
		0.6568	0.6048	0.5482	0.5399	0.5179
	food_insecurity	financial_resource	medication_affordability	resource_strain	social_connections	community_resource
Insurance	uninsured	insurance	charity_care	selfpay	funding	income
		0.52178	0.5168	0.4814	0.4585	0.4543
	medicaid	insurance	bcbs	charity_care	uhc	mcd
		0.6937	0.6758	0.608	0.6046	0.5843
	copay	copays	pay	payments	cost	fees_partials
		0.6237	0.5878	0.5806	0.5635	0.5597
	cheaper	expensive	insurance	payment	cost	insurance_company
		0.5294	0.5286	0.5194	0.4957	0.4848
	selfpay	mcaid	mcd	medicaid	ltc_medicaid	receives_ssi
		0.5222	0.4983	0.4811	0.4779	0.4696
	charity_care	applications	application	nc_medassist	medicaid	pap_applications
		0.6761	0.644	0.613	0.608	0.5855

APPENDIX 5: Terms found on observation after unigram word embedding expansion

Highlighted terms mean they were included

Unigram/Phrase	Reference	Example from EMERSE	Note Type	Corpus Frequency	Number of Patients
financially	observation	"Post-op problems: no symptoms however patient struggles financially to get enough vitamins to last her from one office visit to the next.",	Bariatric post-op established visit	5184	2644
		"Also, as it seems like she is financially dependent on these odd jobs that she works. ",	Pulmonary fellows clinic f/u note		
		"Patient reports that he is not able to financially stabilize himself."	Emergency department provider note		

disability income	observation	"Employment history, current income, financial resources: Pt receives disability income for bipolar d/o. ",	Social work psychosocial assessment	1682	849
		"She reports that a case worker at DSS told her that her disability income is too high to qualify.",	Unnamed note		
		"Source of Income: Disability income of $735.00 a month. "	Social work psychosocial assessment		
lack of transportation	observation	"Recently missed a PT appt at UNC PMR on Fordham Blvd due to lack of transportation.",	Assessment/Plan	3264	1605
		"Ability to Access Community Services: Lack of transportation (Pt voiced that she is currently receiving services via Project Access.)",	Social work psychosocial assessment		
		"Ability to Access Community Services: Lack of transportation, Unfamiliar with options/procedures for obtaining"	Social work psychosocial assessment		
parole	observation	"LEGAL HISTORY Arrests: denies Jail or Prison: denies Probation or Parole: denies",	New patient psychiatric evaluation note	3603	1382
		"Ms X states she gave the info to his parole officer who will go by his house tonight.",	Progress note		
		"Reports longest period of abstinence between May, 2010 and Nov 2011 (though prior notes say Aug 2012), when he was on parole and living at the Oxford House."	Psychiatry emergency service initial consult		

lack of satisfaction with housing	observation	"Discussed pattern of poor decision making, attempts at locating safe housing, patient's lack of satisfaction with housing plans within days of obtaining housing.",		1	1
food pantries	observation	"Today, SW provided patient with list of following food pantries:",	Adult cystic fibrosis clinic	337	212
		"Other education or resources provided: Discussed food pantries and food stamps.",	Unnamed note		
		" He states that he sleeps "all over" and gets food from food pantries, bathes in public restrooms and washes his clothes in the creek. "	Occupational therapy		
food insecurity	observation	"Food insecurity: Worry: Not on file"	Psychiatry initial consult	3193	1598
		"Food insecurity: Worry: Never true "	Psychiatry initial consult		
food stamps	observation	"Couple does not receive disability or food stamps.",	Social work psychosocial assessment	3923	1693
		"I did just receive a call from the primary care nurse stating that pt was saying he was going to sign out AMA because he needed to go take care of his food stamps. ",	Unnamed note		
		"She states that she does receive food stamps in the amount of $504/month."	Social work psychosocial assessment		
lack of caregivers	observation	"Recent Psychological Factors: Lack of Caregivers, Other (Comment) Patient reports he is not married, no children, lives with his partner.",	Social work psychosocial assessment	1057	755

		"Lack of Caregivers / Caregiver burn out"	Social work psychosocial assessment		
		"Recent Psychological Factors: Family issues/concerns, Lack of Caregivers, "	Social work psychosocial assessment		
pays out of pocket	observation	"CM explained that right now, the patient does not qualify for any benefits and we cannot place him in a facility unless the family pays out of pocket.",	Case management note	2216	1504
		"Prescription Coverage: Patient pays out of pocket"	Social work psychosocial assessment		
		"Prescription Coverage: Patient pays out of pocket"	Social work psychosocial assessment		
access to transportation	observation	"Her supports include a history of education, as well as access to transportation."	Occupational therapy	n/a	n/a
		"Her supports include supportive mother, as well as access to transportation and outpatient therapy."	Occupational therapy		
		"Reason: Access to transportation"	Transitional care clinic behavioral health progress note		
self pay	observation	"Reason for referral based on assessment: Mental health issues (Comment), Self-pay/financial issues, Poor health literacy",	Social work psychosocial assessment	2211	1218
		"Reason for Referral: Mental Health Issues, Substance Abuse Issues, Self-Pay / Financial Issues",	Social work psychosocial assessment		

		"He would be self pay."	Unnamed note		
financial stressors	observation	"She said the patient's stressors might include expectations of being "the man of the family", financial stressors, and feelings like "he has to help everyone."",	Unnamed note	4657	1266
		" Pt reports financial stressors in addition to environmental stressors at his parents house. ",	Recreational therapy evaluation		
		"Financial stressors"	Diabetes stressors		
financial concerns	observation	"Stressors: relationship strife, financial concerns, lack of outpatient mental health follow up, recent medication changes",	Psychiatry daily progress note	6929	3949
		"Recent Psychological Factors: Feelings of hopelessness about future and/or goals, Family issues/concerns, Financial concerns",	Social work psychosocial assessment		
		"Will need follow-up with orthotics, has seen Marco before, for offloading; financial concerns, and he will work with staff along these lines"	Assessment/Plan		
hoarder	observation	"Apparently he states" you would think I was a hoarder,and is also overwhelmed with the work that needs to be done to clear his house and sell it.",	Unnamed note	n/a	n/a
		"Pt's son is cautious of him returning home due to fall risk and inaccessible home environment as he claims his father is a "hoarder" and there is no room to negotiate an AD in the home",	Physical therapy note		

		"The patient did report that her home is extremely cluttered and she describes herself as a "hoarder" with some psych component. "	Physician discharge summary		
taken into custody	observation	"Notably, he ran from Raleigh PD at the scene because he was driving without a license and so he was taken into custody upon discharge from the ED.",	Assessment/Plan,	203	128
		"Patient will remain in custody and will be taken into custody when medically stable again and taken to jail.",	ED clinical impression,		
		"Would limit visitors or at least have a sitter Girlfriend in room in ER is the person taken into custody by police at Wake for injecting into his PICC"	Unnamed note		
limited social support	observation	"A thorough evaluation has been completed of risk and protective factors including, but not limited to these risk factors: strong family history, apparent suicide attempt, limited social support-- and these protective factors: female. In my judgment the patient is at an acutely elevated risk of dangerousness to self (and/or others).",	Psychiatry initial consult	10983	1919

		"68-year-old female with psychosocial stressors and limited social support , unwilling to make any adjustments in her living situation due to the fact that she does not want to part with her 3 dogs which appear to be impacting her ability to secure a safe living arrangement",	Psychiatry initial consult		
		"Obstacles: Limited social support in NC;"	Social work psychosocial assessment		
housing insecure	observation	"Psychosocial Stressors: Recent move, Housing Insecure",	Social work psychosocial assessment,	1011	752
		"Psychosocial Stressors: Chronically Homeless, Coping with health challenges/recent hospitalization, Housing Insecure",	Social work psychosocial assessment,		
		"Psychosocial Stressors: Feelings of hopelessness about future and/or goals, Housing Insecure, Trauma (recent or history of)"	Social work psychosocial assessment,		
on disability	observation	"Social History: Lives in Raleigh, on disability",	Diabetes education consultation,	1320	280
		"He has been on disability since 1995.",	Social work psychosocial assessment,		
		"Occupational History RN (cardiac) until 2013 after which she went on disability"	Psychiatry initial consult		
court date	observation	"says he "needs to leave by tomorrow morning because I have a court date"",	Daily progress note,	19430	6535
		"Legal: (pt had past court date for possible eviction notice)",	Social work psychosocial assessment,		

		"Pt. Also stated he is court involved and has court date on 2/22/16 due to addiction related behaviors. "	Continuing care/discharge planning		
eviction	observation	"who presents for evaluation of suicidal ideation with a plan to cut his wrists in the context of job loss, eviction from sober living house and now homeless, and misdemeanor legal issues.",	Psychiatry emergency service initial consult,	2340	617
		"Occupational Profile Summary: Ean reported history of regular suicide attempts since Sept 2017 in the context of leaving his job due to not being able to physically complete it, loss of girlfriend, and eviction from his apartment. ",	Occupational therapy evaluation,		
		" While in the emergency department, the patient was served with an eviction notice from his mother's house."	Psychiatry discharge summary		
no insurance	observation	"No insurance/ no income",	Treatment team review,	8831	4343
		" X, MAC worker, called and asked to speak to patient due to patient having no insurance.",	Unnamed note,		
		"He wrecked a car that mom had cosigned, and there was no insurance on the car"	Psychiatry initial consult		
job loss	observation	"Stressors: Recent diagnosis, loss of job, loss of insurance, end of significant relationship, recent move.",	Psychiatry emergency service initial consult,	1412	427

		"Pt also reports recent job loss, x1 month ago due to use.",	Initial psychosocial assessment evaluation,		
		"Patient has been drinking heavily over the last 6 weeks due to a job loss. "	Unnamed note		
lost insurance	observation	"At this point in time Topamax was increased from 50 mg daily to 100 mg daily and patient was discharged with plan to follow-up at Kernodle clinic for neurology as she reports that she recently lost insurance. ",	Initial consult note,	342	184
		"Barriers to taking medications: Yes (Comment) Lost insurance in June 2017",	Care management initial transition planning assessment,		
		"She has not been working and pt lost insurance coverage at that time."	Unnamed note		

APPENDIX 6: Terms found on observation after bigram word embedding expansion

Highlighted terms mean they were included

Bigram	Sample notes	Clinical Note Type
currently_homeless	"he is currently homeless"	Emergency department encounter
	"Patient is currently homeless"	Emergency department provider note
	"Living situation: the patient is currently homeless, recently lost his home of 20 years due to be unable to make payments"	Crisis and assessment services initial assessment
homeless_shelter	"The pt was recently discharged from 1North on 11/08 to a homeless shelter"	Behavioral health assessment team
	"Post Acute Facility: Homeless shelter, Substance Abuse Treatment Facility, Other (Oxford House.)"	Care management: continued transition planning assessment
	"She reports when she was unable to get into the homeless shelter."	UNC Wakebrook primary care initial consult note
lives_alone	"Pt states that she lives alone, but daughter and granddaughter live nearby (5-10 minutes) and could come by if she needed them to."	Physical therapy
	"Pt lives alone"	N/A
	"Living situation: the patient lives alone."	Palliative care consult

Oxford_house	"Question/Concern for provider (Specific): Patient is in an Oxford House facility and his mother wanted to let Dr. X know that 3 of the guys there with him have been diagnosed with MRSA and she would like a call for advice. "	Patient advice request message
	"Other prior treatment(s): AA, referral to Oxford House after WakeBrook discharge"	Rex psychiatry initial consult
	"The patient was accepted for enrollment into an Oxford House (a substance abuse recovery house) on 04/17/2017,"	UNC health care addictions detoxification unit at wakebrook psychiatric discharge note
Ifc_shelter	"Provided pt with phone number to the IFC shelter where he has been staying so he can notify them of his whereabouts."	UNC health care psychiatry follow-up consult
	"Pt says he wants to go to the IFC shelter and has been trying to get in there for the last few months. "	N/A
	"He is still living at he IFC shelter in Chapel Hill but may soon move back in with his parents."	UNC Health care psychiatry established patient evaluation
Durham_rescue	"Previously staying at Durham Rescue Mission x 1 month"	N/A

	"Client reported that he was not interested in going to the Healing Place or the Durham Rescue Mission"	N/A
	"She could not go back to the Durham Rescue Mission for the same reasons."	Psychiatry discharge summary
Subsidized_housing	"They have also gotten him on some waiting lists for subsidized housing, but he knows those waits are long. "	Rex health care psychiatry follow-up
	"She said pt was evicted from a subsidized housing in November 2017 and has had both housing and medication adherence instability since, with multiple statements to CM about owing people money, which has made it difficult to obtain any stable residence until last week at a boarding house."	Psychiatry emergency service follow-up consult
	"For several months he's been living in subsidized housing in a very isolated area near Asheville North Carolina."	Psychiatry/PATHS discharge summary
unc_charity	"Need for financial assistance?: No - Patient states she has applied for UNC Charity Care program, to help with co-pays. "	Care management initial transition planning assessment
	"UNC Charity Care, ADAP (AIDS Drug Assistance Program) and Pharmaceutical company patient assistance program"	N/A

	" She stated that she has been approved for UNC charity care."	N/A
helping_hands	"Referred to Cancer Society Fresh Start program and Helping Hands Clinic lecture."	N/A
	"SW mention possibility of receiving financial assistance from HNC/HFA-Helping Hands."	N/A
	"Lantus pen to be sent home with patient-send prescription with patient to be filled at Helping Hands Clinic on Monday (they are closed today and will reopen Monday; this will be free of charge to patient.)"	N/A
stealing	"my daughter is stealing money from me - she has my debit card now but said she didn't';"	N/A
	"Counselor called son who reports he and other family do not talk with patient because she always accuses family members of stealing from her."	Psychiatry emergency service initial consult
	"Pt reports she was recently released from jail, discontinued her meds and was living at Urban Ministries in Greensboro until she started accusing staff and peers of stealing. "	Psychiatry emergency service initial consult
infested	"Patient has large scabs all over her body that she believes are infested with bugs, continually scratching the sores."	N/A

	"EMS personnel also expressed concern for living conditions in that the house was ill kempt and 'infested with termites'."	N/A
	" Items such as hats, grooming aids, and towels that come in contact with the hair of an infested person should not be shared"	N/a
robbed	""i have been robbed". "	N/A
	"She does state that she had a head injury from being robbed and beaten in 2005 and that since that time she has trouble getting her thoughts together and then when she starts talking she can't stop. "	Psychiatry initial consult service
	"My husband passed in 2015, last year I was robbed coming out of food lion, and now my children took my keys away and wont let me drive or do anything for myself.""	N/A
pay_rent	""I'll pay rent if they want me to.""	Psychiatry initial consult service
	"Pt expressed concern about getting back to her apartment soon because she needs to pay rent."	Social work psychosocial assessment service
	"Recently discharged from inpatient unit in Statesville, evicted from apartment since son	N/A

	who recently became her payee did not pay rent."	
lost_job	"Vocation Unemployed (lost job last hospital admission. "	Occupational therapy evaluation
	"Not employed; lost job few weeks ago due to background check revealing pending charge for larceny"	Rex express care note
	"Usually lives at home alone- but has son who recently lost job who is living with her currently; uses walker; has 3 children total; retired from NC Museum of Life and Science"	NC Heart and vascular admission history and physical
financial_strain	"She is taking steps to ease their financial strain and hopes to sell their house and find another place to live."	N/A
	"He acknowledged situational depression arising from recent unexpected job loss, with resulting financial strain, but denied	Rex psychiatry initial consult
	that it had affected his self-care, eating habits, rx compliance, or future outlook, and also repeatedly denied that he had ever felt suicidal"	Recreational therapy evaluation
	" Per chart, pt's primary stressors are unemployment, financial strain, and problems with primary supports."	

legal_charges	"Stressors: Substance abuse, legal charges"	Crisis and assessment services initial assessment
	"Per the NC Court Calendars, there are no pending legal charges or court dates."	Social work psychosocial assessment
	"Denies pending legal charges. "	N/A
unstable_housing	"Unstable housing."	Social work psychosocial assessment
	"Unstable housing during last several weeks."	Crisis and assessment
	"Stressors: chronic drug use, unstable housing"	Crisis and assessment initial assessment
condemned	"they may feel condemned and that life is a strain."	Neuropsychological assessment
	"Ms. X says the state of pt's house is so bad that she's surprised it hasn't been condemned."	N/A
	"He admitted that he has "pushed God away" because he has been told and believes that he is condemned for identifying as "pansexual.""	N/A
pay_mortgage	"Patient states that the last two weeks have been difficult, dealing with the anniversary of the death of one of her twins, financial issues with inability to find work despite having 6+ interviews and relying on family to pay mortgage on her and spouse's home, brother is	Psychiatry daily progress note

	in a stressful relationship as well using her as support system."	
	"SW reminded pt that this would likely be a one time assistance and asked pt if this would put her and her partner in a better place next month so they would be able to pay mortgage. "	N/A
	"Presenting problems: despair, hopelessness, grief, loss and shame specifically, illness results in X being able to pay mortgage on the home his wife and kids life in"	Spiritual care progress note
payments	"He has reported that he has medication coverage and money for his co-payments but sometimes needs new prescriptions."	N/A
	"Now she worries about how to make her mortgage payments. "	Psychiatry initial consult
	"Patient reports that she is in need of assistance with her co-payments to see specialist."	N/A
flooded	"Pt reports having hepatitis (he is unsure of the type) after walking in a flooded area over 30 years ago. "	N/A
	"Becky's Motel, where she and her long-term boyfriend had been living, was flooded and the patient has been staying in various places"	Psychiatry initial psychiatric evaluation

	"In the week prior to admission she mentions he has been sleeping 2-3 hours per night, if at all, and has flooded a hotel room after leaving the water on, lit a mobile fire pit on the family's enclosed back porch, and has become erratic in driving with children in the car."	Psychiatry daily progress note
receives_disability	"Pt receives disability and has a cane at home."	Social work psychosocial assessment
	"Pt receives disability and does not drive, family provides transportation."	Social work psychosocial assessment
	"He receives disability. "	Psychiatry initial evaluation
receives_ssi	" Pt receives SSI and Medicaid. "	Social work psychosocial assessment
	"She receives SSI income."	Psychiatry initial consult
	"Employment history, current income, financial resources: Pt reports she receives SSI - $735.00 a month. "	Social work psychosocial assessment
public_housing	"Living situation: Lives with his 2nd wife but they are going through a divorce and he is looking for public housing."	Psychiatry initial consult
	"Pt's niece was able to decipher that pt had questions about public housing, food stamps and a car seat for the daughter. "	Beacon adult abuse consult

	"stated he can't stay with brother because he lives in public housing"	Care management initial transition planning assessment
housing_authority	"worked for NYC housing authority"	Neurology consult note
	"The Housing Authority of the County of Wake's Public Housing Waiting List is open."	N/A
	"Addendum: SW provided pt with list of housing authority resources in Raleigh, Wake and Johnston County. "	N/A
boarding_house	"He rents a room in a boarding house"	Raleigh infectious diseases consult note
	"Patient also requested a boarding house list from case management, reporting that her family might pitch in $400 to have her stay at a place temporarily."	Psychiatry brief progress note
	"The patient had a recent eviction from a placement in a boarding house difficult."	Psychiatry discharge summary
halfway_house	" Freedom House: (Now has intensive out-pt treatment as well as halfway house)"	Physician discharge summary
	" Freedom House: (Now has intensive out-pt treatment as well as halfway house)"	Physician discharge summary
	"A new halfway house that her sponsor is working with has a bed available now."	N/A

healing_transitions	" Mother expressed interest in Healing Transition."	Care coordination discharge planning
	"Patient was picked up by van to be transported to The Healing Transition."	N/A
	"Seen by Dr Zarzar,pt calm and coop,denies SI/HI/AVH,no N/V,no falls,no tremors noted or observed,attended partially 1 group,tol tx well,adequate diet,described mood as good and ready to back to healing transition,no diarrhea,feels safe here,discharge teaching completed and pt verbalized understanding of the teachings,no physical medication given but pt stated that his medications are at the place."	N/A
rescue_mission	"3)RALEIGH RESCUE MISSION"	Wake brook ADU discharge note
	"2. Western Carolina Rescue Mission 225 Patton Avenue Asheville, NC 28804 on Friday, 5/19/17 (check in between 1-3 pm)"	N/A
	"PT reports that she was discharged from PATHS on 1/9 and returned to the rescue mission but that her medications aren't helping with her hallucinations and she is feeling "out of control"."	N/A

losing_job	"Is worried about losing job at the mall (the one she likes) and is worried she has already lost it based on her manager telling her not to come in today"	Psychiatry initial consult
	"Stressors:Employment difficulties and possibility of losing job"	Psychiatry discharge summary
	" Pt report increased depression since losing job and continued struggle with alcohol use. "	Psychiatry emergency service initial consult
medication_affordability	"She also reports medication affordability as a factor in her noncompliance."	Psychiatry initial consult
	"Medication Affordability: Ms. X was approved for the UNC Pharmacy Assistance Program."	Care management progress note
	"He declined the option of initiating Campral therapy for treatment of his alcohol use disorder following completion of detoxification, citing concerns related to medication affordability"	Psychiatric discharge note
resource_strain	"Financial resource strain: Not on file"	Pulmonary and critical care medicine consultation
	"Financial resource strain: Not on file"	Consultation note
	"Financial resource strain: Not on file"	Consultation note
receives_ssdi	"Pt receives SSDI and medicaid."	N/A

	" Pt receives SSDI and has Medicare Advantage plan. "	CM screening assessment
	"Employment history, current income, financial resources:Pt receives SSDI and VA benefits."	Social work psychosocial assessment
Unable to make payments	"(Pt stated she had been	Social work psychosocial assessment
	unable to make payments to insurance due to continued hospitalizations; therefore, insurance was not reinstated.	Heart failure consultation
	Informed she would be able to reapply in approx 6w"	Psychiatry initial consult
	"He previously used to lease a truck but was unable to make payments on it so had to return it. "	
	"Pt then got a bedroom suite, let it go back when unable to make payments""	
Lost his home	"Stressors he reports that he lost his job in June, lost his girlfriend and then lost his home."	Psychiatry discharge summary
	"The anxiety and depression have been exacerbated over the past several months as he has lost his home due to foreclosure."	N/A

	"Living situation: the patient is currently homeless, recently lost his home of 20 years due to being unable to make payments"	Crisis and assessment services initial assessment
Lost her home	"Pt reports she lost her home and all belongings in Hurricane Matthew, her fiance left her yesterday."	N/A
	"65-year-old widowed female who reports that she has too many stressors, she has become homeless and had been living in a motel having lost her home which reportedly had to be condemned."	N/A
	" In the last several months, patient has lost her home, as file for bankruptcy and feels she is feeling children."	Emergency department encounter assessment plan
home_alone	" Recommend SNF following D/C as pt is not safe to return home alone at this time."	Occupational therapy
	"Living situation: home alone"	Consult note
	"Pt has been safe to be home alone in the day while daughter works, Daughter assist with dressing, bathing, meals and pt able to ambulate in house to bathroom Ily."	Occupational therapy
bankruptcy	"He also notes that he does not have health insurance currently and is going through bankruptcy proceedings"	Surgery consultation assessment/plan

	"He filed for bankruptcy six to seven years ago; it has been resolved."	Social work kidney transplant assessment
	"He is going through bankruptcy because his business failed. "	Psychiatry initial consult
Lack of stable housing	"Though pt was denying SI at the time of discharge from Medicine, he remained depressed, and felt that psychiatric hospitalization would be helpful, and given serious suicide attempt and lack of resolution of risk factors like lack of stable housing, he was admitted voluntarily to the UNC Geropsychiatry unit for safety, stabilization, and management of major depressive disorder with suicidal ideation "	Psychiatry discharge summary
	"The patient has several stressors which can trigger her hopelessness, such as lack of stable housing, separation from dog, and recent loss of wallet."	Wakebrook FBC individual note
	"Therapist assessed patient's mood which was reported to be ◊pretty good considering I◊m homeless◊ but stable with some anxiety related to limited finances and lack of stable housing. "	N/A

financial_strain	"This would be 3 times a week, and is an obvious financial strain for them."	N/A
	"She is worried about the financial strain this will cause her family. "	Psychiatry initial consult
	"Stressors: Two young children in the home; recent repeated hospitalizations; rheumatoid arthritis; husband◊s recent job loss and financial strain."	Psychiatry psychotherapy note
no long term housing	1 pt	
transportation_problems	"Son unable to come to hospital at this time due to transportation problems"	Patient care overview
	"PT HAS TRANSPORTATION PROBLEMS INCLUDING DEXA, CHEST CT, DOC APPT ETC."	N/A
	"She has been referred to pain clinic in Raleigh but has been having transportation problems."	N/A
Lack of social support	Already grabbed	
social_isolation	"Social Isolation-pt stays home with her daughter and does not go out."	n/a
	"08/02/16 0123Suicide Risk (Adult,Obstetrics,Pediatric)Suicide Risk: Related Risk Factorsco-occurring disorders;mental health diagnosis;multiple stressors;previous suicide attempt;substance	Patient care overview

	use/abuseSigns and Symptoms (Suicide Risk)overwhelming hopelessness/helplessness;narrow thinking;social isolation;substance use/abuse;suicidal ideation/intent/plan"	
	"He also noted social isolation; most of his friends are dead or moved away, and he is estranged from most of his family; being "totally alone is killing me.""	Psychiatry initial consult
geographical_isolation	1 patient	
finanical_assistance	"Need for financial assistance?No"	N/A
	"Need for financial assistance?N/A"	N/A
	"Need for financial assistance?No"	N/A
urban_ministries	"Alliance Medical Ministry, The Healing Place, Mariam clinic, Open Door Clinic of Urban Ministries, Raleigh Rescue Mission, The Salvation Army"	Primary care discharge consult note
	"SW reviewed qualifications for Urban Ministries with pt, however pt does not meet their criteria because she has medicaid."	Case management note
	"Alliance Medical Ministry, The Healing Place, Mariam clinic, Open Door Clinic of Urban Ministries, Raleigh Rescue Mission, The Salvation Army)"	Primary care discharge consult note

finanical_constraints	"He has been unable to follow-up with Pulmonary due to financial constraints."	Hospitalists discharge summary
	"Feeling depressed most days by current health status, financial constraints and spouse (they are separated) has dementia and is living in home in Ohio."	Follow-up visit alignment CDM programs
	"Care is limited by financial constraints and lack of transportation-uses CARTS transportation"	N/A
not_employed	"X voices that Pt DTR X is not employed, and the entire family has become dependent on Pt financially"	Social work psychosocial assessment
	"Occupational History◊Disabled◊Not Employed"	Psychiatry history & physical
	"He is currently not employed or enrolled in school. "	Social work psychosocial assessment
Loss of job	"Stressors: IV drug use, medical problems, legal problems (contributng to recent loss of job)"	Crisis and assessment services initial assessment
	"Stressors: Divorce of parent, transition to college, loss of job, and loss of friend by suicide in 2012. "	Crisis and assessment services
	"Obstacles: Recent loss of job"	Social work psychosocial assessment

legal_problems	"Using marijuana may cause legal problems."	Anxiety disorder: care instructions
	"He denies access to firearms, legal problems, history of military service or other trauma."	Psychiatry initial consult
	"Client denies legal problems."	Psychotherapy family therapy progress note
mortgage_assistance	"Since our last meeting, the X's have moved forward with seeking mortgage assistance from the NC Mortgage Foreclosure Assistance Program. "	N/A
	"SW contacted pt wife for follow up regarding mortgage assistance and to inform family that CCF will be making mortgage payment. "	N/A
	"SW applied on pt's behalf to CCF for mortgage assistance."	N/A
Does not drive	"Patient does not drive."	Occupational therapy
	"Prior functional status: independent ADL's, no AD, does not drive"	Physical therapy evaluation
	"Pt. does not drive."	Occupational therapy
difficulty maintaining employment	"Primary concerns included difficulty maintaining employment, challenges with emotional self regulation, social communication difficulties, disrespectful language towards grandparents whome he lives	Psychiatric diagnostic interview

	with, and refusal to complete daily living skills."	
	"Vocation Unemployed (previously worked as a nanny but has had difficulty maintaining employment since October)"	Occupational therapy evaluation
	"Without this level of support, Susan would likely have difficulty maintaining employment."	Psychological evaluation report
employment_difficulties	"These financial hardships and employment difficulties have compounded the patient's depressive symptoms."	Psychiatry history & physical
	"Patient became tearful when discussing her current employment difficulties."	Social work consult note
	"Patient's Stressors / Triggers: Per H&P pt with barriers to being able to see his children, employment difficulties, lack of support system as stressors"	Recreational therapy evaluation
financial_issues	" He has psychosocial stressors of financial issues, limited support and grief. "	N/a
	"Reason for referral based on assessment: Self-pay/financial issues,"	Social work psychosocial assessment
	"Reason for Referral: Self-Pay / Financial Issues (Readmit < 30 days)"	Social work psychosocial assessment

fixed_income	"Pt is on a fixed income of $1563/month for disability."	Case management initial assessment
	"She recently retired, relocated from Maryland and started supporting her adult son; an additional dependent on an already fixed income has made it difficult for pt to afford food each month, making it especially challenging to incorporate fruits and vegetables into her diet."	Outpatient adult nutrition-initial assessment
	"Fixed income"	N/A

CHAPTER 3: IDENTIFICATION OF SDH USING MULTI-LABEL CLASSIFICATION OF EHR CLINICAL NOTES

Introduction

Emergency departments (EDs) are often called the 'safety net' of the U.S. health care system. Patients with poor mental health, depression, and high ratings of psychological distress have greater odds of being frequent ED users[1]. As a result, this population makes up one in eight ED visits contributing to ED overcrowding, poor quality of care, and higher costs[2]. Phelan and colleagues argued that social conditions related to poor socioeconomic resources such as money, social ties, and knowledge are "fundamental causes" of disease[3]. Social, psychological, and behavioral factors, or social determinants of health (SDH), are key contributors to health but are rarely measured in a systematic way in health care settings[4,5]. For example, housing insecurity is associated with poor health including chronic diseases, substance abuse, and frequent ED visits[6]. While, Ku and colleagues[7] found that frequent ED patients expressed a variety of other social needs including the inability to meet essential expenses, having a telephone service disconnected, worrying about running out of food, and inability to afford a balanced meal. These related disparities have a direct link to health and often result from overlapping factors[8,9]. Factors such as a patient's housing stability, income level, and insurance status must be considered when providing treatment and care.

In 2014, the National Library of Medicine underscored the importance of capturing these SDH in electronic health records (EHRs) to improve clinical care[5,10]. However, the extent to which SDH are encoded in EHRs is unknown, SDH data may only be available in clinical notes[11–13] in free-text form. Feller and colleagues examined the performance of SDH models using structured and unstructured data

finding that while there was higher performance using both data structures it was not statistically significantly higher than using text only[12]. This finding is supported by Vest and colleagues who recognize the importance of structured data found it difficult to extract and variable across health IT systems[14]. Therefore, hybrid techniques that blend natural language processing (NLP) and machine learning (ML) methodologies are needed to automatically extract SDH information. Various studies effectively applied NLP approaches, including information extraction techniques, to different types of SDH classification including homelessness[13,15,16], employment status[16,17], and exposure to violence[13,18]. These techniques included regular expressions, named entity recognition, and distributional semantic techniques. Several challenges appear during the design of such tools. Patients with SDH, such as someone who has lost their job (employment insecurity), suffer frequently from several SDH in relation to the loss of a job such as the loss of health insurance associated with employment. New approaches in ML, such as multi-label learning (MLL) may be a viable candidate for modeling the profile of a patient affected by several SDH. MLL differs from classical ML by tackling the learning problem from a different perspective. In contrast to the classical classification tasks where each observation belongs to only one mutually exclusive class, in MLL decision areas of labels (i.e. classes) overlap. A traditional approach to solving the multi-label text classification problem is binary relevance, which decomposes the problem into multiple independent binary classification tasks (1 for each label)[19,20]. A review of multi-label learning algorithms can be found in Min-Ling & Zhi-Hua[20].

In this paper, we investigate how to leverage clinical notes using novel applications of MLL to classify financial resource strain and poor social support among MHSUD patients who frequent the ED. We focus on the following: (i) assess the feasibility of developing a model to classify SDH using only clinical notes. We experiment with five approaches to classification: a linear SVM-baseline, K-Nearest Neighbors (KNN), Random Forest (RF), XGBoost, and Bi-LSTM; (ii) develop a multi-label setting (up to 6 labels per instance); (iii) apply the model to single sentences, the most granular level of clinical notes.

We rely on clinical notes from a large academic health system to validate our experiments with a gold-standard corpus; and (iv) highlight the elements in the sentence that explain and support the predicted labels to promote transparency. A characteristics of the healthcare domain is long sentences with a large number of technical words and typos/misspellings. We experiment with simple yet effective preprocessing of the input texts. While research exists for each individual SDH characteristic in our model, we believe we are the first to tackle the multi-label in the clinical domain. Our results show the feasibility of developing ML models to classify clinical note sentences with multiple SDH labels with XGBoost, SVM, and Bi-LSTM yielding the most promising results.

Previous Work

Patient record labeling

Most multi-label patient record classifiers fall in the tasks of phenotyping across multiple conditions, health conditions known to co-occur, at once. Zufferey and colleagues[20] evaluated multi-label classification algorithms for the analysis of clinical data of chronically ill patients and found that decision trees performed optimally considering all the evaluation metrics. However, SVM-based approaches were better when measured by Hamming loss (fraction of the proper labels to the total number of labels). More recently, Hong and colleagues implemented a multi-class and multi-label classification system to identify patients with obesity and multiple comorbidities from semi-structured discharge summaries[21]. Among the four ML classifiers (logistic regression, support vector machine, decision tree, and random forest), the Random Forest (RF) algorithm performed the best with an F1 micro-average of 0.95 to predict obesity and its 15 comorbidities. Liang and Gong effectively adapted multi-label text classification to effectively categorize patient safety reports using de-identified EHR data[22]. This demonstrated the feasibility and efficiency of multi-label algorithms in identifying low frequencies of adverse patient events. However, low prevalence events create a label imbalance problem, particularly in the context of multi-label

classification[22]. Therefore, our study targeted an SDH rich population to create an overly positive dataset and utilized purposeful sampling to address the imbalance of labels.

Biomedical texts

Multi-label classification has broad task applications in the biomedical domain, including biomedical literature indexing. Du and colleagues proposed ML-Net, an ML method that combined a label prediction network with an automated label count prediction mechanism to produce an optimal set of labels[18]. They evaluated this deep learning approach on biomedical literature and diagnosis code assignment using clinical notes and found high precision and recall (0.848, 0.850) for the biomedical literature task. However, the approach performed poorly for the diagnosis code task (precision = 0.577, recall= 0.442). The length of clinical documents and the likelihood of more labels per document may be one cause for the poor diagnosis code task results. ML-Net demonstrated the generalizability of this framework by applying it to multiple data sources, a similar approach was taken in our study.

Diagnosis code assignment

Patient clinical note length combined with large numbers of diagnostic codes challenge the development of automated multi-labeling processes. Baumel and colleagues investigated four models for the task of extreme ($\geq$ 20 labels per instance) multi-label classification on open source clinical notes[23]. They found that using a hierarchical neural network approach to tag a document by first identifying the sentences relevant for each label led to the highest micro-F1 results (55.86%). Although this result is low, their model provided full transparency for classification decisions, a characteristic that may be important for adoption by medical experts. In our study we evaluate our models by each label and as an overall algorithm with a variety of validated metrics to provide as much transparency as possible with a multi-label classification approach.

Methodology

Setting and sample population

Clinical notes were obtained from the clinical data warehouse at the University of North Carolina Health System, a large academic medical center serving much of North Carolina. Clinical notes created between April 2014 to December 2019 were collected, a time period that encompassed the health system's transition to a single EHR. Clinical notes that met the following inclusion criteria were retained: (1) visited University of North Carolina at Chapel Hill Emergency Department (UNC-CH ED) between 2014-2019. Patients who had less than four ED visits in the year 2017 or 2018 within a rolling 365-day period were excluded because they do not meet the standard for frequent vistor of the ED ($\geq$ 4 visits per year)[25,26]. (2) greater than 18 years old in the CDW-H as of 2014, and (3) documented MHSUD "final primary diagnosis" as defined by the International Classification of Diseases and Related Health Problems 10th Revision (ICD-10 CM) code F00-F99 "mental and behavioral disorders". The study was approved by UNC's Institutional Review Board.

Curation of social determinants of health

We created a gold-standard corpus of clinical notes containing SDH characteristics from MHSUD patients who frequent the ED. Sentences with a high likelihood of SDH characteristics were identified through a SDH dictionary that was developed by training two word embedding models (unigram and bigram) using seed terms abstracted from published research studies. These models detected and identified semantically similar terms to characterize financial resource strain and poor social support, yielding 109 terms or phrases (Chapter 2). In this study, we focused on multiple SDH characteristic classification of financial resource strain and poor social support. The selected labels included (1) housing insecurity (homelessness, unstable housing), (2) food insecurity (food stamps, unable to afford food), (3) employment and income insecurity (unemployment, insufficient income), (4) general financial insecurity (lack of transportation, other financial issues), (5) insurance insecurity (uninsured, underinsured), and (6)

poor social support (social isolation, lack of social support) as guided by the IOM's "Capturing Social and Behavioral Domains and Measure in Electronic Health Records"[26].

Gold-standard corpora purposeful sampling

Purposeful sampling is widely used in qualitative research to identify and select information-rich examples related to the target of interest[27]. In contrast, probabilistic or random sampling is used to ensure the generalizability of findings by minimizing the potential for bias in selection. In this study, we developed a data-level hybrid approach to address our imbalanced dataset. This imbalance occurs when one or more classes have very low proportions in the training data as compared to the other classes[28]. In this study, our target classes (SDH domains) are the minority classes and negative sentences are the majority, as in the case in rare event classification tasks.

The frequency of redundant text in clinical notes, created by copy and paste or auto-generation, presents challenges for training and evaluating machine learning models[30]. While this functionality has definite benefits for clinicians, among them more efficient documentation, it has been noted that it might impact the quality of documentation as well as introduce errors in the documentation process[30,31]. To derive an unbiased estimate of likely SDH documentation, we removed auto-generated and copy and paste entries that appeared to duplicate sentences. We removed sentences that were exactly the same as another sentence within an individual patient's clinical record regardless of the time period between the occurrences of these entries. We then isolated two corpora (one unigram, one bigram) with a combined 1,596,166 sentences with likely SDH documentation based on dictionary of terms and phrases developed through a SME driven word embedding expansion approach, described in Chapter 2. We then took a randomized sampling of 150-200 sentences from each SDH class pre-labeled by their associated dictionary term using the NLTK shuffle library. A randomized sampling of negative sentences (i.e., lacking a SDH term) were added to the dataset to adjust for over representation of SDH in the dataset. No

duplicates were found between the two corpora that were then annotated and combined for model training. This newly formed dataset was used by annotators to complete the annotation process and create a gold-standard corpus.

Gold-standard annotation guidelines

To produce higher quality SDH analysis and downstream applications, we chose to obtain sentence-level annotations rather than document-level annotations because we wanted to evaluate the feasibility of classifying SDH on a granular-level. For example, we observed mentions of SDH, such as "patient's current stressors include: unemployment, homelessness and recent relapse on illicit substances" and "patient reports that he lost his job in June, lost his girlfriend and then lost his home," that would not be amenable to extraction by document or named-entity recognition.

Two annotators manually reviewed extracted clinical note sentences to classify documentation using six SDH characteristic categories described earlier. A third annotator adjudicated disagreements to determine the final classification. The annotators represented an interdisciplinary group of health professionals that serve the study population: a clinical social worker for UNC ED (MH), a paramedic and PhD candidate in Health Informatics (RS), and a registered nurse and clinical informatician (RK). Annotators (RS, MH) read each clinical note sentence in its entirety to assess the presence of SDH documentation. Any confirmatory mention of SDH associated, regardless of status was treated as a positive finding, for example, "patient is currently homeless" or "patient states he has been homeless in the past," resulted in a positive label for housing insecurity. Detailed annotation guidelines are in Appendix 1.

Collection of clinical notes

Clinical notes were obtained from the clinical data warehouse at the University of North Carolina (UNC-CDW), North Carolina's largest academic health system. We initially isolated two corpora that

were hypothesized to contain documentation of financial resource strain and poor social support (Chapter 2). Sentences were derived from a variety of note types such as "Emergency Department progress note," "Psychiatry initial consult," and "social work psychosocial assessment." Only a small proportion of all notes collected from the EHR system were used in this study, as a result of a word embedding terminology expansion approach was used to identify a subset of notes likely to contain SDH documentation (Chapter 2). The output of the word embedding expansion approach was used by annotators to complete the annotation process and significantly increased the yield of SDH positive annotations compared to traditional manual annotation of all documents in a corpus[12,31]. The implementation details and evaluation results of our word embedding expansion approach are described at length in the previous chapter.

Input Texts

For clinical notes, we completed the following preprocessing steps: (i) tokenized all input texts using Natural Language Toolkit (NLTK)[32]; (ii) removed all punctuation from each sentence; (iii) removed all non-alphabetical tokens; (iv) converted all letters to lower case; (v) normalized text through stemming and lemmatization that transform words to their root forms; and (vi) removed English stop-words (i.e. me, my, myself, etc.).

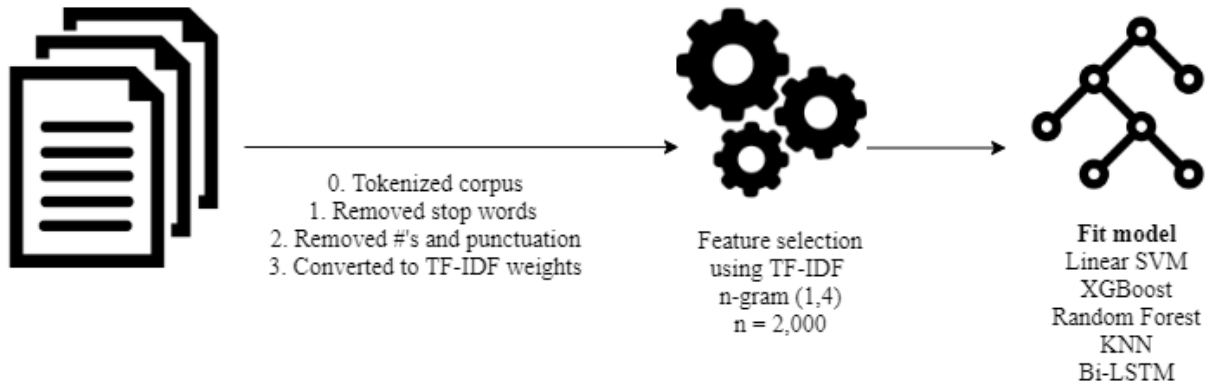
Outcome variables

In these experiments we trained a binary relevance model, an ensemble of single-label binary classifiers, one for each class. Each classifier predicts either the membership or non-membership of one class. The union of all classes that were predicted is taken as the multi-label output[19]. We classified whether a given SDH topic (e.g. housing insecurity and/or food insecurity) was either documented or not document in the clinical note sentence.

Experimental design

We developed and examined the outcome of five models: Random Forest (RF), XGBoost, K-Nearest Neighbors (KNN), Bi-directional Long Short-Term Memory (LSTM), and Support Vector Machine (SVM) as our baseline. These models were trained (80%, N= 3250) and tested (20%, N= 813) on a randomized gold-standard corpus. To reduce bias or a less optimistic estimate of model performance a five-fold cross validation was used on the test dataset[34,35]. This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k-1 folds[34]. The choice of k is usually 5 or 10, but there is no formal rule[36]. Inputs into the classification models included a single free-text clinical note sentence. An overview of methods for developing a machine learning classifier to identify SDH in clinical notes is shown in Figure 1.

FIGURE 1. Overview of methods for machine learning



Evaluation

Precision, recall, and F1 scores were computed across the SDH models using five-fold cross validation. Because the decision to optimize precision or recall depends on the specific clinical

application, we considered F1 as the primary evaluation metric[12]. F1 represents the harmonic mean of precision and recall and takes both metrics into account. Since a multi-label neural network lacks a computational library from which to measure each label, we adopted 5 common evaluation measures: accuracy, Average Precision-Recall (AP), Area under Curve Receiver Operating Characteristic (AUC-ROC), Hamming loss, and log loss to compare the performance of different methods for multi-label SDH classification[19,20,36]. Accuracy calculates ratio of the prediction of true labels. In multi-label classification, this function computes subset accuracy; the set of labels predicted for a sample must exactly match the corresponding set of labels in the test set[37]. AP summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. Because the label distribution of SDH is skewed as described earlier, the micro average Area under Curve Receiver Operating Characteristic (AUC-ROC) was calculated for each class. Hamming loss represents the fraction of labels that are incorrectly predicted. Finally, log loss or cross-entropy loss, the loss function used in (multinomial) logistic regression and neural networks, is defined as the negative log-likelihood of a logistic model that returns testing probabilities for its training data. A full mathematical description of all evaluation metrics are found in Figure 2. All source code can be found at *www.github.com/rstem/dissertation*.

FIGURE 2: Evaluation metrics

$$\text{precision} = \frac{tp}{tp + fp}$$

$$\text{recall} = \frac{tp}{tp + fn}$$

$$F_\beta = (1 + \beta^2)\frac{\text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}$$

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

$$\text{Average Precision-Recall} = \sum_{n} (R_n - R_{n-1}) P_n$$

$$\text{AUC-ROC} = \frac{2}{c(c-1)} \sum_{j=1}^{c} \sum_{k>j}^{c} p(j \cup k)(\text{AUC}(j|k) + \text{AUC}(k|j))$$

$$L_{Hamming}(y, \hat{y}) = \frac{1}{n_{\text{labels}}} \sum_{j=0}^{n_{\text{labels}}-1} 1(\hat{y}_j \neq y_j)$$

$$L_{\log}(Y, P) = -\log \Pr(Y|P) = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} y_{i,k} \log p_{i,k}$$

In addition, we conducted an error analysis to gain insight into model performance for SDH labels. We reviewed all incorrectly labeled sentences and analyzed false negatives using a classification matrix and attempted to classify each error as an incorrect annotation, unrecognized negation, or confusing auto-generated structure.

Feature extraction. Text analysis is a major application field for machine learning. Unlike numeric text, raw text data, essentially a sequence of symbols, cannot be fed directly into algorithms. Most algorithms expect numeric feature vectors with a fixed size rather than the raw text documents with variable length[37]. Therefore, features were extracted through a bag of words, with tf*idf weights (determined from the corpus of clinical notes) for each label. To reduce the bias generated by various lengths of individual clinical note sentences we applied sublinear tf scaling (i.e. replace tf with 1 + log(tf))[34]. The lower and upper boundary of the range of n-values for the different n-grams to be extracted were (1, 4). For example an n-gram range of (1, 1) means only unigrams for the word occurrence matrix associated with characteristics of each label. The maximum features (2,000) were ordered by the term

frequency across the corpus. Since our corpus has exceptionally rare words, we applied a fitted vectorizer that initially added unwanted dimensions to inputs.

SVM. We used the Scikit Learn[37] to implement a one-vs-all, multi-label binary SVM classifier. The model fits a binary SVM classifier for each label (SDH category) against the rest of the labels. In practice, the SVM algorithm was implemented using a kernel that transformed an input data space into the required form. SVM uses a technique called the kernel trick where the kernel takes a low-dimensional input space and transforms it into a higher dimensional space. We utilized a linear Kernel Trick that tries to find a decision boundary between labels. A linear kernel can be used as normal dot product (product of the Euclidean magnitudes of the two vectors and the cosine of the angle between them) of any two given observations. This is a common baseline algorithm for developing machine learning based classifiers using clinical notes[12,38,39].

Random Forest. A RF classifier is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting[37]. A number of hyperparameters were optimized including a maximum depth for each tree (10), bootstrapping sub-samples, and sampling features to determine the best split at each node. RF classifiers have shown success in previous multi-label clinical note classification tasks[18,20].

XGBoost. The XGBoost classifier implements machine learning algorithms under the Gradient Boosting framework. Gradient boosting is a technique for regression and classification problems, that produces a prediction model in the form of an ensemble of weak prediction models such as decision trees[40]. We utilized the gbtree booster that uses a version of regression trees as a weak learner to learn on the ensemble residuals of the previous iteration of a decision tree. A number of hyperparameters were optimized to include a learning rate (0.5) used to prevent overfitting, calculating a maximum depth for each tree (10), and scaling the positive and negative weights of unbalanced classes

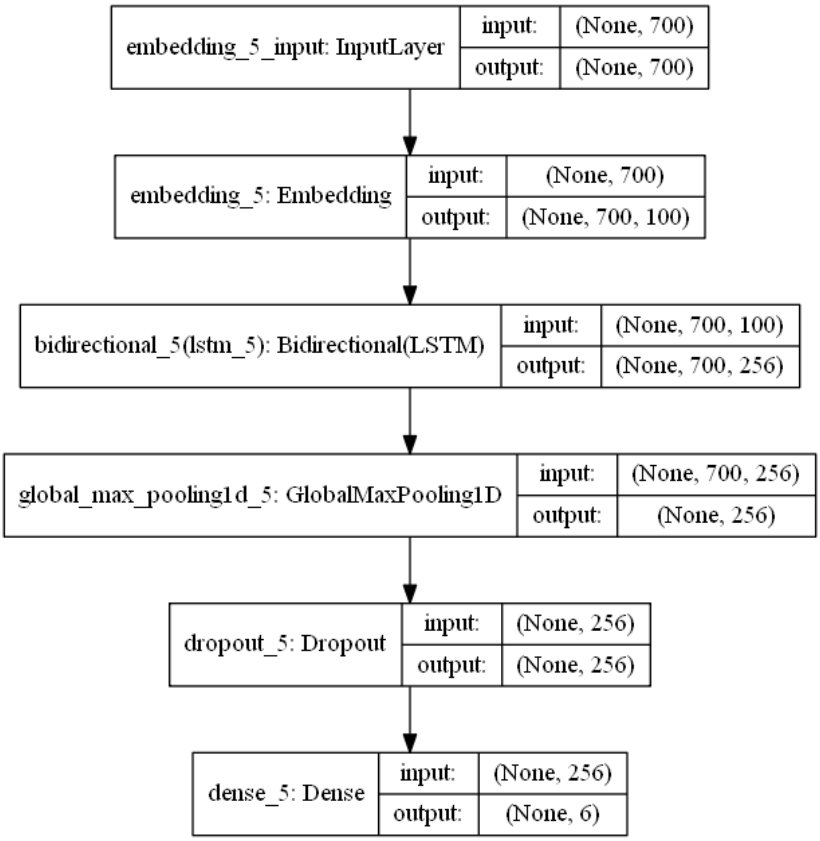
(sum(negative instances) / sum(positive instances)). Feller and colleagues[12] found that a gradient boosting tree algorithm had the highest performance when classifying SDH among patients with sexually transmitted infections.

KNN. K-Nearest Neighbors classifier is a type of instance-based learning or non-generalized learning: KNN does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned to the data class which has the most representatives within the nearest neighbors of point. A number of hyperparameters were optimized to include the number of neighbors (4) to use for kneighbors queries and a distance weight function used in prediction.

Bi-directional LSTM. An LSTM has a similar control flow as a recurrent neural network (RNN), as data are processed, it passing on information as it propagates forward. LSTM does not take into account the words that come after the current word in the sequence but rather uses only the information from previous words in the sequence. However, a Bi-directional LSTM takes information from both left-to-right and right-to-left directions covering both information for the previous words and future words into consideration at a particular step. For this reason, we have used a Bi-LSTM model to obtain more contextual information. The basic idea behind this approach is to encode the premise and hypothesis sentences separately and then combine those using a neural network classifier[41]. Recently, this deep learning technique was successfully applied to various fields, such as disease prediction and diagnosis code labeling, as it provides a more efficient learning mechanism for classification problems than classical machine learning methods[18,42,43]. Lipton and colleagues[44] used a LSTM to recognize patterns in multivariate time series of clinical measurements to train a model to classify 128 diagnosis thus demonstrating promise for clinical multi-label problems.

When applying Bi-LSTM, the words in the input sentence are first mapped to numerical vectors. These vectors can be random valued, a pre-trained word embedding, domain-specific word features or any combination of them[45]. A word embedding maps a word to a numerical vector in a vector space, where semantically-similar words are expected to be assigned similar vectors. To perform this mapping, we used a well-known algorithm, GloVe[46]. GloVe learns word embeddings by looking at the co-occurrences of the word in the training data. GloVe assumes that a word's meaning is mostly defined by its context and, therefore, words having similar contexts should have similar embeddings. GloVe can be trained from large, general-purpose datasets such as Wikipedia, Gigaword5 or Common Crawl without the need for any manual supervision. In this work, we experimented with different general-purpose, pre-trained word embeddings from the official GloVe website[46] and observed that the embeddings trained with Wikipedia 2014 and Gigaword 5 6B tokens, 400,000 vocabulary, uncased and 300 dimension vectors (cc) provided the best results with minimal training time (< 30 minutes). LSTMs have a chain like structure with four interacting layers designed to remove or add information to the cell state. Motivated by the strong results of the BiLSTM max pooling network by Conneau and colleagues[47] and Zhou and colleagues[48], we experimented with a global max pooling layer that takes the average of each feature map and feeds the resulting vector into the Sigmoid layer. We chose a Sigmoid function because it can be used to forget or remember the information that better facilitates low-memory modeling[49] (Figure 3). Low memory was of significant importance as this study was computed on a standard processor (GPU was not available); 10th Generation Intel Core i7 with 8GB RAM. Final hyperparameters were 20,000 features, vocabulary size 8315, maximum length of sentence characters 700, batch size of 128, and 9 epochs.

FIGURE 3: Bi-LSTM model layers



Results

Study population

Our gold-standard corpus (N= 4063) clinical note sentences represented 1119 patients of which 44.6% had at least one positive documentation of SDH. Characteristics of study patients are shown in Table 1. Half (N=548, 50.2%) were between the ages of 39 – 62 and primarily White or Caucasian (N=726, 66.9%).

TABLE 1. Study population characteristics

Characteristic	All (%)	SDH positive (%)
N patients	1119	1066
Age	51.4 (±15.9)	51.2 (±16)
18-29	106 (9.5%)	104 (9.8%)
30-39	180 (16.1%)	173 (16.2%)
40-49	222 (16.1%)	215 (20.2%)
50-59	263 (19.8%)	251 (23.5%)
60-69	203 (23.5%)	186 (17.4%)
70-79	95 (8.5%)	88 (8.3%)
>80	50 (4.5%)	49 (4.6%)
Sex		
Male	573 (51.2%)	542 (50.8%)
Female	546 (48.8%)	524 (49.2%)
Race		
American Indian or Alaska Native	6 (0.0%)	6 (0.0%)
Asian	4 (0.0%)	4 (0.0%)
Black or African American	304 (27.2%)	285 (26.7%)
Native Hawaiian or other Pacific Islander	1 (0.0%)	1 (0.0%)
Other race	29 (0.0%)	29 (0.0%)
Patient refused	1 (0.0%)	1 (0.0%)
Unknown	24 (0.0%)	24 (0.0%)

White or Caucasian	750 (67.0%)	716 (67.2%)

*Percentage based on non-missing data

Characteristics of gold-standard corpus

A total of 4063 clinical note sentences associated with 1119 patients treated at a large academic medical system were manually reviewed for characteristics of SDH; 502 clinical note sentences were associated with housing insecurity, 530 with poor social support, 321 with food insecurity, 686 with employment and/or income insecurity, 437 insurance insecurity, and 428 with general financial insecurity. 19.7% of SDH clinical note sentences in the entire corpus had two or more SDH labels documented; however, of positive SDH sentences (N=1066) 75.0% of them had two or more SDH labels documented (Figure 4). To balance an initially overly positive dataset, an additional 2252 negative SDH sentences were added to the corpus. Each sentence had an average length of 83.2 words with top N (frequent) words being patient, discharge, care, and history. All clinical note sentences in the corpus were double annotated by clinical experts, with an overall Kappa statistic of 86.6% agreement (79.1%-90.6%) (Table 2). The mean time our abstractors spent reviewing and coding notes was 65 seconds per clinical note sentence (~58 per hour). Annotators disagreed about 255 sentences. Disagreement occurred among sentences with the greatest length, an average of 145.8 words as compared to the corpus average of 83.2. Figure 5 shows the Pearson correlation coefficient between SDH labels with highest between general financial insecurity and poor social support (0.29).

Figure 4. Count of SDH per sentence

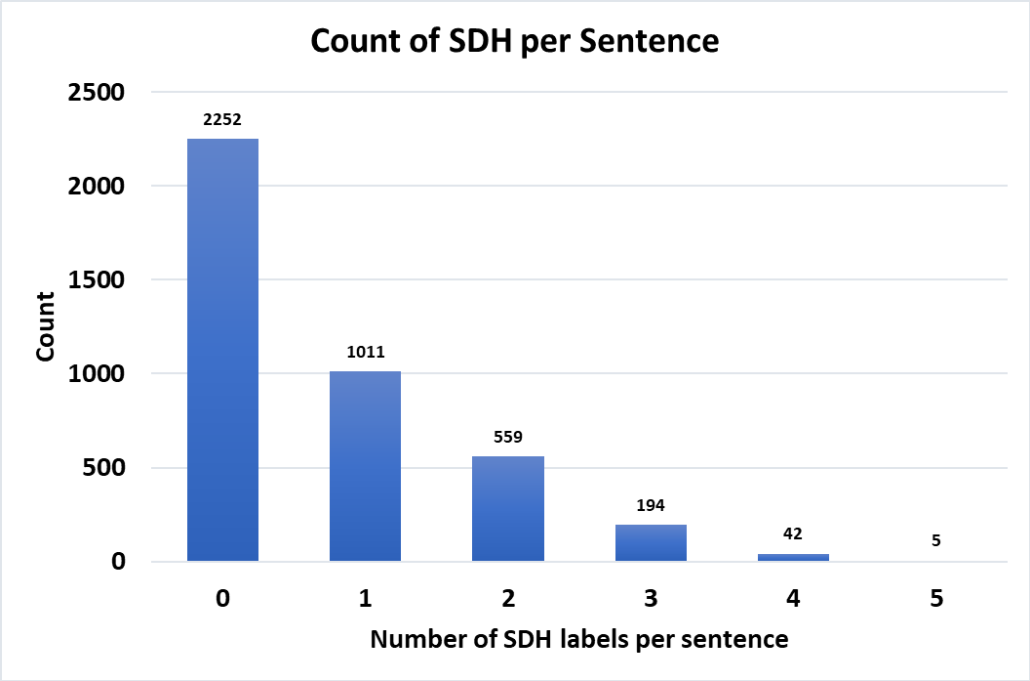
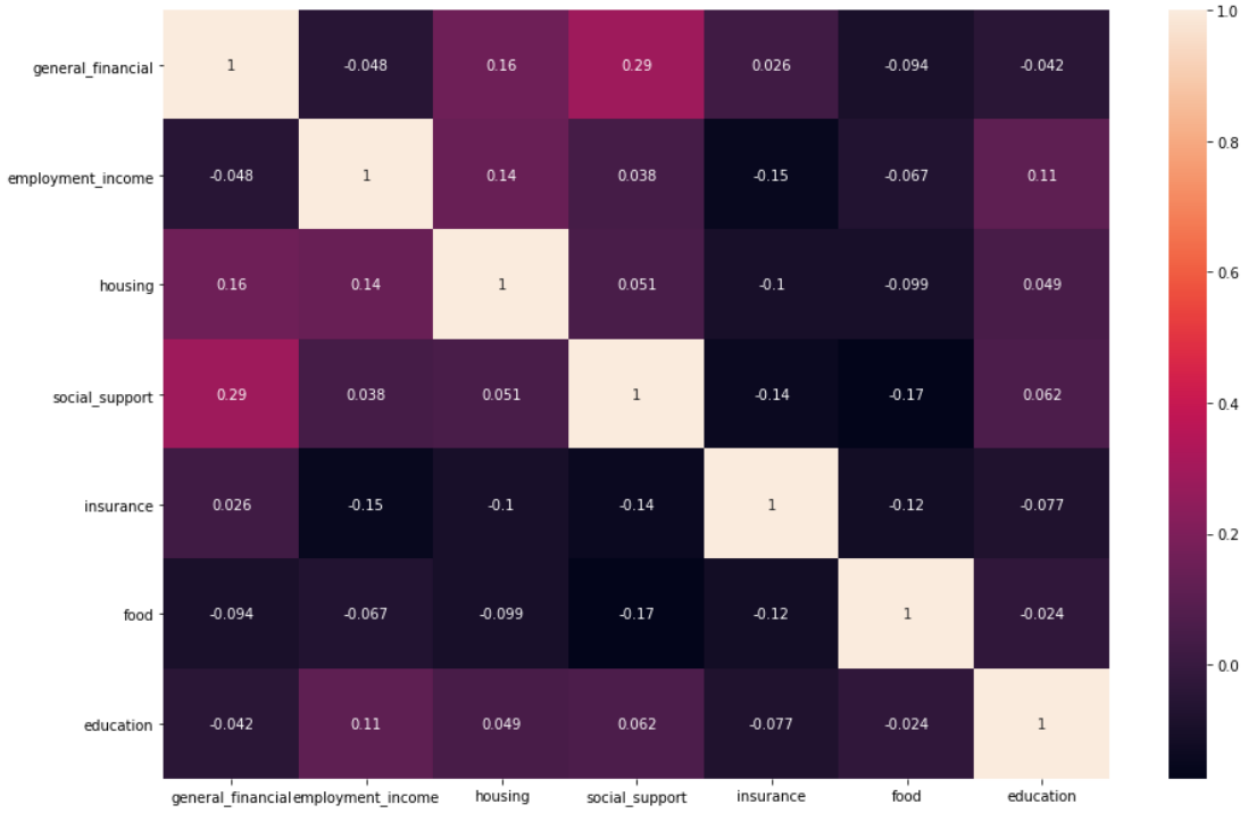


TABLE 2. Inter-rater reliability for individual SDH classes

General financial insecurity	Employment Income insecurity	Housing insecurity	Insurance insecurity	Poor social support	Food insecurity
84.4%	89.1%	89.2%	79.1%	87.1%	90.6%

Figure 5. Correlation matrix of SDH labels



Features used for SDH label classification

The features for the SDH classifiers are presented in Appendix 2. Text features used by the classifiers included explicit indicators of SDH, as well as co-occurring determinants. For example, top features for a sentence classified as poor social support and employment/income insecurity included "limited social support," "alcohol use disorder," "cocaine use disorder," and "financial concerns." Meanwhile other SDH labels had text features that were more closely associated with SDH risk factors. For example, top features for a sentence classified as food insecure included "food stamp," "afford," and "money."

Classifier performance

Classification results inferring the presence of topic-specific SDH documentation are presented in Table 3 and ranged from F1=0.82 for XGBoost to F1=0.45 for KNN (F1 micro-averaged across all labels). XGBoost had the highest average precision micro-averaged across all labels (0.85), while the highest average recall micro-averaged across all labels was SVM (0.88). XGBoost had the lowest Hamming loss (4.13) with all other algorithms having nearly double the loss (Table 3). When comparing the precision-recall micro-averaged across all labels the Bi-LSTM (0.76) (Figure 6) out-performed all other classifiers, had the lowest Hamming and log loss (0.12, 0.17), and the highest average ROC (93.5).

TABLE 3. Performance of models using five-fold cross validation

Metric	SVM	XGBoost	KNN	RF	Bi-LSTM
Hamming loss	7.18	4.13	9.51	8.79	0.12
Accuracy	70.92	81.38	64.46	62.62	93.3
Log loss	2.71	3.98	4.76	5.16	0.17
Average precision-recall	0.58	0.69	0.31	0.34	0.76
Average ROC	90.5	88.2	65.6	65.8	93.9
Micro-average precision	0.64	0.85	0.70	0.81	N/A
Micro-average recall	0.88	0.78	0.33	0.33	N/A
Micro-average F1	0.74	0.82	0.45	0.46	N/A

*Averages are across all labels

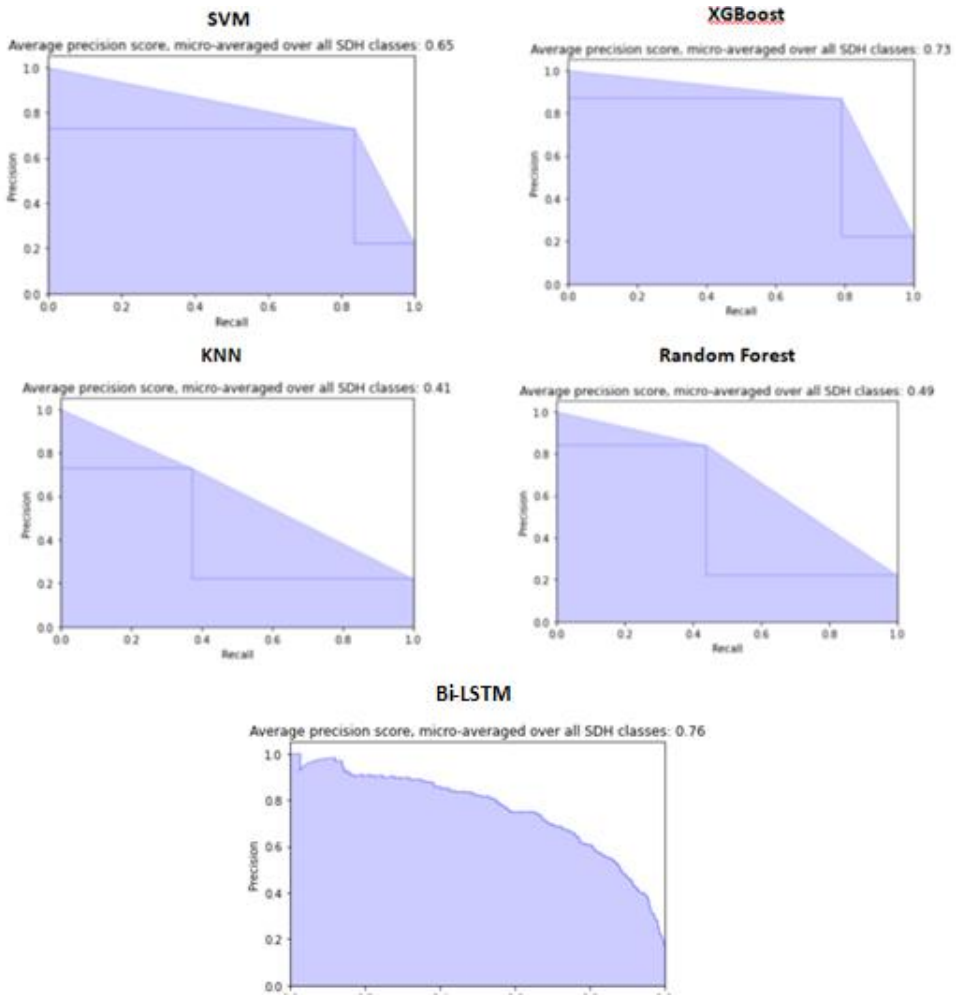
The XGBoost poor social support model was the best performing SDH classifier using (F1=0.89; Table 4), while the RF insurance insecurity model was the lowest performing model (F1=0.17). The best performing algorithms and their respective hyperparameters are presented in Appendix 1.

TABLE 4. Performance of models inferring SDH labels using five-fold cross validation

Algorithm	Label	Precision	Recall	F1	Support
SVM	General financial insecurity	0.51	0.91	0.66	69
	Employment/income insecurity	0.66	0.81	0.73	98
	Housing insecurity	0.62	0.89	0.73	83
	Poor social support	0.7	0.97	0.82	89
	Insurance insecurity	0.62	0.84	0.71	73
	Food insecurity	0.92	0.82	0.87	44
XGBboost	General financial insecurity	0.84	0.71	0.77	69
	Employment/income insecurity	0.85	0.76	0.8	98
	Housing insecurity	0.88	0.69	0.77	83
	Poor social support	0.85	0.93	0.89	89
	Insurance insecurity	0.81	0.75	0.78	73
	Food insecurity	0.93	0.86	0.89	44
KNN	General financial insecurity	0.66	0.33	0.44	69
	Employment/income insecurity	0.76	0.3	0.43	98
	Housing insecurity	0.53	0.12	0.2	83
	Poor social support	0.68	0.46	0.55	89
	Insurance insecurity	0.56	0.3	0.39	73

RF	Food insecurity	1	0.59	0.74	44
	General financial insecurity	0.78	0.42	0.55	69
	Employment/income insecurity	0.78	0.29	0.42	98
	Housing insecurity	0.83	0.23	0.36	83
	Poor social support	0.8	0.57	0.67	89
	Insurance insecurity	0.7	0.1	0.17	73
	Food insecurity	1	0.34	0.51	44

FIGURE 6. Average precision-recall score, micro-averaged over all SDH classes

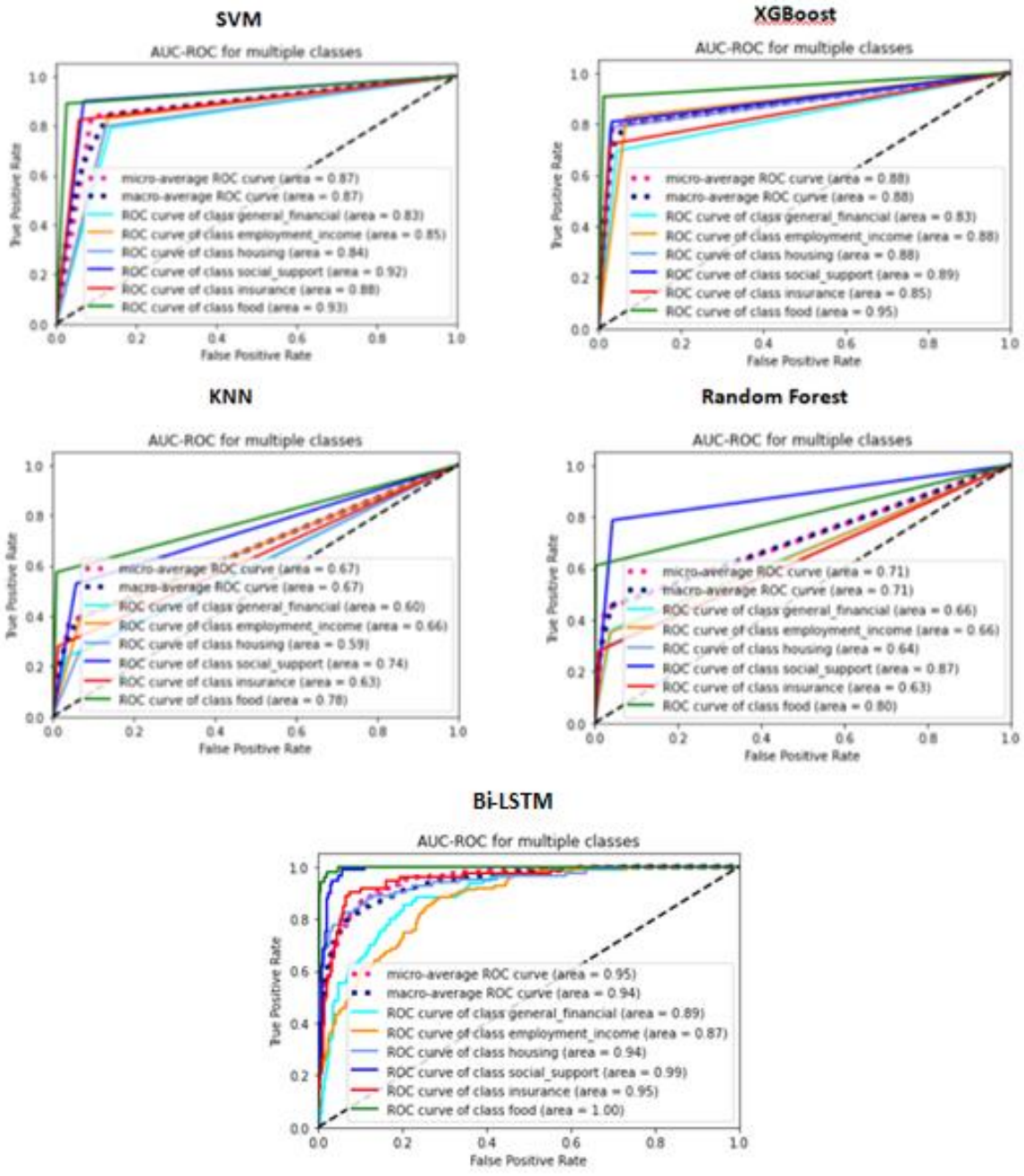


Because the prediction task relies on imbalanced data, we report Area under ROC curve (AUC). We observed that the Bi-LSTM outperformed all other models (Figure 7) with exception of the XGBoost classifying employment and/or income insecurity (Table 5). Figure 7 shows the ROC-curve of the best model on each outcome. We achieve relatively high AUC's with the SVM, XGBoost, and Bi-LSTM

TABLE 5. Model performance (AUC) by SDH label

SDH Label	SVM	XGBoost	KNN	Random Forest	Bi-LSTM
General financial insecurity	0.83	0.83	0.6	0.66	0.89
Employment or income insecurity	0.85	0.88	0.66	0.66	0.87
Housing insecurity	0.84	0.88	0.59	0.64	0.94
Poor social support	0.92	0.89	0.74	0.87	0.99
Insurance insecurity	0.88	0.85	0.63	0.63	0.95
Food insecurity	0.93	0.95	0.78	0.8	1

FIGURE 7. AUC-ROC for multiple classes



Error analysis

SVM. Among the 118 incorrect sentences classified by the SVM-baseline, 52 were false negatives and 20 were attributed to confusion between general financial insecurity and employment/income insecurity. Many false negatives (9) were within the insurance insecurity label due "Medicaid" being the only included text feature. However, many abbreviations for Medicaid were found in false negative insurance insecurity sentences. For example, "mcd" and "mcaid" were common abbreviations found in sentences. SVM had the highest false positives of any model with 228 due to a tendancy to label SDH classes' employment or income insecurity, general financial inseucirty, and housing insecurity as positive when one of those classes was labeled an actual true positive.

XGBoost. The lowest incorrect classified sentences was seen by the XGBoost model (88). While the XGBoost model had 120 false negatives it had the lowest number of false negatives of any model within the food insecurity (5) class. Many false negatvies in the housing insecurity (23) class were lengthy sentences that listed ICD-10 code diagnosis (e.g. "homeless"). Similar to the SVM model, several false negatives (28) were due to abbreviations of "Medicaid" in insurance insecurity, in addition to at times not recognizing the term "Medicaid" itself when bound within no sentence structure such as an answer to auto-generated questions. A total of 62 false positives were classified by this model with the lowest number of all models in the poor social support (7) class.

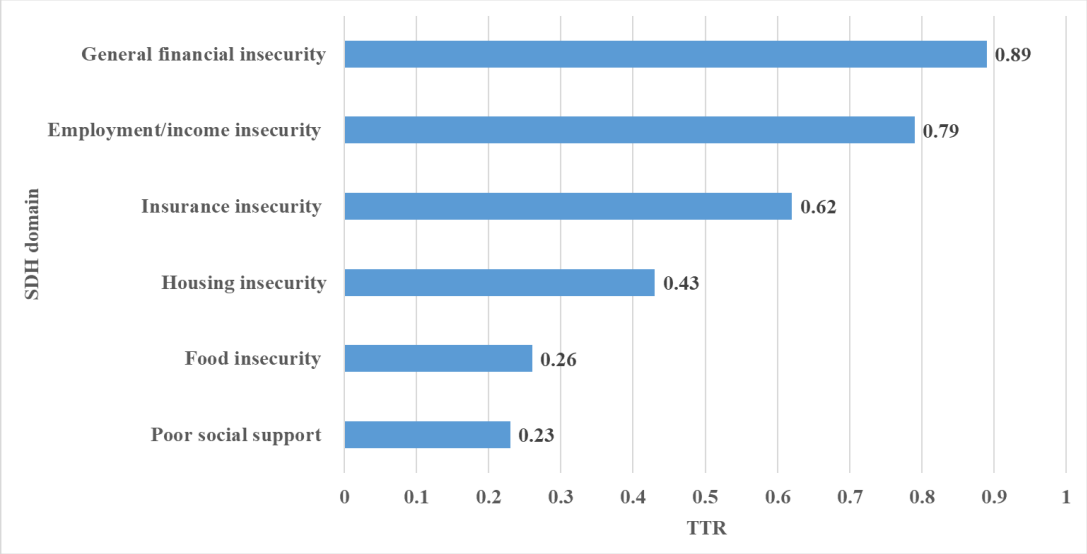
KNN. This model had low performance with 443 incorrect sentences classified including 258 false negatives and 79 false positives. KNN model had the highest number of false negatives of any model in the SDH classes of general financial insecurity (46) and poor social support (42).

Random Forest. This model had the highest number of false negatives (290) among the 398 incorrect sentences classified. Of all the models the Random Forest had the highest false negatives in the employment/income insecurity (79), housing insecurity (60), insurance insecurity (51), and food

insecurity (27). The lowest false positives in any model were classified by the Random Forest model in the insurance insecurity (5), housing insecurity (7), and food insecurity (0). The Random Forest had the lowest false positives of any model (52).

Although previous studies of SDH free text qualitative note lexical diversity in clinical documentation, few extend this analysis to quantitative measures. One of the most commonly used approaches to measure lexical diversity is to use the ratio of unique lexical items divided by the total number of words in a sample known as type-token ratio or TTR[51]. This is computed after standardizing the length of the text. Fergadiotis and colleagues used TTR to gauge lexical diversity in patients presenting with aphasia and found that shorter documents appeared to be richer when rendering comparisons across documents with different lengths[52]. A document or sentence length of 62 words or tokens was chosen based on the average length of a sentence after stop words were removed. Although not an aim of this study, we examined lexical diversity using TTR. Figure 8 shows the TTR by SDH class for the gold-standard corpus. The highest lexical diversity was seen in general financial and employment or income insecurity classes with the least among food insecurity and poor social support.

FIGURE 8. Lexical diversity of SDH classes



Discussion

This study aimed to develop a multi-label learning (MLL) model for identifying the SDH characteristics financial resource strain and poor social support using only clinical notes. Our findings suggest that a MLL approach trained on an SDH rich corpus can produce a high performing model. We also provide evidence that model performance is associated with lexical diversity by health professionals and the auto-generation of clinical note sentences to document SDH.

Based on our results, we recommend the neural network model, Bi-LSTM, because it performed well across all evaluation metrics. However, if the classification task requires transparency, gradient decision tree algorithms such as XGBoost, performed well across traditional evaluation metrics precision, recall, and micro-averaged F1 across all SDH labels. Our model outperformed (F1; 0.89-0.43) a similar study by Feller and colleagues[12] who used a multi-class gradient boosting tree to classify SDH sexual risk factors with F1 ranging from 79.2 for LGBT status to 27.3 for intravenous drug abuse. This outperformance is most likely due to our significantly larger training and testing dataset. Additionally, our model outperformed a similar MLL algorithm classification task by Zufferey and colleagues[21] whose top-performing algorithm (SVM) had a Hamming loss of 16.94 and AP of 0.72, as compared to our model (0.12, 0.76). The disparity in results may be attributed to our use of only 6 labels as opposed to Zufferey and colleagues' 15 labels. On the other hand, the poor performance of our KNN algorithm suggests that in multi-label medical domains the correlation between features may be an import characteristic to take into consideration. The successful results of the binary relevance SVM (F1= 0.74) approach assumes independence among SDH characteristics and suggests the features are not as correlated as previously assumed. An unintended consequence of our approach using an SDH rich corpus for model development may have led precision to be a more reliable evaluation measure. It is difficult to give a complete explanation about these results; however, it may be that the feature extraction process cannot optimally model the correlation between the features in a manner that an MLL approach can exploit. To confirm

this, we suggest further studies use algorithms and processing methods that deepen the analysis of SDH interdependence.

Our study is innovative in the following aspects. First, we developed a classifier to identify SDH on sentence-level data. The sentence-level scope can reduce the ambiguity of SDH characteristics and increase the agreement in the classification task as opposed to document level where SDH can still be buried within large quantities of text. Additionally, this allows for more granular results and thus a better understanding of the SDH documentation within clinical notes. Second, our model performance shows the feasibility of classifying SDH using only clinical notes and no structured features of the EHR. Feller and colleagues[12] found that the combination of clinical notes and structured data yielded better but not statistically significantly better performance than either data source alone when inferring SDH sexual risk factors. Future studies should explore whether structured data elements such as demographics and medical codes could enhance performance. Finally, we used an MLL approach to reduce information loss[20,53] and take advantage of the cardinality and label dependency[53,54].

We observed a positive correlation between model performance and the prevalence of each specific SDH. This demonstrates the necessity of building gold-standard corpora of adequate size, especially for infrequently documented SDH such as food insecurity. However, the poor social support label had higher precision and recall than other labels with similar prevalence, likely reflecting the limited lexical diversity used to express this SDH. For example, poor social support was often referenced as "limited social supports" or "lacking social support." Higher lexical diversity was observed in the general financial and employment or income insecurity labels. This increase in lexical diversity may have contributed to the lower performance of these individual labels in their respective models. The high performance of the food insecurity and poor social support labels may be related to the low lexical diversity attributed to these labels. This study contributes to our understanding of how SDH is

documented in the EHR and supports the qualitative analysis by previous studies[12–14] suggesting lexical diversity exists in SDH documentation. An unintended consequence of low annotator agreement in the general financial (84.4%) and insurance insecurity (79.1%) labels may have contributed to their overall poor performance (F1). Little research has analyzed the correlation between lower inter-rater reliability and machine learning performance. Future research should explore the relationship of lexical diversity and lower inter-rater reliability on model performance. These finding further emphasizes the importance of standardization in documentation and collection of SDH factors within the EHR.

The results of our error analysis suggest several areas for improvement in automated SDH classification. The labels general financial insecurity and employment/income insecurity had the highest false negatives (N=13). We believe this is due to the high lexical diversity associated with these labels, suggesting that clinicians lack a standardized way of expressing those SDH. Potentially, further sub-domains describing these SDH could be used to enhance future classification tasks. Our findings also suggest benefit from standardized approaches to collecting SDH data in EHRs [27,55,56].

Limitations

First, our SDH classifier was trained using data from a single institution limiting its generalizability, although our data comprise input from geographically distributed, rural and urban, academic and non-academic EDs. Future work should focus on using corpus developed from multiple institutions or publicly available sources. Second, our overall modest results may have resulted from data quality issues in the documentation of SDH and/or inaccurate annotation. Third, most approach this problem as a named-entity recognition (NER) task but because we approached the problem as a sentence labeling task, our experimental design does not allow for direct comparisons to previous work. Future work may explore the proficiency of SDH identification as an NER task. Fourth, our model performance may have been improved by considering negation or by correcting misspelling in text; we did not

consider negation due to the fact that not all SDH studied would have benefitted from this addition (i.e. "the patient denied homelessness despite living in a tent in the woods"). Fifth, our patient population was comprised of those who with MHSUD frequented the ED creating an overly positive dataset of SDH; these records likely differ the general population of a health system, potentially compromising the generalizability of the classification models. Additionally, this corpus does not represent prevalence of SDH and analyzing prevalence of SDH was not an aim of this study. Sixth, we did not use a "holdout" dataset that was never used in model training since we did not have the requisite volume of data to create training, validation, and test sets and thus the observed model performance may be inflated. If possible, future studies should use a holdout set to estimate unbiased model performance. Seventh, we did not explore other problem transformation approaches to MLL such as classifier chains or label powerset. Eighth, to balance our overly positive dataset we added negative sentences that were reviewed by one annotator leading to the possibility of hidden SDH among the negative. Ninth, there is limited ability to understanding neural networks as they use a hidden layer for pattern recognition in feature selection and thus full explanation of the Bi-LSTM results are not possible at this time. For full transparency into feature selection and decision tree decisions future work could explore transforming this problem into a multi-class classification task despite the information loss risks.

Conclusions

We investigated five common ML models for the task of multi-label classification of SDH using only clinical notes. Unlike previous work, we evaluated our models sentence-level data that contained multiple instances of SDH documentation, labels thus making sure our models could be used for real-world SDH clinical decision support tasks. Our MLL approach is a first concrete step towards SDH phenotyping across EHRs as SDH characteristics cross multiple domains and future studies may approach SDH phenotyping as an extreme MLL task. The study findings suggest that SDH prevalence and the

lexical diversity used to express a given SDH characteristic have an impact on the performance of classification algorithms. Future studies should explore the standardization of SDH collection and computational methods that can effectively learn models of diverse rare features.

REFERENCES

1. Smith JL, De Nadai AS, Storch EA, Langland-Orban B, Pracht E, Petrila J. Correlates of length of stay and boarding in florida emergency departments for patients with psychiatric diagnoses. *Psychiatr Serv*. 2016;67(11):1169-1174. doi:10.1176/appi.ps.201500283

2. Data on behavioral health in the United States. http://www.apa.org/helpcenter/data-behavioral-health.aspx. Accessed June 2, 2018.

3. Phelan JC, Link BG. Fundamental social causes of disease and mortality. In: *Encyclopedia of health and behavior*. 2455 Teller Road, Thousand Oaks California 91320 United States : SAGE Publications, Inc.; 2004. doi:10.4135/9781412952576.n102

4. Matthews KA, Adler NE, Forrest CB, Stead WW. Collecting psychosocial "vital signs" in electronic health records: Why now? What are they? What's new for psychology? *Am Psychol*. 2016;71(6):497-504. doi:10.1037/a0040317

5. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing social and behavioral domains in electronic health records: phase 1*. Washington (DC): National Academies Press (US); 2014. doi:10.17226/18709

6. Chang G, Weiss AP, Orav EJ, Rauch SL. Predictors of frequent emergency department use among patients with psychiatric illness. *Gen Hosp Psychiatry*. 2014;36(6):716-720. doi:10.1016/j.genhosppsych.2014.09.010

7. Ku BS, Fields JM, Santana A, Wasserman D, Borman L, Scott KC. The urban homeless: super-users of the emergency department. *Popul Health Manag*. 2014;17(6):366-371. doi:10.1089/pop.2013.0118

8. Foege WH. Actual Causes of Death in the United States-Reply. *JAMA*. 1994;271(9):660. doi:10.1001/jama.1994.03510330037023

9. Singh GK, Siahpush M, Kogan MD. Neighborhood socioeconomic conditions, built environments, and childhood obesity. *Health Aff (Millwood)*. 2010;29(3):503-512. doi:10.1377/hlthaff.2009.0730

10. Gold R, Cottrell E, Bunce A, et al. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *J Am Board Fam Med*. 2017;30(4):428-447. doi:10.3122/jabfm.2017.04.170046

11. Hatef E, Rouhizadeh M, Tia I, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system. *JMIR Med Inform*. 2019;7(3):e13802. doi:10.2196/13802

12. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data. *Appl Clin Inform*. 2020;11(1):172-181. doi:10.1055/s-0040-1702214

13. Bejan CA, Angiolillo J, Conway D, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc*. 2018;25(1):61-71. doi:10.1093/jamia/ocx059

14. Vest JR, Grannis SJ, Haut DP, Halverson PK, Menachemi N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int J Med Inform*. 2017;107:101-106. doi:10.1016/j.ijmedinf.2017.09.008

15. 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.

16. Hollister BM, Restrepo NA, Farber-Eger E, Crawford DC, Aldrich MC, Non A. Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. *Pac Symp Biocomput*. 2017;22:230-241. doi:10.1142/9789813207813_0023

17. Lindemann EA, Chen ES, Rajamani S, Manohar N, Wang Y, Melton GB. Assessing the Representation of Occupation Information in Free-Text Clinical Documents Across Multiple Sources. *Stud Health Technol Inform*. 2017;245:486-490.

18. Blosnich JR, Marsiglio MC, Dichter ME, et al. Impact of Social Determinants of Health on Medical Conditions Among Transgender Veterans. *Am J Prev Med*. 2017;52(4):491-498. doi:10.1016/j.amepre.2016.12.019

19. Du J, Chen Q, Peng Y, Xiang Y, Tao C, Lu Z. ML-Net: multi-label classification of biomedical texts with deep neural networks. *J Am Med Inform Assoc*. 2019;26(11):1279-1285. doi:10.1093/jamia/ocz085

20. Zhang M-L, Zhou Z-H. A Review on Multi-Label Learning Algorithms. *IEEE Trans Knowl Data Eng*. 2014;26(8):1819-1837. doi:10.1109/TKDE.2013.39

21. Zufferey D, Hofer T, Hennebert J, Schumacher M, Ingold R, Bromuri S. Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Comput Biol Med*. 2015;65:34-43. doi:10.1016/j.compbiomed.2015.07.017

22. Hong N, Wen A, Stone DJ, et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J Biomed Infor*m. 2019;99:103310. doi:10.1016/j.jbi.2019.103310

23. Liang C, Gong Y. Automated Classification of Multi-Labeled Patient Safety Reports: A Shift from Quantity to Quality Measure. *Stud Health Technol Inform*. 2017;245:1070-1074.

24. Baumel T. Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment.

25. Krieg C, Hudon C, Chouinard M-C, Dufour I. Individual predictors of frequent emergency department use: a scoping review. *BMC Health Serv Res*. 2016;16(1):594. doi:10.1186/s12913-016-1852-1

26. Brennan JJ, Chan TC, Hsia RY, Wilson MP, Castillo EM. Emergency department utilization among frequent users with psychiatric visits. *Acad Emerg Med*. 2014;21(9):1015-1022. doi:10.1111/acem.12453

27. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing social and behavioral domains and measures in electronic health records: phase* 2. Washington (DC): National Academies Press (US); 2015. doi:10.17226/18951

28. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm Policy Ment Heal*th. 2015;42(5):533-544. doi:10.1007/s10488-013-0528-y

29. He H, Ma Y, eds. *Imbalanced learning: foundations, algorithms, and applications*. Wiley; 2013. doi:10.1002/9781118646106

30. Cohen R, Aviram I, Elhadad M, Elhadad N. Redundancy-aware topic modeling for patient record notes. *PLoS On*e. 2014;9(2):e87555. doi:10.1371/journal.pone.0087555

31. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*. 2013;14:10. doi:10.1186/1471-2105-14-10

32. Lingren T, Deleger L, Molnar K, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J Am Med Inform Assoc*. 2014;21(3):406-413. doi:10.1136/amiajnl-2013-001837

33. Natural Language Toolkit — NLTK 3.5 documentation. https://www.nltk.org/. Accessed May 12, 2020.

34. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learnin*g. Vol 103. New York, NY: Springer New York; 2013. doi:10.1007/978-1-4614-7138-7

35. Manning CD, Raghavan P, Schutze H. *Introduction to information retrieval*. Cambridge: Cambridge University Press; 2008. doi:10.1017/CBO9780511809071

36. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York, NY: Springer New York; 2013. doi:10.1007/978-1-4614-6849-3

37. Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recogni*t. 2012;45(9):3084-3104. doi:10.1016/j.patcog.2012.03.004

38. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Resear*ch. 2011.

39. Zhang R, Pakhomov SV, Lee JT, Melton GB. Using language models to identify relevant new information in inpatient clinical notes. *AMIA Annu Symp Pro*c. 2014;2014:1268-1276.

40. Zhang Y, Zhang OR, Li R, et al. Psychiatric stressor recognition from clinical notes to reveal association with suicide. *Health Informatics J*. October 2018:1460458218796598. doi:10.1177/1460458218796598

41. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD' ' '16*. New York, New York, USA: ACM Press; 2016:785-794. doi:10.1145/2939672.2939785

42. Talman A, Yli-Jyrä A, Tiedemann J. Sentence embeddings in NLI with iterative refinement encoders. *Nat Lang Eng*. 2019;25(4):467-482. doi:10.1017/S1351324919000202

43. Zhang X, Zhao H, Zhang S, Li R. A Novel Deep Neural Network Model for Multi-Label Chronic Disease Prediction. *Front Genet*. 2019;10:351. doi:10.3389/fgene.2019.00351

44. Tran T, Kavuluru R. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. *J Biomed Inform*. 2017;75S:S138-S148. doi:10.1016/j.jbi.2017.06.010

45. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to Diagnose with LSTM Recurrent Neural Networks. *arXiv*. November 2015.

46. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *J Biomed Inform*. 2017;76:102-109. doi:10.1016/j.jbi.2017.11.007

47. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014:1532-1543. doi:10.3115/v1/D14-1162

48. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised Learning of Universal Sentence Representations from

' ' Natural Language Inference Data. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural*

' ' *Language Processin*g. Stroudsburg, PA, USA: Association for Computational Linguistics; 2017:670-680. doi:10.18653/v1/D17-1070

49. Zhou P, Qi Z, Zheng S, Xu J, Bao H, Xu B. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. *arXiv*. November 2016.

50. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735

51. Hess CW, Ritchie KP, Landry RG. The Type-Token Ratio and Vocabulary Performance. *Psychol Rep*. 1984;55(1):51-57. doi:10.2466/pr0.1984.55.1.51

52. Fergadiotis G, Wright HH, West TM. Measuring lexical diversity in narrative discourse of people with aphasia. *Am J Speech Lang Pathol*. 2013;22(2):S397-408. doi:10.1044/1058-0360(2013/12-0083)

53. Luaces O, Díez J, Barranquero J, del Coz JJ, Bahamonde A. Binary relevance efficacy for multilabel classification. *Prog Artif Intel*l. 2012;1(4):303-313. doi:10.1007/s13748-012-0030-x

54. Bromuri S, Zufferey D, Hennebert J, Schumacher M. Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. *J Biomed Inform*. 2014;51:165-175. doi:10.1016/j.jbi.2014.05.010

55. Monsen KA, Rudenick JM, Kapinos N, Warmbold K, McMahon SK, Schorr EN. Documentation of social determinants in electronic health records with and without standardized terminologies: A comparative study. *Proceedings of Singapore Healthca*re. 2018;28(1):201010581878564. doi:10.1177/2010105818785641

56. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Me*d. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

57. Social Determinants of Health | Healthy People 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health. Accessed March 13, 2020.

58. Brooker A-S, Eakin JM. Gender, class, work-related stress and health: toward a power-centred approach. *J Community Appl Soc Psychol*. 2001;11(2):97-109. doi:10.1002/casp.620

59. Braveman PA, Egerter SA, Mockenhaupt RE. Broadening the focus: the need to address the social determinants of health. *Am J Prev Med*. 2011;40(1 Suppl 1):S4-18. doi:10.1016/j.amepre.2010.10.002

60. Heiman H, Artiga S. Beyond Health Care: The Role of Social Determinants in Promoting Health and Health Equity.

61. Alegría M, NeMoyer A, Falgàs Bagué I, Wang Y, Alvarez K. Social determinants of mental health: where we are and where we need to go. *Curr Psychiatry Rep*. 2018;20(11):95. doi:10.1007/s11920-018-0969-9

62. Simmons LA, Swanberg JE. Psychosocial work environment and depressive symptoms among US workers: comparing working poor and working non-poor. *Soc Psychiatry Psychiatr Epidemiol*. 2009;44(8):628-635. doi:10.1007/s00127-008-0479-x

63. Doran KM, Kunzler NM, Mijanovich T, et al. Homelessness and other social determinants of health among emergency department patients. *J Soc Distress Homeless*. 2016;25(2):71-77. doi:10.1080/10530789.2016.1237699

APPENDIX 1: SDH Annotation Guidelines

What are social determinants of health (SDH)?

The Centers for Disease Control defines SDH as the conditions in the places where people live, learn, work, and play that affect a wide range of health risks and outcomes.[54]

We know that poverty limits access to healthy foods and safe neighborhoods and that more education is a predictor of better health.[26,54] We also know that differences in health are striking in communities that experience pronounced SDH such as unstable housing, low income, unsafe neighborhoods, or substandard education.[54–56] By applying what we know to reverse SDH, we can not only improve individual and population health but also advance health equity.[57,58]

The purpose of this study is to identify and classify two SDH categories: financial resource strain and poor social support systems.

Financial resource strain encompasses both the subjective sense of strain as the result of economic difficulties and the specific sources of strain, including employment insecurity, income insecurity, housing insecurity, and food insecurity.[26] Financial resource strain is a characteristic of a household or family unit and not simply individual, and may take the form of housing insecurity, employment insecurity, and general characteristics of financial insecurity such as lack of insurance.

Social relationships have been identified as a major psychosocial risk factor for health, and they have been identified as potential resources or buffers mitigating the impact of other risk factors for health, such as stress, and facilitating recovery from acute and chronic diseases.[26] Add some vocabulary examples as you did for financial strain.

Objective

To classify individual sentences of a clinical note record as positive or negative for the SDH category of financial resource strain and poor social support.

SDH concept examples

1. Housing insecurity: can take many forms including multiple moves, foreclosure, and homelessness.[26]

Examples:

"she has also had multiple, different living situations including several assisted living facilities, all of which she left within 24 hours and, at times, has been homeless" – *positive for SDH*

"patient arranged for personal transportation to signature healthcare and to a hotel" –*negative for SDH*

2. Employment/Income insecurity: this includes joblessness with an expressed desire to work, but available to work, underemployed (poverty-wage employment and intermittent unemployment, intermittent employment), unemployment caused by incarceration, and dependent on subsidies.[26,59]

a. Intersection with criminal justice system

"he was diagnosed 2005 at age 27 during prison screening lab exam" – *positive for SDH*

"patient left ama citing need to go to a court date" – *positive for SDH*

"90% ostial d1 stenosis - in stent jail" –*negative for SDH*

"denies pets, tb in self or family, time in prison/jail" –*negative for SDH*

3. Food insecurity: food insecurity includes financial struggle in securing food, prioritizing or compromising other necessities such as housing standards to gain food security,.[26] reliance on food subsidies. The presence of government-funded food assistance programs does not necessarily diminish food insecurity due to many patients needing help navigating the administrative responsibilities or managing resources.[26,54,60]

"pt also has food stamps" – *positive for SDH*

"continues to endorse food insecurity at home" – *positive for SDH*

"food insecurity: patient reports no food insecurity" –*negative for SDH*

4. General financial resource strain: encompasses all other characteristics of financial instability including disability and lack of transportation.

"relationship status: seperated children: yes; one son (age 16); son lives with pt's sister due to social services involvement in the past after which pt lost custody of son education: some high school income/employment/disability: disability military service: no abuse/neglect/trauma: none " – *positive for SDH*

"rone has missed 2 appointments with id this week due to lack of transportation" – *positive for SDH*

"no epileptiform discharges were seen" –*negative for SDH*

5. Insecurity due to insurance: encompasses underinsured (high copays or deductibles, poor coverage), lack of insurance, and poor socioeconomic status based on insurance (Medicaid, Charity Care).

"she has medicare part a and b and medicaid as her insurances" – *positive for SDH*

"part b copayments may vary by type of service" –*negative for SDH*

6. Poor social support: quantity and quality of an individual social relationships can be described as social isolation or the perceived support or benefit a person derives from interpersonal relationships.[26]

"living situation: the patient lives alone" – *positive for SDH*

"per h&p limited social support, complex medical issues, multiple suicide attempts, single mother, family discord, medication refractory, history of trauma and chronic pain goals: 1" – *positive for SDH*

Multiple SDH

Some SDH characteristics cross different categories. For example, a high insurance copay may strain financial or making a choice to pay for food over medications. Under the current health care insurance system in the United States, where insurance is tightly tied to employment, employment has obvious consequences for health care insurance for an individual's immediate family members and for their ability to access prescription drugs.[26]

Procedure

1. Open the Excel file “MH_annotation.csv”

2. Resize your formula bar to your preference.

3. Freeze the top row of the excel sheet: Click View → Freeze panes → Freeze top row

4. Read the entire text in Column C, “RPT_TEXT”.

5. Determine if the text is indicative of the patient having a characteristic of SDH.

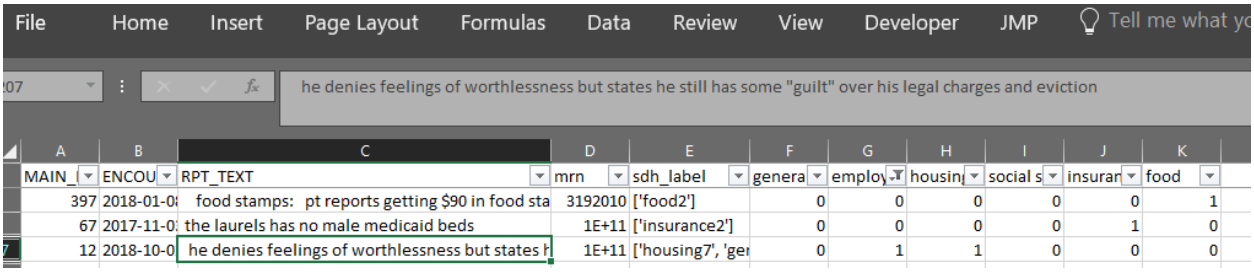
6. Place a 1 or 0 in Column F-K, “label”. 1 for SDH positive, 0 for SDH negative (Figure 1).

- Only multi-label for sentences with multiple SDH characteristics **NOT** multiple meanings

“carrboro, nc 27510 phone: 919-942-8741 wake county human service dental clinic routine dental services for children and pregnant women of wake county medicaid and nc health choice; sliding scale based on household income” – *positive for insurance*, *negative for financial*

“he denies feelings of worthlessness but states he still has some "guilt" over his legal charges and eviction” – *positive for employment/income and housing insecurity*

Figure 1. Example of annotation process



	A	B	C	D	E	F	G	H	I	J	K
	MAIN_	ENCOU	RPT_TEXT	mrn	sdh_label	genera	employ	housin	social s	insuran	food
	397	2018-01-0	food stamps: pt reports getting $90 in food sta	3192010	['food2']	0	0	0	0	0	1
	67	2017-11-0	the laurels has no male medicaid beds	1E+11	['insurance2']	0	0	0	0	1	0
	12	2018-10-0	he denies feelings of worthlessness but states h	1E+11	['housing7', 'ger	0	1	1	0	0	0

6. Sentences that are not clear or you feel require more explanation for your label of 1 or 0 please put a short (2-3 words) comment in Column L, “comment”.

7. Save frequently and keep track of your time.

Challenges and Barriers

1. Auto-generated sentences: Some sentences appear to be auto-generated by the EHR, such as the content found in referrals. These sentences should be classified based on the indication of an SDH not the patient’s underlying mental health or substance use disorder conditions.

"you will not receive extra refills for lost or stolen medications, regardless of the circumstances” – *negative for SDH*

“ check with medicaid or your plan” –*negative for SDH*

2. Potentially subjective sentences: Use your expertise and err on the side of inclusion rather than exclusion.

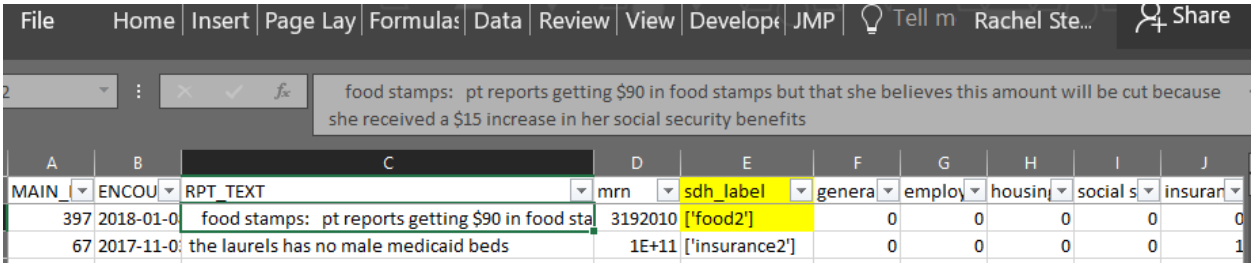
"pt stated his possessions were "stolen" " – *positive for SDH*

3. Lengthy notes

"problem: patient care overview (adult) goal: plan of care review outcome: progressing 04/04/16 2215 plan of care review plan of care reviewed with patient progress improving goal: individualization and mutuality outcome: progressing 04/04/16 2215 individualization patient specific goals "to find a place to live" goal: discharge needs assessment outcome: progressing 04/04/16 2215 discharge needs assessment concerns to be addressed denies needs/concerns at this time;discharge planning concerns;financial/insurance concerns;home safety concerns;mental health concerns readmission within the last 30 days lack of support;previous discharge plan unsuccessful outpatient/agency/support group needs assisted living facility (specify) anticipated changes related to illness inability to care for self equipment currently used at home wheelchair;walker, rolling discharge facility/level of care needs assisted living facility current discharge risk substance abuse;psychiatric illness;lack of support system/caregiver problem: pressure ulcer risk (braden scale) (adult,obstetrics,pediatric) goal: identify related risk factors and signs and symptoms related risk factors and signs and symptoms are identified upon initiation of human response clinical practice guideline (cpg) outcome: progressing 04/04/16 2215 pressure ulcer risk (braden scale) (adult,obstetrics,pediatric) related risk factors (pressure ulcer risk (braden scale)) body weight extremes;mobility impaired goal: skin integrity patient will demonstrate the desired outcomes by discharge/transition of care" – *positive for SDH; poor social support and housing insecurity*

- Use the sdh_label from the dataset to help you find the key term that the algorithm used to identify the sentence. See Figure 1 below.

Figure 2. Example of dataset



	A	B	C	D	E	F	G	H	I	J
	MAIN_	ENCOU	RPT_TEXT	mrn	sdh_label	genera	employ	housin	social s	insuran
397		2018-01-0	food stamps: pt reports getting $90 in food sta	3192010	['food2']	0	0	0	0	0
67		2017-11-0	the laurels has no male medicaid beds	1E+11	['insurance2']	0	0	0	0	1

APPENDIX 2: Feature names

N=2000

['abdomen', 'abdomin', 'abil', 'abl', 'abl afford', 'abnorm', 'abov', 'abov inform', 'absenc', 'absenc caregiv', 'absenc caregiv assist', 'absenc caregiv assist discharg', 'absent', 'absent manag', 'abus', 'abus histori', 'abus loz', 'abus treatment', 'abuseneglecttrauma', 'abuseus', 'accept', 'access', 'access commun', 'access commun servic', 'access firearm', 'act', 'action', 'activ', 'activ file', 'activ file topic', 'activ file topic concern', 'activ particip', 'activ particip think', 'activ particip think process', 'activ problem', 'activ problem list', 'activ

problem list diagnosi', 'acut', 'addit', 'address', 'address citi', 'address citi counti', 'address citi counti state', 'adher', 'adher safeti', 'adher safeti consider', 'adher safeti consider self', 'adjust', 'adl', 'administ', 'admiss', 'admiss day', 'admiss day anticip', 'admiss day anticip chang', 'admiss equip', 'admiss equip current', 'admiss equip current use', 'admiss inpati', 'admiss inpati psychiatr', 'admiss inpati psychiatr unit', 'admit', 'admit date', 'admit inpati', 'admit inpati psychiatr', 'admit inpati psychiatr unit', 'adult', 'adult goal', 'adult intervent', 'adultobstetricspediatr', 'advanc', 'affect', 'afford', 'afford medic', 'age', 'age onset', 'age onset loz', 'agenc', 'agenc namephon', 'ago', 'alcohol', 'alcohol abus', 'alcohol use', 'alcohol use disord', 'alert', 'allergi', 'allianc', 'allianc support', 'allianc support success', 'allianc support success transit', 'alon', 'alon absenc', 'alon absenc caregiv', 'alon absenc caregiv assist', 'ambul', 'amput', 'andor', 'andor famili', 'andor famili provid', 'andor famili provid choic', 'anemia', 'ani', 'ani necessari', 'ani necessari medic', 'ani necessari medic clearanc', 'anticip', 'anticip chang', 'anticip chang relat', 'anticip chang relat ill', 'anticip comment', 'anticip servic', 'anticip servic transit', 'anticip transit', 'anxieti', 'anxieti loz', 'apart', 'appear', 'appli', 'applic', 'appoint', 'appropri', 'appropri meet', 'appropri meet post', 'appropri meet post hospit', 'area', 'arrang', 'arteri', 'ask', 'assault', 'assess', 'assess concern', 'assess concern address', 'assess outcom', 'assess outcom progress', 'assess outcom progress discharg', 'assess patient', 'assess patient person', 'assess patient person interview', 'assess scale', 'assess scale estim', 'assess scale whoda', 'assist', 'assist discharg', 'assist discharg home', 'assist discharg home care', 'assist discharg need', 'assist discharg need assess', 'assist requir', 'assist ye', 'assist ye type', 'assist ye type financi', 'associ', 'asthma', 'attempt', 'attend', 'attest', 'attest review', 'attest review abov', 'attest review abov inform', 'autorefreshmetadatabegin', 'autorefreshmetadatabegin autorefreshmetadataend', 'autorefreshmetadataend', 'avail', 'avail appropri', 'avail appropri meet', 'avail appropri meet post', 'awar', 'balanc', 'barrier', 'barrier medic', 'barrier medic prior', 'barrier medic prior overnight', 'barrier medic ye', 'barrier medic ye comment', 'base', 'basic', 'basic need', 'bear', 'becaus', 'bed', 'bedsid', 'befor', 'behavior', 'belief', 'benefit', 'benzodiazepin', 'bid', 'bilater', 'biopsi', 'bipolar', 'bipolar disord', 'bleed', 'blood', 'bodi', 'bodi structur', 'bone', 'borderlin', 'bowel', 'bp', 'breath', 'brother', 'calm', 'cancer', 'cannabi', 'capsul', 'car', 'care', 'care home', 'care home na', 'care manag', 'care manag assess', 'care manag assess patient', 'care manag initi', 'care manag initi transit', 'care need', 'care overview', 'care overview goal', 'care review', 'care review outcom', 'care review outcom progress', 'care review patient', 'care self', 'care servic', 'care servic place', 'care servic place prior', 'caregiv', 'caregiv assist', 'caregiv assist discharg', 'caregiv assist discharg home', 'caregiv respons', 'caregiv respons complet', 'caregiv respons complet homemak', 'carolina', 'carolina guardianpaye', 'case', 'cell', 'center', 'central', 'chang', 'chang relat', 'chang relat ill', 'chang relat ill equip', 'chang relat ill inabl', 'chapel', 'chapel hill', 'chapel hill nc', 'charg', 'chariti', 'chariti care', 'check', 'chest', 'chest pain', 'chief', 'chief complaint', 'child', 'children', 'children educ', 'choic', 'choic facil', 'choic facil servic', 'choic facil servic avail', 'chronic', 'chronic mental', 'chronic mental ill', 'chronic obstruct', 'chronic obstruct pulmonari', 'chronic pain', 'cigarett', 'citi', 'citi counti', 'citi counti state', 'clear', 'clearanc', 'clearanc patient', 'clearanc patient admit', 'clearanc patient admit inpati', 'client', 'client factor', 'client factor spiritu', 'client factor spiritu belief', 'clinic', 'clinic practic', 'clinic practic guidelin', 'clinic practic guidelin cpg', 'clinic risk', 'clinic risk factor', 'clinic risk factor multipli', 'close', 'club', 'cm', 'cmshcc', 'cmshcc loz', 'cocain', 'cocain use', 'cocain use disord', 'cognit', 'collect', 'colleg', 'come', 'comment', 'commun', 'commun agenc', 'commun mobil', 'commun resourc', 'commun servic', 'complaint', 'complet', 'complet homemak', 'complianc', 'complic', 'compon', 'concern', 'concern address', 'concern clinic risk factor', 'concern loz', 'concern loz file', 'concern loz file social', 'concern recommend', 'concern recommend follow', 'concern recommend follow ani', 'condit', 'conf', 'confirm', 'conflict', 'consid', 'consider', 'consider self', 'consult', 'contact', 'continu', 'continu admiss', 'continu admiss inpati', 'continu admiss inpati psychiatr', 'control', 'cooper', 'coordinationcar', 'copay', 'copd', 'cope', 'cope skill', 'cope skill respons', 'cope skill respons life', 'cope strategi', 'coronari', 'cost', 'counti', 'counti health', 'counti health depart', 'counti state', 'court', 'court date', 'cover', 'coverag', 'cpg', 'crisi', 'ct', 'cultur', 'current', 'current discharg', 'current discharg risk', 'current everi', 'current everi day', 'current everi day smoker', 'current homeless', 'current live', 'current receiv', 'current receiv outpati', 'current receiv outpati dialysi',

'current use', 'current use home', 'current use home current', 'current use home equip', 'currentprior', 'currentprior legal', 'cut', 'daili', 'damag', 'date', 'date loz', 'daughter', 'day', 'day anticip', 'day anticip chang', 'day anticip chang relat', 'day previou', 'day previou admiss', 'day previou admiss day', 'day readmiss', 'day readmiss day', 'day readmiss day previou', 'day smoker', 'day ye', 'dc', 'debrid', 'decis', 'decis maker', 'decreas', 'deficit', 'deg', 'delus', 'delus note', 'demonstr', 'demonstr desir', 'demonstr desir outcom', 'demonstr desir outcom dischargetransit', 'deni', 'dental', 'depart', 'depend', 'depend rafhcc', 'depress', 'depress disord', 'depress loz', 'describ', 'desir', 'desir outcom', 'desir outcom dischargetransit', 'desir outcom dischargetransit care', 'detox', 'develop', 'developsparticip', 'developsparticip therapeut', 'developsparticip therapeut allianc', 'developsparticip therapeut allianc support', 'devic', 'diabet', 'diabet mellitu', 'diagnos', 'diagnos activ', 'diagnos activ problem', 'diagnos chronic', 'diagnos patient', 'diagnos patient activ', 'diagnos patient activ problem', 'diagnos princip', 'diagnos princip problem', 'diagnosi', 'diagnosi date', 'diagnosi date loz', 'diagnosi loz', 'dialysi', 'dialysi financi', 'dialysi financi inform', 'dialysi financi inform patient', 'diet', 'differ', 'difficulti', 'direct', 'disabl', 'disabl assess', 'disabl assess scale', 'disabl assess scale estim', 'disabl assess scale whoda', 'discharg', 'discharg date', 'discharg discharg', 'discharg discharg facilitylevel', 'discharg discharg facilitylevel care', 'discharg disposit', 'discharg disposit patient', 'discharg facilitylevel', 'discharg facilitylevel care', 'discharg facilitylevel care need', 'discharg home', 'discharg home care', 'discharg need', 'discharg need assess', 'discharg need assess concern', 'discharg need assess outcom', 'discharg plan', 'discharg plan need', 'discharg plan need identifi', 'discharg plan screen', 'discharg plan screen discharg', 'discharg plan unsuccess', 'discharg risk', 'dischargetransit', 'dischargetransit care', 'discord', 'discuss', 'diseas', 'diseas loz', 'diseas rafhcc', 'disord', 'disord loz', 'disord rafhcc', 'disord rafhcc loz', 'disord sever', 'disord stressor', 'disp', 'dispens', 'disposit', 'disposit patient', 'divorc', 'doe', 'doesnt', 'domest', 'domest violenc', 'domin', 'dose', 'dr', 'drink', 'drive', 'drive commun', 'drive commun mobil', 'drug', 'drug use', 'drug use loz', 'drug use loz sexual', 'durat', 'dure', 'durham', 'dvt', 'dvt pe', 'dx', 'eat', 'ed', 'ed visit', 'ed visit day', 'ed visit day readmiss', 'edema', 'educ', 'educ high', 'educ high school', 'educ level', 'educ na', 'effect', 'elev', 'emerg', 'emot', 'emot regul', 'emot regul skill', 'employ', 'encount', 'encourag', 'endo', 'endors', 'endoscopydiagnosi', 'engag', 'environ', 'environ homeless', 'episod', 'equip', 'equip current', 'equip current use', 'equip current use home', 'equip need', 'equip need discharg', 'equip need discharg discharg', 'essenti', 'estim', 'estim discharg', 'etoh', 'evalu', 'event', 'everi', 'everi day', 'everi day smoker', 'evict', 'evid', 'exam', 'excess', 'exercis', 'expect', 'exposur', 'express', 'extrem', 'eye', 'eye contact', 'eye contact good', 'eye contact good hallucin', 'facil', 'facil servic', 'facil servic avail', 'facil servic avail appropri', 'facilitylevel', 'facilitylevel care', 'facilitylevel care need', 'factor', 'factor multipli', 'factor multipli diagnos', 'factor multipli diagnos chronic', 'factor sign', 'factor sign symptom', 'factor sign symptom identifi', 'factor sign symptom relat', 'factor spiritu', 'factor spiritu belief', 'failur', 'fair', 'fall', 'fall risk', 'famili', 'famili contact', 'famili histori', 'famili histori famili', 'famili histori famili histori', 'famili histori problem', 'famili histori problem relat', 'famili member', 'famili provid', 'famili provid choic', 'famili provid choic facil', 'father', 'fee', 'fee base', 'feel', 'femal', 'file', 'file social', 'file social histori', 'file social histori narr', 'file topic', 'file topic concern', 'final', 'financ', 'financi', 'financi assist', 'financi assist discharg', 'financi assist discharg need', 'financi assist requir', 'financi assist ye', 'financi assist ye type', 'financi difficulti', 'financi inform', 'financi inform patient', 'financi inform patient sourc', 'financi resourc', 'financi resourc strain', 'financi strain', 'firearm', 'flowsheet', 'fluid', 'follow', 'follow ani', 'follow ani necessari', 'follow ani necessari medic', 'followup', 'food', 'food insecur', 'food insecur worri', 'food pantri', 'food stamp', 'foot', 'fractur', 'frame', 'frame week', 'free', 'frequenc', 'friend', 'function', 'fund', 'gastroenterolog', 'gastroenterolog loz', 'gastroenterolog loz pr', 'gastroenterolog loz pr upper', 'gastroesophag', 'gastroesophag reflux', 'gastroesophag reflux diseas', 'gener', 'gerd', 'gi', 'gi endoscopydiagnosi', 'goal', 'goal adult', 'goal discharg', 'goal discharg need', 'goal discharg need assess', 'goal identifi', 'goal identifi relat', 'goal identifi relat risk', 'goal individu', 'goal individu mutual', 'goal individu mutual outcom', 'goal plan', 'goal plan care', 'goal plan care review', 'good', 'good hallucin', 'grade', 'grammarmetadatabegin', 'grammarmetadatabegin grammarmetadataend', 'grammarmetadatabegin

grammarmetadataend autorefreshmetadatabegin', 'grammarmetadatabegin grammarmetadataend autorefreshmetadatabegin autorefreshmetadataend', 'grammarmetadataend', 'grammarmetadataend autorefreshmetadatabegin', 'grammarmetadataend autorefreshmetadatabegin autorefreshmetadataend', 'grandfath', 'grandmoth', 'grossli', 'group', 'group need', 'guardianpaye', 'guidelin', 'guidelin cpg', 'ha', 'habit', 'habit domin', 'habit impoverish', 'habit use', 'hallucin', 'health', 'health care', 'health depart', 'health servic', 'healthcar', 'heart', 'heart failur', 'help', 'hepat', 'hi', 'hi mother', 'high', 'high school', 'highest', 'highest educ', 'highest educ level', 'hill', 'hill nc', 'hip', 'histor', 'histori', 'histori diagnosi', 'histori diagnosi date', 'histori diagnosi date loz', 'histori famili', 'histori famili histori', 'histori famili histori problem', 'histori file', 'histori loz', 'histori loz marit', 'histori loz marit statu', 'histori main', 'histori main topic', 'histori main topic loz', 'histori marit', 'histori marit statu', 'histori narr', 'histori narr live', 'histori past', 'histori past medic', 'histori past medic histori', 'histori past surgic', 'histori past surgic histori', 'histori problem', 'histori problem relat', 'histori problem relat age', 'histori procedur', 'histori procedur later', 'histori procedur later date', 'histori social', 'histori social histori', 'histori social histori loz', 'histori socioeconom', 'histori socioeconom histori', 'histori socioeconom histori marit', 'histori substanc', 'histori substanc use', 'ho', 'home', 'home care', 'home care servic', 'home care servic place', 'home current', 'home current receiv', 'home current receiv outpati', 'home equip', 'home equip need', 'home equip need discharg', 'home home', 'home home care', 'home home care servic', 'home live', 'home na', 'homeless', 'homemak', 'hospit', 'hospit care', 'hospit care need', 'hospit stay', 'hospit stay ed', 'hospit stay ed visit', 'hotel', 'hour', 'hous', 'howev', 'hpi', 'hprh', 'hprh servic', 'hr', 'human', 'human respons', 'human respons clinic', 'human respons clinic practic', 'husband', 'hx', 'hx loz', 'hypertens', 'hypertens loz', 'ideat', 'identifi', 'identifi anticip', 'identifi anticip comment', 'identifi initi', 'identifi initi human', 'identifi initi human respons', 'identifi relat', 'identifi relat risk', 'identifi relat risk factor', 'ii', 'ill', 'ill equip', 'ill equip current', 'ill equip current use', 'ill equip need', 'ill equip need discharg', 'ill inabl', 'ill inabl care', 'immun', 'impair', 'impoverish', 'improv', 'impuls', 'inabl', 'inabl care', 'inadequ', 'includ', 'incom', 'incomeemploymentdis', 'increas', 'independ', 'indic', 'individu', 'individu mutual', 'individu mutual outcom', 'individu mutual outcom progress', 'individu patient', 'individu patient specif', 'induc', 'infect', 'inform', 'inform live', 'inform patient', 'inform patient sourc', 'inform patient sourc incom', 'infus', 'inhal', 'initi', 'initi human', 'initi human respons', 'initi human respons clinic', 'initi transit', 'initi transit plan', 'initi transit plan assess', 'inject', 'injuri', 'inpati', 'inpati psychiatr', 'inpati psychiatr unit', 'inpati psychiatr unit safeti', 'insecur', 'insecur worri', 'insight', 'instruct', 'insulin', 'insur', 'insur payor', 'intact', 'intak', 'interact', 'interact skill', 'intern', 'intervent', 'interview', 'interview patient', 'intraven', 'involv', 'issu', 'iv', 'jail', 'job', 'joint', 'judgement', 'kidney', 'knee', 'know', 'know problem', 'lab', 'lack', 'lack support', 'later', 'later date', 'later date loz', 'lcsw', 'leav', 'leave', 'leave procedur', 'leg', 'legal', 'legal histori', 'leisur', 'level', 'level orient', 'level orient provid', 'level orient provid care', 'life', 'life stressor', 'like', 'limit', 'limit social', 'limit social support', 'line', 'list', 'list diagnosi', 'list diagnosi loz', 'list potenti', 'list potenti problem', 'list potenti problem absent', 'listen', 'live', 'live alon', 'live alon absenc', 'live alon absenc caregiv', 'live environ', 'live environ homeless', 'live hi', 'live situat', 'live situat patient', 'live situat patient live', 'liver', 'local', 'locat', 'locat main', 'locat main unch', 'locat main unch servic', 'locationdetail', 'long', 'lose', 'loss', 'low', 'lower', 'loz', 'loz alcohol', 'loz alcohol abus', 'loz alcohol use', 'loz anxieti', 'loz anxieti loz', 'loz chronic', 'loz depress', 'loz depress loz', 'loz diabet', 'loz drug', 'loz drug use', 'loz drug use loz', 'loz file', 'loz file social', 'loz file social histori', 'loz hypertens', 'loz hypertens loz', 'loz know', 'loz know problem', 'loz marit', 'loz marit statu', 'loz nbsp', 'loz nbsp loz', 'loz nbsp loz nbsp', 'loz pr', 'loz pr upper', 'loz pr upper gi', 'loz psychiatr', 'loz sexual', 'loz sexual activ', 'loz smoke', 'loz smoke statu', 'loz smoke statu smoker', 'loz smokeless', 'loz smokeless tobacco', 'loz smokeless tobacco use', 'loz substanc', 'loz year', 'loz year educ', 'loz year educ na', 'lung', 'main', 'main topic', 'main topic loz', 'main topic loz smoke', 'main unch', 'main unch servic', 'main unch servic trauma', 'maintain', 'major', 'major depress', 'major depress disord', 'make', 'make progress', 'make progress outcom', 'maker', 'male', 'manag', 'manag assess', 'manag assess patient', 'manag assess patient person', 'marijuana', 'marit', 'marit statu', 'marit statu singl', 'marit statu singl spous', 'marri', 'matern', 'md', 'md locat', 'md locat

main', 'md locat main unch', 'md mg', 'md mg loz', 'mdd', 'meal', 'meal prep', 'mean', 'measur', 'med', 'medic', 'medic afford', 'medic assist', 'medic clearanc', 'medic clearanc patient', 'medic clearanc patient admit', 'medic histori', 'medic histori diagnosi', 'medic histori diagnosi date', 'medic histori past', 'medic histori past medic', 'medic manag', 'medic prior', 'medic prior overnight', 'medic prior overnight hospit', 'medic record', 'medic ye', 'medic ye comment', 'medicaid', 'medicar', 'medicin', 'meet', 'meet post', 'meet post hospit', 'meet post hospit care', 'mellitu', 'member', 'memori', 'mental', 'mental health', 'mental health servic', 'mental ill', 'mg', 'mg loz', 'mg mg', 'mg mg oral', 'mg oral', 'mg oral daili', 'mg tablet', 'mg tablet mg', 'mg tablet tablet', 'mg tablet tablet mg', 'mg total', 'mg total mouth', 'mgdl', 'middot', 'mild', 'militari', 'militari servic', 'militari servic abuseneglecttrauma', 'min', 'minut', 'ml', 'mlhr', 'mobil', 'moder', 'mom', 'money', 'monitor', 'month', 'mood', 'mood disord', 'mother', 'mother loz', 'motiv', 'motor', 'motor skill', 'mouth', 'mouth daili', 'multipl', 'multipli', 'multipli diagnos', 'multipli diagnos chronic', 'mutual', 'mutual outcom', 'mutual outcom progress', 'na', 'na loz', 'na patient', 'na procedur', 'namephon', 'narr', 'narr live', 'nbsp', 'nbsp loz', 'nbsp loz nbsp', 'nbsp loz nbsp loz', 'nc', 'nc phone', 'necessari', 'necessari medic', 'necessari medic clearanc', 'necessari medic clearanc patient', 'need', 'need assess', 'need assess concern', 'need assess concern address', 'need assess outcom', 'need assess outcom progress', 'need assist', 'need discharg', 'need discharg discharg', 'need discharg discharg facilitylevel', 'need financi', 'need financi assist', 'need financi assist discharg', 'need financi assist ye', 'need financi resourc', 'need financi resourc strain', 'need identifi', 'need identifi anticip', 'need identifi anticip comment', 'need medic', 'need patient', 'neg', 'neg hx', 'neg hx loz', 'neurolog', 'new', 'nightli', 'nonadher', 'noncompli', 'nonmed', 'normal', 'north', 'north carolina', 'north carolina guardianpaye', 'note', 'number', 'number children', 'nurs', 'nurs facil', 'nutrit', 'obes', 'object', 'obstruct', 'obtain', 'occup', 'occup histori', 'old', 'onc', 'ongo', 'onli', 'onset', 'onset loz', 'open', 'opioid', 'optim', 'optim cope', 'optim cope skill', 'optim cope skill respons', 'option', 'oral', 'oral daili', 'orang', 'order', 'organ', 'orient', 'orient level', 'orient level orient', 'orient level orient provid', 'orient provid', 'orient provid care', 'orient provid care home', 'ot', 'outcom', 'outcom dischargetransit', 'outcom dischargetransit care', 'outcom progress', 'outcom progress discharg', 'outcom progress discharg need', 'outcom progress overarch', 'outcom progress overarch goal', 'outpati', 'outpati dialysi', 'outpati dialysi financi', 'outpati dialysi financi inform', 'outpati provid', 'outpatientagencysupport', 'outpatientagencysupport group', 'outpatientagencysupport group need', 'outpatientcommun', 'outsid', 'overarch', 'overarch goal', 'overarch goal adult', 'overnight', 'overnight hospit', 'overnight hospit stay', 'overnight hospit stay ed', 'overview', 'overview goal', 'packsday', 'pain', 'pain loz', 'pantri', 'parent', 'particip', 'particip activ', 'particip activ particip', 'particip activ particip think', 'particip think', 'particip think process', 'partner', 'past', 'past medic', 'past medic histori', 'past medic histori diagnosi', 'past month', 'past surgic', 'past surgic histori', 'past surgic histori procedur', 'patern', 'patient', 'patient activ', 'patient activ problem', 'patient activ problem list', 'patient admit', 'patient admit inpati', 'patient admit inpati psychiatr', 'patient andor', 'patient andor famili', 'patient andor famili provid', 'patient care', 'patient care overview', 'patient care overview goal', 'patient demonstr', 'patient demonstr desir', 'patient demonstr desir outcom', 'patient doe', 'patient famili', 'patient ha', 'patient inform', 'patient inform live', 'patient live', 'patient live alon', 'patient person', 'patient person interview', 'patient person interview patient', 'patient report', 'patient risk', 'patient risk readmiss', 'patient risk readmiss ye', 'patient sourc', 'patient sourc incom', 'patient specif', 'patient specif goal', 'patient specif intervent', 'patient state', 'patient wa', 'patientfamili', 'patientfamili anticip', 'patientfamili anticip servic', 'patientfamili anticip servic transit', 'patientfamili anticip transit', 'pattern', 'pattern habit', 'pattern habit use', 'pay', 'pay financ', 'payor', 'payor medicaid', 'pcp', 'pe', 'pediatr', 'pend', 'peopl', 'perform', 'perform pattern', 'perform pattern habit', 'perform pattern habit use', 'perform routin', 'perform skill', 'perform skill motor', 'perform skill motor skill', 'person', 'person disord', 'person interview', 'person interview patient', 'pharmaci', 'phone', 'physic', 'physician', 'place', 'place prior', 'place prior admiss', 'place prior admiss equip', 'placeholdermetadatabegin', 'placeholdermetadataend', 'placeholdermetadataend grammarmetadatabegin', 'placeholdermetadataend grammarmetadatabegin grammarmetadataend', 'placeholdermetadataend grammarmetadatabegin

grammarmetadataend autorefreshmetadatabegin', 'placement', 'plan', 'plan assess', 'plan care', 'plan care review', 'plan care review outcom', 'plan care review patient', 'plan medicaid', 'plan need', 'plan need identifi', 'plan need identifi anticip', 'plan patient', 'plan safeti', 'plan safeti concern', 'plan safeti concern recommend', 'plan safeti continu', 'plan safeti continu admiss', 'plan screen', 'plan screen discharg', 'plan screen discharg plan', 'plan unsuccess', 'play', 'pm', 'po', 'point', 'poor', 'posit', 'possibl', 'post', 'post acut', 'post hospit', 'post hospit care', 'post hospit care need', 'posttraumat', 'posttraumat stress', 'posttraumat stress disord', 'potenti', 'potenti problem', 'potenti problem absent', 'potenti problem absent manag', 'pr', 'pr upper', 'pr upper gi', 'pr upper gi endoscopydiagnosi', 'practic', 'practic guidelin', 'practic guidelin cpg', 'precaut', 'prefer', 'prefer pharmaci', 'prep', 'prescrib', 'prescript', 'prescript coverag', 'present', 'pressur', 'prevent', 'previou', 'previou admiss', 'previou admiss day', 'previou admiss day anticip', 'previous', 'primari', 'primari care', 'primari contact', 'primari insur', 'primari insur payor', 'princip', 'princip problem', 'prior', 'prior admiss', 'prior admiss equip', 'prior admiss equip current', 'prior overnight', 'prior overnight hospit', 'prior overnight hospit stay', 'prison', 'privat', 'privat resid', 'prn', 'probat', 'problem', 'problem absent', 'problem absent manag', 'problem list', 'problem list diagnosi', 'problem list diagnosi loz', 'problem patient', 'problem patient care', 'problem patient care overview', 'problem relat', 'problem relat age', 'problem relat age onset', 'problem suicid', 'procedur', 'procedur later', 'procedur later date', 'procedur later date loz', 'procedur ugi', 'procedur ugi endo', 'procedur ugi endo includ', 'process', 'process skill', 'product', 'product type', 'product type product', 'product type product type', 'product type secondari', 'product type secondari insur', 'program', 'progress', 'progress discharg', 'progress discharg need', 'progress discharg need assess', 'progress improv', 'progress outcom', 'progress overarch', 'progress overarch goal', 'progress overarch goal adult', 'promot', 'protect', 'provid', 'provid care', 'provid care home', 'provid care home na', 'provid choic', 'provid choic facil', 'provid choic facil servic', 'psych', 'psychiatr', 'psychiatr care', 'psychiatr histori', 'psychiatr hospit', 'psychiatr unit', 'psychiatr unit safeti', 'psychiatr unit safeti stabil', 'psychiatri', 'psycholog', 'psychosi', 'psychosoci', 'psychot', 'pt', 'pt ha', 'pt live', 'pt report', 'pt state', 'ptsd', 'public', 'public transport', 'pulmonari', 'pulmonari diseas', 'puls', 'question', 'quit', 'rafhcc', 'rafhcc activ', 'rafhcc activ problem', 'rafhcc loz', 'rafhcc stressor', 'raleigh', 'rash', 'rate', 'rd', 'reach', 'readmiss', 'readmiss day', 'readmiss day previou', 'readmiss day previou admiss', 'readmiss ye', 'readmiss ye discharg', 'readmiss ye discharg plan', 'reason', 'reason admiss', 'receiv', 'receiv food', 'receiv food stamp', 'receiv outpati', 'receiv outpati dialysi', 'receiv outpati dialysi financi', 'recent', 'recommend', 'recommend follow', 'recommend follow ani', 'recommend follow ani necessari', 'record', 'recoveri', 'recurr', 'ref', 'ref ring', 'refer', 'referr', 'refil', 'reflux', 'refus', 'regard', 'regul', 'regul skill', 'regular', 'relat', 'relat age', 'relat age onset', 'relat age onset loz', 'relat ill', 'relat ill equip', 'relat ill equip current', 'relat ill equip need', 'relat ill inabl', 'relat ill inabl care', 'relat risk', 'relat risk factor', 'relat risk factor sign', 'relationship', 'relationship statu', 'releas', 'remain', 'remiss', 'remov', 'renal', 'repair', 'report', 'report ha', 'request', 'requir', 'resid', 'resid privat', 'resid privat resid', 'resolv', 'resourc', 'resourc strain', 'respiratori', 'respons', 'respons clinic', 'respons clinic practic', 'respons clinic practic guidelin', 'respons complet', 'respons complet homemak', 'respons life', 'respons life stressor', 'responsibilitiesdepend', 'responsibilitiesdepend home', 'responsibilitiesdepend home home', 'responsibilitiesdepend home home care', 'rest', 'rest sleep', 'restrict', 'result', 'retir', 'return', 'review', 'review abov', 'review abov inform', 'review outcom', 'review outcom progress', 'review patient', 'review plan', 'rex', 'right', 'ring', 'risk', 'risk adult', 'risk factor', 'risk factor multipli', 'risk factor multipli diagnos', 'risk factor sign', 'risk factor sign symptom', 'risk readmiss', 'risk readmiss ye', 'risk readmiss ye discharg', 'rn', 'ro', 'role', 'room', 'roundsfamili', 'roundsfamili conf', 'rout', 'routin', 'routin damag', 'routin satisfi', 'rule', 'run', 'safeti', 'safeti concern', 'safeti concern recommend', 'safeti concern recommend follow', 'safeti consider', 'safeti consider self', 'safeti continu', 'safeti continu admiss', 'safeti continu admiss inpati', 'safeti judgement', 'safeti stabil', 'safeti stabil treatment', 'satisfi', 'say', 'scale', 'scale estim', 'scale whoda', 'schedul', 'schizoaffect', 'schizoaffect disord', 'schizophrenia', 'school', 'screen', 'screen discharg', 'screen discharg plan', 'screen discharg plan need', 'secondari', 'secondari insur', 'secur', 'seizur', 'self', 'selfcar', 'sensori', 'sensori function', 'servic', 'servic abuseneglecttrauma', 'servic avail', 'servic avail appropri',

'servic avail appropri meet', 'servic gastroenterolog', 'servic gastroenterolog loz', 'servic gastroenterolog loz pr', 'servic place', 'servic place prior', 'servic place prior admiss', 'servic transit', 'servic trauma', 'session', 'set', 'sever', 'sex', 'sexual', 'sexual activ', 'sexual activ file', 'sexual activ file topic', 'shelter', 'shop', 'short', 'short goal', 'si', 'sig', 'sign', 'sign symptom', 'sign symptom identifi', 'sign symptom identifi initi', 'sign symptom list', 'sign symptom list potenti', 'sign symptom relat', 'sign symptom relat risk', 'signific', 'signssymptom', 'sinc', 'singl', 'singl spous', 'sister', 'sit', 'site', 'situat', 'situat patient', 'situat patient live', 'skill', 'skill impair', 'skill motor', 'skill motor skill', 'skill respons', 'skill respons life', 'skill respons life stressor', 'skin', 'sleep', 'smartlistmetadatabegin', 'smartlistmetadatabegin smartlistmetadataend', 'smartlistmetadatabegin smartlistmetadataend wildcardmetadatabegin', 'smartlistmetadataend', 'smartlistmetadataend wildcardmetadatabegin', 'smoke', 'smoke statu', 'smoke statu current', 'smoke statu current everi', 'smoke statu smoker', 'smokeless', 'smokeless tobacco', 'smokeless tobacco use', 'smokeless tobacco use loz', 'smoker', 'smoker packsday', 'snf', 'social', 'social histori', 'social histori loz', 'social histori loz marit', 'social histori main', 'social histori main topic', 'social histori narr', 'social histori narr live', 'social histori social', 'social histori social histori', 'social histori socioeconom', 'social histori socioeconom histori', 'social interact', 'social interact skill', 'social need', 'social need financi', 'social need financi resourc', 'social support', 'social work', 'socioeconom', 'socioeconom histori', 'socioeconom histori marit', 'socioeconom histori marit statu', 'soft', 'son', 'sourc', 'sourc incom', 'sp', 'speak', 'specif', 'specif goal', 'specif intervent', 'specifi', 'speech', 'spend', 'spiritu', 'spiritu belief', 'spous', 'spous na', 'sq', 'sq cm', 'ssi', 'stabil', 'stabil treatment', 'stabl', 'stamp', 'standard', 'start', 'state', 'statu', 'statu current', 'statu current everi', 'statu current everi day', 'statu divorc', 'statu marri', 'statu singl', 'statu singl spous', 'statu smoker', 'stay', 'stay ed', 'stay ed visit', 'stay ed visit day', 'step', 'stomach', 'stop', 'strain', 'strategi', 'street', 'strength', 'stress', 'stress disord', 'stressor', 'stressor chronic', 'stroke', 'structur', 'studi', 'subject', 'subscrib', 'subscrib number', 'substanc', 'substanc abus', 'substanc use', 'success', 'success transit', 'suicid', 'suicid attempt', 'suicid ideat', 'suicid risk', 'summari', 'supervis', 'suppli', 'support', 'support famili', 'support famili member', 'support measur', 'support success', 'support success transit', 'surgeon', 'surgeri', 'surgeri loz', 'surgic', 'surgic histori', 'surgic histori past', 'surgic histori procedur', 'surgic histori procedur later', 'surrog', 'surrog decis', 'surrog decis maker', 'sw', 'symptom', 'symptom identifi', 'symptom identifi initi', 'symptom identifi initi human', 'symptom list', 'symptom list potenti', 'symptom list potenti problem', 'symptom relat', 'symptom relat risk', 'symptom relat risk factor', 'syndrom', 'tablet', 'tablet mg', 'tablet mg mg', 'tablet mg mg oral', 'tablet mg total', 'tablet mg total mouth', 'tablet tablet', 'tablet tablet mg', 'tablet tablet mg total', 'talk', 'tbd', 'team', 'temp', 'test', 'therapeut', 'therapeut allianc', 'therapeut allianc support', 'therapeut allianc support success', 'therapi', 'therapist', 'thi', 'thi time', 'think', 'think process', 'thyroid', 'time', 'time day', 'time discharg', 'time frame', 'time frame week', 'time patient', 'tissu', 'tobacco', 'tobacco use', 'tobacco use disord', 'tobacco use loz', 'tobacco use loz alcohol', 'tobacco use smoke', 'tobacco use smoke statu', 'today', 'toe', 'toler', 'topic', 'topic concern', 'topic concern loz', 'topic concern loz file', 'topic loz', 'topic loz smoke', 'topic loz smoke statu', 'total', 'total mouth', 'transfer', 'transit', 'transit home', 'transit plan', 'transit plan assess', 'transport', 'transport anticip', 'transport avail', 'transport need', 'transport need medic', 'trauma', 'trauma loz', 'travel', 'treatment', 'treatment plan', 'treatment team', 'tri', 'trial', 'trigger', 'tx', 'type', 'type financi', 'type financi assist', 'type financi assist requir', 'type product', 'type product type', 'type product type secondari', 'type resid', 'type resid privat', 'type resid privat resid', 'type secondari', 'type secondari insur', 'ugi', 'ugi endo', 'ugi endo includ', 'ugi endo includ esophagu', 'ulcer', 'unabl', 'unc', 'unch', 'unch servic', 'unch servic trauma', 'uncl', 'understand', 'unemploy', 'uninsur', 'unit', 'unit safeti', 'unit safeti stabil', 'unit safeti stabil treatment', 'unknown', 'unspecifi', 'unsuccess', 'updat', 'upper', 'upper gi', 'upper gi endoscopydiagnosi', 'urin', 'use', 'use disord', 'use disord sever', 'use home', 'use home current', 'use home current receiv', 'use home equip', 'use home equip need', 'use loz', 'use loz alcohol', 'use loz alcohol use', 'use loz sexual', 'use loz sexual activ', 'use smoke', 'use smoke statu', 'use ye', 'valu', 'valu ref', 'valu ref ring', 'vascular', 'vaya', 'vaya health', 'vehicl', 'verbal', 'veri', 'violenc', 'visit', 'visit day', 'visit day readmiss', 'visit day readmiss day', 'vital', 'vocat', 'voic', 'vs', 'vte', 'vte dvt', 'vte dvt pe', 'wa', 'wake', 'wake counti',

'wake north', 'wake north carolina', 'walk', 'walker', 'want', 'week', 'weight', 'weight bear', 'wfl', 'whoda', 'wife', 'wildcardmetadatabegin', 'wind', 'wish', 'withdraw', 'women', 'work', 'worri', 'wwo', 'ye', 'ye comment', 'ye discharg', 'ye discharg plan', 'ye discharg plan screen', 'ye type', 'ye type financi', 'ye type financi assist', 'year', 'year ago', 'year educ', 'year educ na', 'year old']

CHAPTER 4: IDENTIFYING SOCIAL DETERMINANTS OF HEALTH IN CLINICAL NOTES: COMPARISON OF TEXT MINING AND MACHINE LEARNING BASED APPROACHES

Introduction

In the United States, health and illness are not equally distributed due to social and economic disparities[1–3]. The root cause of such disparities commonly referred to as social determinants of health (SDH) involves indirect factors such as accessible healthcare and education systems, safe communities and environments, and availability of food. Increasingly, health policy organizations and health researchers call for the inclusion of SDH within all health care systems[4–6], as health care continues to transition to value-based care and the widespread adoption of electronic health records (EHRs). Recent research evaluating the SDH integration into the EHR[7–9] found rare instances of SDH documentation and lack of clinical notes standardization as well as data collection[10,11]. Researchers propose that institution-specific language used to express SDH may limit the usefulness of current standardization clinical terminology tools such as the Unified Medical Language System (UMLS)[10,12]. A more popular approach to SDH identification in clinical text includes novel natural language processing (NLP) methods using machine learning or advanced text mining techniques also commonly referred to as rule-based NLP[12–14].

Traditional NLP applications used either rule-based or machine learning approaches. Rule-based NLP systems are primarily based on pre-defined vocabularies that include complex clinical logic. While rule-based systems often have lower performance in comparison with machine learning systems[15–17] they remain popular due to the less intensive technical skills required for the development or use of machine learning systems[17,18]. However, developing hand-crafted rule-based systems require a subject matter expert (SME) making it time consuming, expensive, and less generalizable[18]. On the other hand, machine

learning NLP systems are often based on probabilistic statistical approaches that have been shown to be more accurate and generalizable than rule-based systems[15,19,20]. With so few research studies exploring SDH in clinical notes it is unknown which approach would yield the highest performance and a set of generalizable SDH terminologies.

If simpler, lower cost methods such as auto-encoding from a dictionary of SDH terminologies perform with high specificity and sensitivity remains an open research question. The UMLS framework, including the four major medical vocabularies: Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), Logical Observation Identifiers Names and Codes (LOINC), International Classification of Diseases, 10th Revision (ICD-10), and Current Procedural Terminology (CPT); contain approximately 1,000 unique codes related to SDH[21]. Despite the vast number of existing codes in common medical coding vocabulary that could facilitate documentation of SDH in the clinical setting, Aarons and colleagues found significant gaps noting high specificity but low sensitivity in housing insecurity and substance abuse[21]. For example, they found 14 codes under the subdomain of housing "Utilities" (e.g. SNOMED CT codes 423798004 "Lack of cooling in house;" 105535008 "Lack of heat in house) but not a code that captured the inability to pay for utilities. While advanced NLP approaches have had more success there has been limited research into the costs of developing and maintaining machine learning models in the clinical setting [22]. Additionally, even a highly accurate machine learning system is not necessarily sufficient in and of itself to be routinely utilized and endorsed by clinical staff[23,24]. Lundberg and colleagues found that physicians using machine learning models in clinical practice preferred simpler models that were more interpretable at the expense of lower accuracy[24].

In this study we (1) evaluate the performance of auto-encoding, rule-based, and machine learning based NLP systems to automatically classify the financial resource strain and poor social support SDH characteristics in clinical notes using three approaches, (2) describe the associated costs of these different

approaches;, and (3) evaluate the generalizability of SDH terminologies in open source clinical notes. Unlike previous studies, our experiment were designed to improve clinician decision making about algorithm adoption based on performance, interpretability, and cost.

Background

Related work

Automated encoding. Prior studies[25–29] explored methods that automatically map clinical text to concepts within a standardized coding system, such as the UMLS. The UMLS encompasses terminology libraries such as SNOMED-CT and LOINC[28] that describe conditions (i.e. asthma) and modifiers (i.e. severe). These approaches, primarily have been applied to indexing applications, used methods based on string matching, statistical processing, or linguistic processing such tagging (i.e. classifying words as nouns, verbs, etc.). This encoding approach is typically leveraged for machine learning development or more advanced NLP techniques[25,26,30,31]. Hatef and colleagues successfully developed a hybrid analysis approach to identify homelessness using SNOMED-CT terminologies and LOINC codes in addition to text-mining techniques such a word matching. They successfully found evidence of homeless characteristics, without investing in costly manual annotation. The authors concluded that off-the-shelf data extraction solutions were insufficient to identify SDH data in contrast to standardized terminologies for conditions or diagnostic codes. Automated encoding also proved successful for Zeng and colleagues who developed a tool to identify local recurrences of breast cancer in clinical notes by concept mapping UMLS terms and report for an AUC score of 0.93[32]. The IBM Watson Health group used word embedding models to find new SDH terminology from public corpora such as Wikipedia articles and compared SDH term concepts with the UMLS and SNOMED-CT[33]. Overall, a large number of new terms did not match either (on average 76% in SNOMED-CT and 83% in UMLS), although SNOMED-CT did provide a better coverage throughout[33]. Our study uses the concept of terminology matching utilizing an SDH dictionary of generalized terms developed in Chapter 2.

Rule-based NLP systems. Clinical text classification tasks that used common rule-based NLP methods have been successfully applied in other clinical domains[15,34–36]. Carrell and colleagues developed a 1288 term dictionary to create NLP rules to identify opioid use disorder data within unstructured text for computer aided manual annotation[35]. Hollister and colleagues extracted socioeconomic status information (e.g., insurance status, education, and homelessness) from clinical notes using a rule-based NLP system. Algorithms performance on 50 patient records varied, reported PPV results ranged from 87.5% for unemployment to 33.3% for homelessness. These findings suggest that some categories of socioeconomic status may easier to extract using NLP rules than others[14] most likely due to the limited associated vocabulary. Using Australian domestic violence police reports, others[37] designed and implemented a rule-based model that combined language expression patterns with dictionary terms to identify abuse types and victim injuries from text. The authors reported a precision of 90.2% and 85.0% for abuse type and victim injuries, respectively. Therefore, we used a similar approach of using a dictionary and SME to build custom NLP rules for a model to classify SDH.

Machine-learning NLP systems. Machine learning methods successfully identified specific SDH characteristics in clinical notes including smoking status[38,39], suicide risk factors[40–42], and homelessness[10,43]. Feller and colleagues found that gradient boosting algorithms yielded the highest performance (F1=92.6) when inferring homelessness using only unstructured EHR data[10]. A meta-analysis of studies identifying smoking status using NLP machine learning approaches using an open source n2c2 dataset from the 2006 NLP smoking challenge found 12 approaches with micro-average F1 measures above 0.84[39].

Methods

Study corpus, annotation schema, and gold-standard development

For this study, we used clinical notes from the University of North Carolina Clinical Data Warehouse (UNC-CDW) and n2c2 NLP unstructured notes (formerly i2b2 NLP data sets) from the Partners Healthcare Research Patient Data Repository at Harvard University[44]. We developed two corpora to evaluate the generalizability of our NLP approach for identifying SDH. To provide a granular analysis of SDH representations in EHR documentation, we chose to develop our SDH detection algorithm for sentence-level vs document-level data. This study was approved by the UNC Institutional Review Board.

The UNC corpus was comprised of 49,171 adult patients who frequently used the ED (≥ 4 visits in 365 days) and diagnosed at least once with a mental health or substance use disorder (MHSUD) final primary diagnosis. We queried all clinical notes for this population available between April 2014-December 2019, resulting in approximately 15 million clinical notes. Using terms and regular expressions previously validated in literature[12–14,31,45–47], we explored these notes for SDH characteristics and expanded the SDH term list using a word embedding model validated by SMEs (Chapter 2) to build a comprehensive SDH ontology dictionary (Appendix 1) to describe financial resource strain and poor social support. To create a gold-standard corpus of clinical notes, ontology was used to identify a small proportion of likely positive clinical note sentences (N= 2,045), to which we added an additional 2,018, randomly selected negative SDH sentences to balance an initially overly positive dataset (total N=4,063) comprising 1,119 patients from the target population. .

Two SME annotators manually reviewed the gold standard corpus (N=4,063) to label sentences as (1) housing insecurity (homelessness, unstable housing), (2) food insecurity (food stamps, unable to afford food), (3) employment and income insecurity (unemployment, insufficient income), (4) general financial insecurity (lack of transportation, other financial issues), (5) insurance insecurity (uninsured,

underinsured), and (6) poor social support (social isolation, lack of social support) as guided by the IOM's "Capturing Social and Behavioral Domains and Measure in Electronic Health Records"[48]. A third SME adjudicated 255 disagreements. Since we achieved an overall inter-rater reliability (IRR) agreement of 86.6% we shifted to a single reviewer for negation sampling. Using a single reviewer after a sample of annotations with high IRR has shown to reduce costs without losing accuracy[19,49].This SME classification served as the gold standard for comparison with the auto-coded classification models.

Publicly available, the n2c2 database (National NLP Clinical Challenges) was a comprehensive and anonymized unstructured clinical notes collected from the Partners Healthcare System in Boston from 2004-2014. These datasets are associated with specific NLP tasks such as de-identification, and identifying smoking status and heart disease risk factors. We selected the entire corpus used in the 2006 smoking challenge, 2008 obesity challenge, and 2014 heart disease risk factor challenge to create a corpus of 2,797 clinical notes representing 814 patients (2-5 records per patient).

Text pre-processing

We first pre-processed all clinical notes using the following routine: we lowercased all the letters, removed all metadata and/or special characters, and split into individual sentences using Natural Language ToolKit (NLTK)[50]. We removed all sentences that containing only numbers or a single word. A number of redundant sentences were found in the UNC clinical notes, created by copy and paste or auto-generation, a situation that presents a number of challenges for training and evaluating text mining and machine learning models[51–53]. These duplicates were removed to derive an unbiased estimate of SDH documentation.

Model development

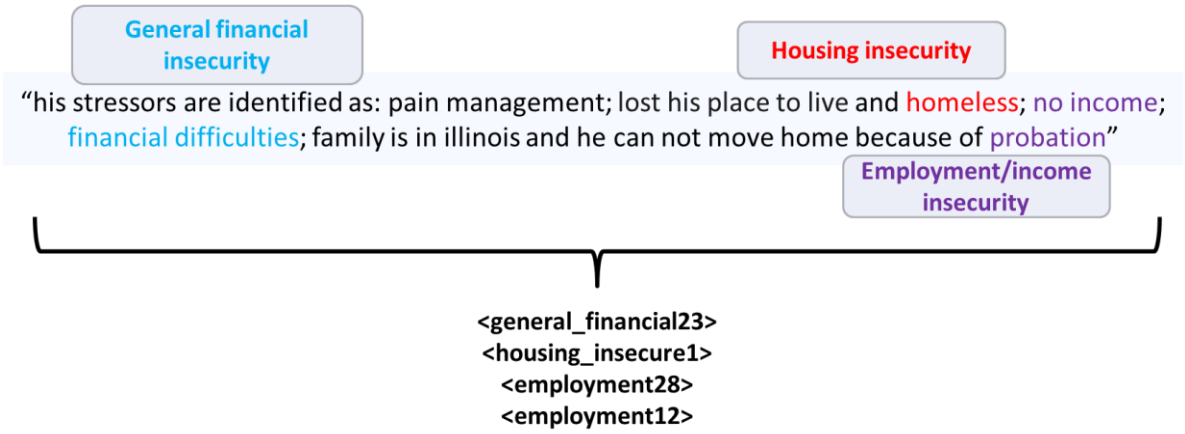
The rule-based system and auto-encoding model was based on a comprehensive SDH clinical vocabulary of 110 terms and phrases to accurately identify financial resource strain and poor social

support (Appendix 1). Creating the SDH dictionary for the models was one of the most time consuming steps of this project (Chapter 1). Development of the dictionary used a combination of literature review of previously validated SDH terms[12–14,45,46,54,55], standardized vocabulary libraries (UMLS, SNOMED-CT, and LOINC), and observational terms that were then reviewed by the study experts. Selection of terms was based on pre-analysis work evaluating frequency of occurrence, evaluation by SME, and generalizability. For example, the term *homeless* has the same meaning regardless of region; however, concepts surrounding public transportation reliance may only be considered a SDH characteristic in a suburban or rural region where public transit may not be adequate.

SDH encoding

First, we investigated the low-cost, high-yield text mining approach of exact and partial string matching to identify and classify SDH documentation. Exact string matching exactly matches a term or phrase to the target (case-sensitive). For example, if the phrase "unemployed" appears in a sentence it would automatically be encoded as employment insecure with its associated unique code regardless of the context of words surrounding the phrase including negation. Partial string matching uses word stems to detect when the target word appears within a sentence. The Porter stemming algorithm was utilized through the NLTK Python library[56]. This method more efficiently captures root forms of target words. For example, the partial match of "homeless" within "homelessness". This encoding approach is the foundation for more complex concept and term mapping schemas found in the UMLS framework[28]. The encoding phase maps a target term in the SDH dictionary to its unique code so that the output can be stored in a clinical repository or other structured architecture (Figure 1).

Figure 1. Encoding example



Rule-based system development

Our method involved the design and implementation of rule-based language expression patterns combined with dictionary terms for the recognition of financial resource strain and poor social support at the sentence level. We based our rules on syntactical patterns identified in the training and development sets, indicating the presence of a SDH characteristic.

1. SDH concept-based rules: The SDH dictionary constituted a repository of concepts relevant to the task. Presence of these concepts in sentences under certain conditions resulted in generation of SDH annotations. These rules were completely based on SDH concepts found in the UNC dataset and did not rely on regular expressions. For example, the term *afford* was categorized as general financial insecurity; however, the term *afford* surrounded by the term medication, co-pay, and/or therapy was tagged as insurance insecurity.

2. *Regular expression-based rules*: irregular language, such as non-standard abbreviations and misspellings commonly found in clinical notes may cause medical dictionaries to fail [16,34]. For example, the term *Medicaid* was often expressed as *mcd, mcaid,* and *Medcaid.* These expressions may be regionally dependent causing a generalizable SDH dictionary to fail to associate the

expressions with insurance insecurity. Therefore, we included regular expressions to account for such cases.

3. *Negation detection*: auto-encoding systems may generate false positives when the concept of interest may be reported as a negated finding in the note (i.e. "The patient was screened negative for food insecurity"). The concept may also be associated with someone other than the patient, as is commonly found in the family history section of a typical note[57,58]. Finally, auto generated text prompts left blank or marked as 'none' (i.e. "transportation: none, financial assistance: none, living arrangements: none, who provides care at home?: none, patient information: none") may contribute to false findings. Therefore, we developed an additional custom negation rule to address these variations.

We did not explore temporal markers such as *past*, *present*, or *future* concerns related to SDH. Rule-based model was not developed for the n2c2 dataset as there were not enough SDH positive annotations to develop custom rules.

Machine learning approach

In our supervised machine learning strategy we implemented a binary relevance Support Vector Machine (BR-SVM) and the gradient boosting tree algorithm XGBoost using Sci-kit Learn[59]. Each model was trained (80%, UNC = 3,250, n2c2=4,816) and tested (20%, UNC=813, n2c2=1,205) on their respective gold-standard dataset that were developed using the approach described earlier; however, the n2c2 datasets were manually inspected and annotated by only one SME. Inputs into the classification models included a single free-text clinical note sentence. Both of these models use a bag-of-words feature representation in which each clinical note sentence was converted to a high-dimensional vector in which each dimension represented the term frequency-inverse document frequency (TF-IDF) score of all 1, 2, 3, and 4 g (n-grams) in the note[60]. The TF-IDF weights were used for feature selection, limited to a

maximum of 2,000 features, the maximum features parameter is used to set a limit on the number of feature to select. To reduce bias in the validation models, a five-fold cross validation was used to evaluate model results. Performance metrics were micro-averaged across all folds.

Evaluation

Our text mining system was evaluated on the training set, previously unseen, and randomly chosen clinical note sentences. The UNC trained models (auto-encoder, rule-based, machine learning) were evaluated against a gold-standard dataset. The inter-rater reliability was calculated for overall and by label agreement. Performance of the methodology was evaluated at the sentence level. For both methods, we calculated the precision (the number of true positives against the number of true positives and false positives), recall (the number of true positives against the number of true positives and false negatives), and F1-score (the harmonic mean between precision and recall) across all labels and per label for each model. The decision to optimize precision or recall may depend on the specific clinical application, therefore F1 was considered the primary evaluation metric[10,61] because it combines both. We defined true positive as the detection of a correct mention in an event; false positive as the extraction of any unrelated mention that has not been annotated manually; false negative as the correct mention that was not detected by our method; and true negative as the case where our method did not identify any mentions when none were annotated. Only 10% negative sampling was completed on the n2c2 dataset so true negatives were not calculated for those models. Additionally, the n2c2 auto-encoding model used is a reflection of the entire dataset (N=6,021 sentences) for better insights how a larger sample size would perform (machine learning models are only evaluated on 20% training dataset). We conducted an error analysis to gain insight into model performance for SDH labels. We manually reviewed all incorrectly labeled sentences and analyzed false negatives using a classification matrix and attempted to evaluate each error as an incorrect annotation, unrecognized negation, or confusing auto-generated structure.

To assess annotation costs we calculated the mean total personnel cost of employing annotators to annotate the corpora. We arrived at this cost by multiplying the mean number of hours annotators spent annotating by the mean personnel cost of the same annotators in current (2020) dollars[49,62]. Time and cost estimates were based on actual methodology we used in this study, reducing from two to a single annotator after the first gold-standard dataset was completed and respectable agreement rate achieved. All source code can be found at *www.github.com/rstem/dissertation*.

Results

Corpora characteristics

UNC corpus. A total of 4,063 clinical note sentences associated with 1,119 patients who frequented the ED with a MHSUD diagnosis were manually reviewed for characteristics of SDH that describe financial resource strain and poor social support (Table 1). Of the positive SDH sentences (N=1,066) 75.0% of them had two or more annotated SDH labels. SME annotators had high inter-rater reliability agreement with an 86.6% overall; insurance insecurity had the lowest agreement (79.1%) and food insecurity had the greatest (90.6%). On average, each sentence was 83.2 words long, top N words were *patient*, *discharge*, *care*, and *history*.

n2c2 dataset. A total of 6,021 sentences were extrapolated from the original dataset of 2,797 clinical notes containing 209,318 sentences representing 814 patients. A total of 256 SDH characteristics within the corpus were categorized by our SME. Poor social support (N=152) was the most frequent, food insecurity least frequent (N=) across all n2c2 datasets. In general, the combined datasets had a low rate of SDH characteristics (< 1%), the obesity challenge dataset had the highest percentage of SDH (11%) (Table 2). When the datasets were combined, the average sentence length was 30.3 words, top N words were *patient*, *discharge*, and *admission*. The top bigrams for the combined dataset were *the patient*, *patient was*, and *lives alone* with *she lives alone* in the top trigrams.

TABLE 1. Manually annotated SDH characteristics

SDH label	UNC dataset	n2c2 dataset
General financial insecurity	428	5
Employment and/or income insecurity	686	52
Housing insecurity	502	15
Insurance insecurity	437	10
Food insecurity	321	1
Poor social support	530	108
Total	2904	191

TABLE 2. n2c2 manually annotated SDH characteristics

	n2c2 dataset			
	Obesity	**Smoking status**	**Heart disease**	**Total**
# of clinical notes	1118	889	790	2797
# of sentences	97720	67270	44328	209318
% SDH positive	0.11%	0.06%	0.10%	0.09%
SDH labels				
General financial insecurity	1	3	1	5
Employment/income insecurity	29	13	10	52
Housing insecurity	8	2	5	15
Insurance insecurity	4	3	3	10
Food insecurity	0	0	1	1
Poor social support	67	18	23	108

Total	109	39	43	191

Text mining and ML model performance

When comparing the overall performance of the models XGBoost algorithm had the highest F1 score for both datasets (UNC; F1=0.80, n2c2; F1=0.86) (Table 3). For overall evaluation, all models were evaluated on their respective training datasets (UNC=813, n2c2=1,205). For both data sources, XGBoost had the highest precision and SVM had the highest recall micro-averaged across all labels. The model with the lowest fraction of wrong labels to the total number of labels, also known as the Hamming loss, was XGBoost (UNC=4.67, n2c2=0.21).

TABLE 3. Overall performance of SDH models

Data source	Algorithm	Avg precision	Avg recall	Avg F1	Hamming loss	Avg precision-recall
UNC	auto-encoded	0.77	0.75	0.76	5.6	0.61
	rule-based	0.74	0.77	0.76	5.86	0.6
	SVM	0.64	0.84	0.72	7.97	0.56
	XGBoost	0.86	0.75	0.8	4.67	0.68
n2c2	auto-encoded	0.85	0.8	0.83	0.32	0.68
	SVM	0.87	0.85	0.86	0.23	0.74
	XGBoost	0.97	0.78	0.86	0.21	0.75

Auto-encoding performance

Auto-encoding of the corpora was completed using a SDH dictionary (Appendix 1). The SDH dictionary contained a total of 110 terms or phrases categorized as general financial insecurity (17), employment and/or income insecurity (31), housing insecurity (28), insurance insecurity (12), food insecurity (5), and poor social support (17). The UNC data source had 84 terms represented while only 19 were represented in the n2c2 data source. The most represented in the SDH dictionary terms were *homeless* (335) and *afford* (298) within the UNC data source and *lives alone* (134) and *unemployed* (26) within the n2c2 data source (Appendix 1). Analysis of the UNC data source true positive rate (TPR) resulted in 18 dictionary terms with >90% TPR and 5 terms = 100% TPR (lack of transportation, disability income, unemployment, trespassing, and on disability). The same analysis of the n2c2 data source resulted in 4 terms > 90% TPR and 3 terms = 100% TPR (on disability, unemployed, and homeless).

Overall, performance of auto-encoding models was successful with F1 scores of 0.76 and 0.83 for the UNC and n2c2 data sources respectively (Table 3). Because food insecurity was only represented once in the n2c2 data source, it was not included in any of the n2c2 models. Auto-encoding was very effective in detecting poor social support SDH label, producing the highest metrics with exception of recall in the n2c2 housing insecurity label (Table 4). General financial insecurity had the lowest performance across all performance metrics.

TABLE 4. Auto-encoding model performance by SDH label

Data source	SDH label	Precision (PPV)	Recall (sensitivity)	F1	TP	TN	FP	FN
UNC	General financial insecurity	0.43	0.54	0.48	231	3328	309	195

n2c2	Employment/income insecurity	0.91	0.67	0.77	3333	458	43	229
	Housing insecurity	0.79	0.85	0.82	426	3447	114	76
	Insurance insecurity	0.8	0.74	0.74	301	3552	76	134
	Food insecurity	0.81	0.87	0.87	297	3674	69	23
	Poor social support	0.96	0.91	0.91	435	3511	21	76
	General financial insecurity	0.2	0.2	0.2	1	N/A	4	4
	Employment/income insecurity	0.81	0.85	0.83	60	N/A	14	11
	Housing insecurity	0.89	0.94	0.91	16	N/A	2	1
	Insurance insecurity	1	0.73	0.84	8	N/A	0	3
	Food insecurity	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Poor social support	0.97	0.87	0.92	132	N/A	4	20

Rule-based model performance

When evaluating by SDH label, the micro-averaged F1 improved among all labels with exception of employment and/or income insecurity and poor social support. However, the implemented custom NLP rules only increased F1 micro-averaged by 0.1 across all labels among the UNC data source. The rule-based model did not outperform auto-encoding across all SDH label performance measures. For example, the rule-based model increased the recall significantly, but at the sacrifice of precision (Table 5). This is due to a high number of false negatives (FN=733) that were reduced by the implementation of the rules (FN=565). Custom rules could not be developed for the n2c2 data source since too few SDH characteristics identified by the SME annotator.

TABLE 5. Comparison of UNC data source model performance by SDH label

Model	SDH label	Precision (PPV)	Recall (sensitivity)	F1	TP	TN	FP	FN
Auto-encode	General financial insecurity	0.43	0.54	0.48	231	3328	309	195
	Employment/income insecurity	0.91	0.67	0.77	3333	458	43	229
	Housing insecurity	0.79	0.85	0.82	426	3447	114	76
	Insurance insecurity	0.80	0.74	0.74	301	3552	76	134
	Food insecurity	0.81	0.87	0.87	297	3674	69	23
	Poor social support	0.96	0.91	0.91	435	3511	21	76
Rule-based	General financial insecurity	0.46	0.64	0.54	271	3318	319	155
	Employment/income insecurity	0.66	0.80	0.72	3096	549	280	138
	Housing insecurity	0.79	0.89	0.84	445	3444	117	57
	Insurance insecurity	0.79	0.71	0.75	307	3544	84	128
	Food insecurity	0.91	0.93	0.92	297	3714	29	23
	Poor social support	0.90	0.88	0.89	467	3482	50	64

Feature selection for machine learning models

To gain insights into feature selection for predicted positive labels by the algorithms we outputted the TF-IDF matrix for each input (i.e. sentence). Text features used by the classifiers included explicit indicators of SDH, as well as co-occurring determinants. For example, top features for poor social support, the SDH label with the largest sample size in the n2c2 dataset, was *lives alone*, *home care*, and *social history*. Top features in positive labeled SDH sentences within the UNC data source were similar in that they often were bigrams and vocabulary terminology from the SDH dictionary. Mutual top features included *disabled*, *social history*, and *homeless*.

Machine learning model performance

XGBoost was the highest performer for the UNC data source, but equal in F1 performance for the n2c2 data source (Table 3), outperforming the SVM model. All labels had an improvement in precision with the XGBoost model (Table 6), but a slight decrease in recall performance within the UNC data source. Minimal differences occurred between the machine learning models developed with the n2c2 data source.

TABLE 6. Machine learning performance by SDH label

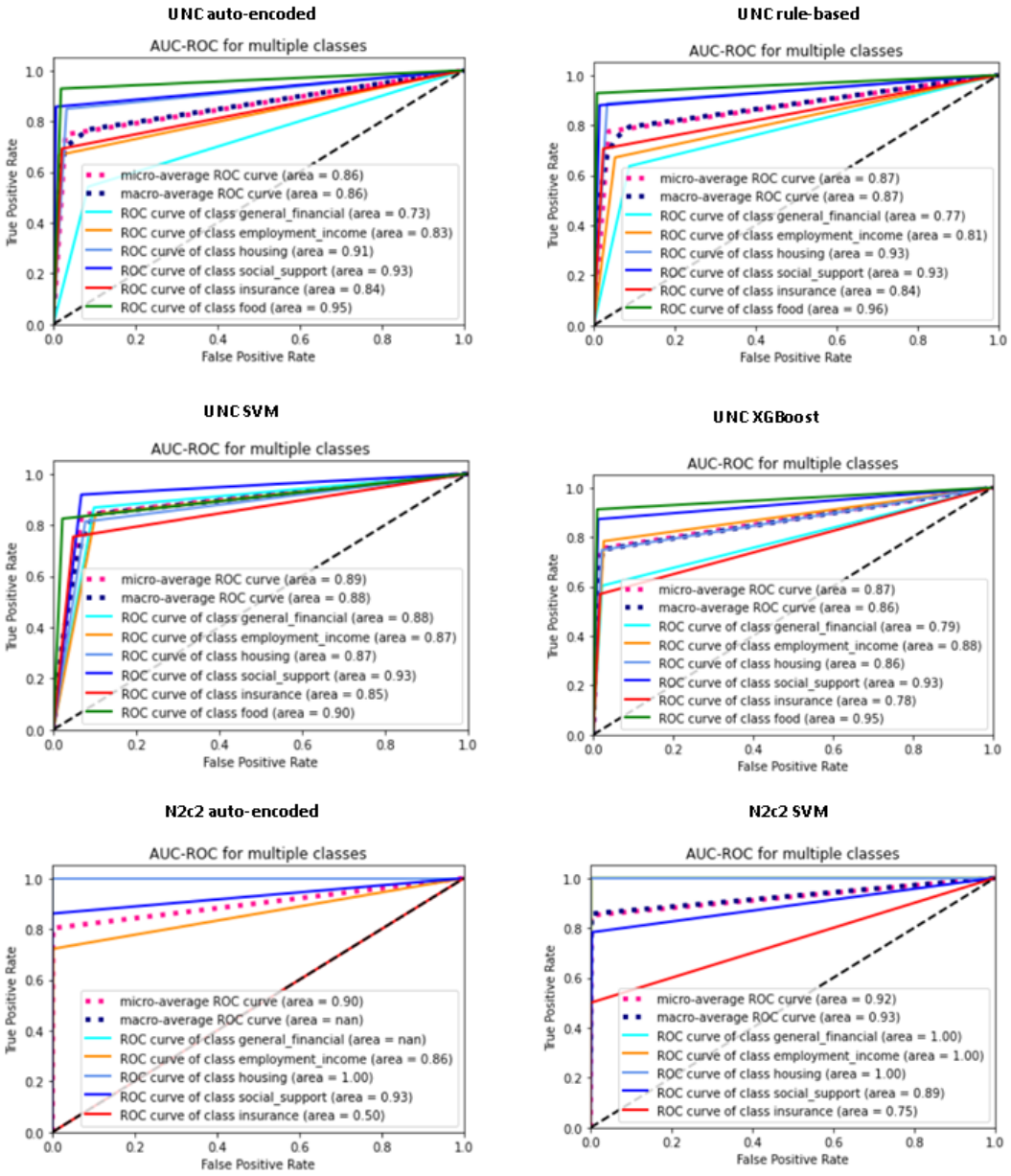
Data source	Model	SDH label	Precision (PPV)	Recall or sensitivity	F1	TP	TN	FP	FN
UNC	SVM	General financial insecurity	0.51	0.87	0.64	59	525	57	9
		Employment/income insecurity	0.65	0.84	0.73	101	476	54	19
		Housing insecurity	0.63	0.81	0.71	74	516	43	17
		Insurance insecurity	0.64	0.75	0.69	49	557	28	16

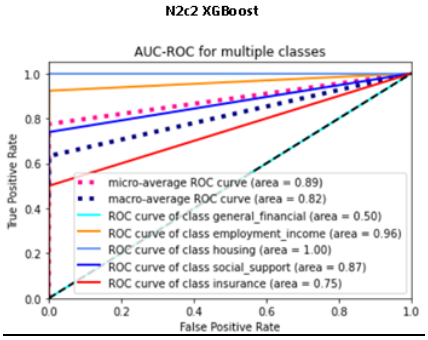
n2c2	XGBoost	Food insecurity	0.78	0.82	0.8	47	580	13	10
		Poor social support	0.68	0.92	0.78	79	526	38	7
		General financial insecurity	0.84	0.71	0.77	50	569	13	18
		Employment/income insecurity	0.87	0.78	0.82	94	516	14	26
		Housing insecurity	0.85	0.75	0.8	68	547	12	23
		Insurance insecurity	0.79	0.57	0.66	75	557	7	11
	SVM	Food insecurity	0.9	0.91	0.9	37	575	10	28
		Poor social support	0.91	0.87	0.89	52	587	6	5
		General financial insecurity	0.5	1	0.67	1	962	1	0
		Employment/income insecurity	0.93	1	0.96	13	950	1	0
		Housing insecurity	1	1	1	1	963	0	0
		Insurance insecurity	1	0.5	0.67	1	962	0	1
	XGBoost	Food insecurity	n/a	n/a	n/a	n/a	n/a	n/a	n/a
		Poor social support	0.86	0.78	0.82	18	938	3	5
		General financial insecurity	0	0	n/a	0	963	0	1
		Employment/income insecurity	0.92	0.92	0.92	12	950	1	1
		Housing insecurity	1	1	1	1	963	0	0
		Insurance insecurity	1	0.5	0.67	1	962	0	1

	Food insecurity	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Poor social support	1	0.74	0.85	17	941	0	6

The Area under the Receiver Operating Characteristic (AUC-ROC) curve in Figure 2 shows the performance of the models at all classification thresholds by plotting the true positive rate vs false positive rate. All models have very similar AUC, however the SVM model for both data sources was more consistent in terms of true-positive rate vs. false-positive rate (for all thresholds).

FIGURE 2. AUC-ROC for SDH labels



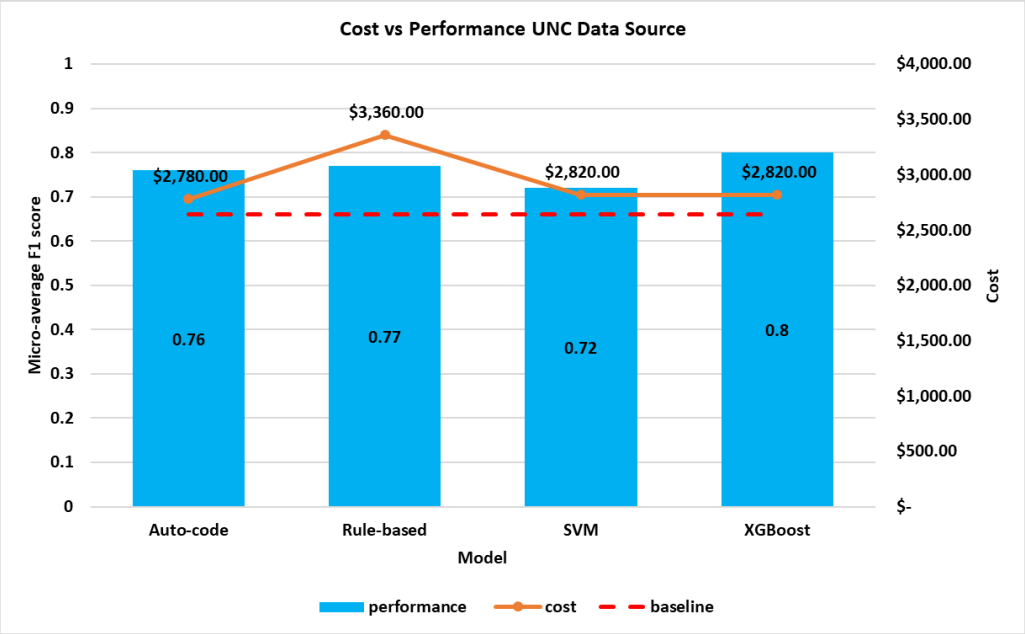


Cost vs performance

Mean time spent annotating the UNC corpus by individual annotators was 42 hours. With a mean total personnel cost of $40.00 per hour (in 2020 dollars) the mean per annotator cost of annotating the corpus was $840.00 (42 * 40). Additional annotation costs were accrued during third annotator adjudication for disagreements and the single annotator for negative sampling. These costs were totaled with labor hours (6) and personnel cost per hour ($20.00) resulting in the total costs for creating the gold-standard were $1,800.00. Given that the 2,045 sentences contained 1,066 SDH positive instances, this resulted in annotation costs of $1.69 per SDH instance and $0.88 per annotated sentence.

Figure 3 shows the estimated cost of developing each model and their respective micro-averaged F1 score. Estimated costs of developing our models were derived from two different labor tasks of research (i.e. literature review) and coding. We developed a baseline cost for model derived from the gold-standard annotation costs and SDH dictionary development that was the foundation for all model methodology. All labor costs were $20 per hour based on the hourly stipend for a research assistant with the necessary skills required for machine learning model development. Rule-based model ($3,360.00) had the highest costs associated with it due to the additional coding demand for the custom rules and linguistic research required to develop SDH rules. XGBoost had the best performance for the least cost, adding an additional $40.00 for 0.04 improvement in F1.

FIGURE 3. Cost vs performance of UNC data source



Error analysis

A number of false negatives (FN=733) occurred for the auto-encoded UNC data source primarily with the employment and/or income insecurity (FN=229) label followed closely by general financial insecurity (FN=195) (Table 4). Many false negatives related to the term *afford* occurred using the auto-encoding model. These errors occurred when *afford* in conjunction with different label characteristic identifiers such as *medications* for insurance insecurity or *living expenses* for employment and/or income insecurity. Complex concepts surrounding housing and income insecurity lead to the majority of false negatives when auto-encoding general financial insecurity. For example, "stressors: unemployment, limited financial resources, and conflict with mother leading to unstable housing". Overall a low number of false negatives were seen among all models for the n2c2 data source; however, only 191 actual positives were found by annotators. Poor social support (FN=20) and employment and/or income insecurity (FN=11) had the highest number of false negatives for any model. Social support false negatives primarily occurred due to complex documentation about discharge concerns to include *no*

supportive assistance surround poor were due to variations of being unable or concerns surrounding discharging a patient home. Many employment false negatives were attributed to health professionals' references to Veterans Day, not veteran status.

The SVM model trained on the UNC data source incorrectly labeled 118 sentences. The 52 false negatives were mostly attributed to confusion between general financial insecurity and employment and/or income insecurity. Error analysis for n2c2 data source is limited due to the small positive label sample size. Poor social support had the highest number of false negatives for both machine learning models (SVM=5, XGBoost=6) also related to general financial insecurity and employment and/or income insecurity.

Discussion

This study compared three information extraction approaches at identifying SDH characteristics among clinical notes from two different data sources: auto-encoding, rule-based, and machine learning (SVM and XGBoost). We demonstrated that a variety of approaches, both simple and complex, can have high performance when classifying SDH characteristics in clinical notes. Our study is the first to our knowledge to evaluate the generalizability of SDH terminology in open source clinical notes.

The first approach, auto-encoding, was based only on term matching from a SDH dictionary developed by SMEs in the clinical setting and informatics. No other studies were found that built a model based only term matching for SDH identification in clinical notes. However, our auto-encoding models demonstrated similar or better performance compared to other studies using matching dictionary term in other domains[37]. For example, a study that examined domestic violence in police reports reported similar F1 (0.85) and recall (0.86)[37]. Our auto-encoding model showed that it is possible to use a simple term matching method with a comprehensive SDH dictionary in both an SDH rich (F1=0.76) and general (F1=0.83) population. Within the UNC data source, this approach outperformed the SVM model

(F1=0.72) discussed later in approach three. The programming required for this approach was lowest (< 1 hour) and required the least amount of technical skill; however, the development of the SDH dictionary required a significant amount of time and subject matter expertise. The high precision (0.85), recall (0.80), and F1 (0.83) of this approach on the open source n2c2 data set suggests that researchers may utilize the SDH dictionary without customization in similar settings, populations, or institution. At a minimum, this SDH dictionary provides a comprehensive foundation for any future work.

Our second approach used a rule-based NLP method with hand-crafted logical rules for each SDH category by a SME. In traditional clinical NLP rule-based systems, rules are built based on clinical criteria such as heart failure diagnosis[63]; however, SDH language is more expressive and nuanced[10,33,64] suggesting that a SDH rule-based system may be limited to the corpus that it was developed on. In a recent study, a rule-based system was developed to extract social risk factors from the Veteran's Affair EHR clinical text using a similar approach to crafting rules through linguistic patterns and SDH term matching from vocabulary found in the UMLS framework[64]. The resulting system classified poor social support at similar overall performance (F1=0.84) to our model, but had much lower PPV (0.78). Our model's PPV (0.79) outperformed multiple rule-based studies system designed to classify homeless or marginally housed patients (0.66, 0.33)[14,64]. Rule-based methods are often selected over machine learning approaches because they are typically substantially more interpretable, when classification decisions can be clearly articulated.

Rule-based NLP systems are often crafted to fit a certain domain and type of clinical language. This was also the case in our study: our rule-based system achieved better F1 performance than the auto-encoding (0.76) and SVM (0.72) models. While our rule-based model had higher F1 when classifying food, housing, and insurance insecurity classes; however, the overall model had lower precision (PPV) in all SDH labels with exception of food insecurity. This is due to the reduced number of false positives in

the model as a result of the negation rule we developed to deal with auto-generated text. Linguistic pattern rules reduced false negatives in all SDH labels. For example, a rule was created for the employment label where a positive instance was auto-labeled for the text prompt “financial difficulties?:” followed by “yes”. This, in combination with other custom rules lead to a reduction of 91 false negatives in comparison with the auto-encoding model. However, these SME hand-crafted NLP rules were time consuming to create and therefore expensive. Development of rules also requires a moderate sample size of positive instances, something we could not achieve with the n2c2 data set due to insufficient numbers. Therefore, the generalizability of NLP rules is a concept that still needs to be further explored.

Our third NLP approach introduced a probabilistic (SVM) and ensemble decision tree (XGBoost) model for SDH text classification. In the open data source provided by the n2c2 challenges performance was high for both the machine learning models (F1=0.86). XGBoost model achieved higher precision (0.97) and lower Hamming loss (0.21); however SVM achieved higher recall (0.85) and an average ROC (92.5%). Both machine learning approaches achieved similar performance across metrics on the UNC data source. Due to no one single evaluation measure capturing all the desirable properties of a model[65], several measures were reported to provide a more accurate representation of clinical efficacy. We chose a multi-label classification as our model because, in the UNC data source, 75% of SDH positive sentences had more than one SDH label, but only 12 of the 244 positive sentences in the n2c2 data set were multi-label. Unfortunately, the overall feature importance matrix is not currently available in open source computing libraries for multi-label learning models. However, the feature matrix is computable per input (i.e. sentence) and may provide granular-level insights. The largest cost associated with the machine learning method was the labor and time required to development of a gold-standard corpora for training, development, and testing of models.

Limitations

This study has several important limitations. First, the n2c2 data set is not considered gold-standard. In our study, only one SME annotator completed the labeling and negative sampling. Second, the sample size for the n2c2 data set was too small for custom rules to be developed and limited the analysis of the machine learning models in this dataset. Future work should continue to evaluate generalizability by exploring other open source datasets (e.g. Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)) or a multi-site research proposal. Third, no food insecurity representations were identified in the manual review of the n2c2 data set and therefore was not included as a SDH category in any of the models. This may be solved through a larger data source as stated above. Fourth, we did not conduct a formal evaluation of interpretability from clinicians of the models instead guided by informal focus groups with a diverse background SMEs (i.e. physicians, clinical researchers, social workers, etc.). Future work should explore a user-centered machine learning design to formally test interpretability.

Conclusion

This study evaluated the performance, interpretability, generalizability, and cost of different approaches to text classification of SDH characteristics in clinical notes. Our machine learning model outperformed (F1 measure) all other approaches in identifying the SDH characteristics financial resource strain and poor social support in data sets from a single health care system and open source clinical notes. Auto-encoding had slightly lower performance, but also cost less and was easily interpretable. All models achieved excellent results in the challenging task of identifying SDH characteristics from multiple source clinical notes. We believe that exploring a range of approaches from simple to complex are of high interest to researchers and clinical practitioners, especially for health care systems with less resources and access to machine learning expertise. Importantly, our promising results suggest that clinical text mining and machine learning can be implemented on any dataset and potentially without the need of large labeled datasets typically necessary for machine learning. Finally, our auto-encoding system may potentially be

used by almost any institution or software without special informatics training, a limitation experience by many existing health systems.

RERFERENCES

1. Chetty R, Stepner M, Abraham S, et al. The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA*. 2016;315(16):1750-1766. doi:10.1001/jama.2016.4226

2. Daniel H, Bornstein SS, Kane GC, Health and Public Policy Committee of the American College of Physicians. Addressing social determinants to improve patient care and promote health equity: an american college of physicians position paper. *Ann Intern Med*. 2018;168(8):577-578. doi:10.7326/M17-2441

3. Social Determinants of Health | Healthy People 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health. Accessed March 13, 2020.

4. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. Washington (DC): National Academies Press (US); 2014. doi:10.17226/18709

5. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Med*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

6. Advancing Health Equity: Principles to Address the Social Determinants of Health in Alternative Payment Models. https://www.aafp.org/about/policies/all/socialdeterminants-paymentmodels.html. Accessed July 9, 2020.

7. Gold R, Cottrell E, Bunce A, et al. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *J Am Board Fam Med*. 2017;30(4):428-447. doi:10.3122/jabfm.2017.04.170046

8. Freij M, Dullabh P, Lewis S, Smith SR, Hovey L, Dhopeshwarkar R. Incorporating social determinants of health in electronic health records: qualitative study of current practices among top vendors. *JMIR Med Inform*. 2019;7(2):e13849. doi:10.2196/13849

9. Gottlieb L, Tobey R, Cantor J, Hessler D, Adler NE. Integrating social and medical data to improve population health: opportunities and barriers. *Health Aff (Millwood)*. 2016;35(11):2116-2123. doi:10.1377/hlthaff.2016.0723

10. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data. *Appl Clin Inform*. 2020;11(1):172-181. doi:10.1055/s-0040-1702214

11. Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. *Health Aff (Millwood)*. 2018;37(4):585-590. doi:10.1377/hlthaff.2017.1252

12. Bejan CA, Angiolillo J, Conway D, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc*. 2018;25(1):61-71. doi:10.1093/jamia/ocx059

13. Feller DJ, Zucker J, Don't Walk OB, et al. Towards the Inference of Social and Behavioral Determinants of Sexual Health: Development of a Gold-Standard Corpus with Semi-Supervised Learning. *AMIA Annu Symp Proc*. 2018;2018:422-429.

14. Hollister BM, Restrepo NA, Farber-Eger E, Crawford DC, Aldrich MC, Non A. Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. *Pac Symp Biocomput*. 2017;22:230-241. doi:10.1142/9789813207813_0023

15. Topaz M, Murga L, Gaddis KM, et al. Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *J Biomed Inform*. 2019;90:103103. doi:10.1016/j.jbi.2019.103103

16. Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. *JAMA Netw Open*. 2018;1(6):e183451. doi:10.1001/jamanetworkopen.2018.3451

17. Malmasi S, Ge W, Hosomura N, Turchin A. Comparison of natural language processing techniques in analysis of sparse clinical data: insulin decline by patients. *AMIA Jt Summits Transl Sci Proc*. 2019;2019:610-619.

18. Blom M, Nobile N, Suen CY, eds. *Frontiers In Pattern Recognition And Artificial Intelligence*. World Scientific; 2019.

19. Shivade C, Malewadkar P, Fosler-Lussier E, Lai AM. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. *J Biomed Inform*. 2015;58 Suppl:S103-10. doi:10.1016/j.jbi.2015.08.025

20. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review. *J Biomed Inform*. 2018;77:34-49. doi:10.1016/j.jbi.2017.11.011

21. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. *JAMIA Open*. 2019;2(1):81-88. doi:10.1093/jamiaopen/ooy051

22. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. 2020;368:l6927. doi:10.1136/bmj.l6927

23. Masino AJ, Harris MC, Forsyth D, et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PLoS One*. 2019;14(2):e0212665. doi:10.1371/journal.pone.0212665

24. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-760. doi:10.1038/s41551-018-0304-0

25. Weng W-H, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak*. 2017;17(1):155. doi:10.1186/s12911-017-0556-8

26. Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Med Inform Decis Mak*. 2018;18(Suppl 3):74. doi:10.1186/s12911-018-0654-2

27. Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. *J Am Med Inform Assoc*. 2001;8(1):80-91. doi:10.1136/jamia.2001.0080080

28. Introduction to the UMLS. https://www.ncbi.nlm.nih.gov/books/NBK9675/. Published 2009. Accessed July 6, 2020.

29. Peterson KJ, Liu H. Automating the Transformation of Free-Text Clinical Problems into SNOMED CT Expressions. *AMIA Jt Summits Transl Sci Proc*. 2020;2020:497-506.

30. Doan S, Maehara CK, Chaparro JD, et al. Building a Natural Language Processing Tool to Identify Patients With High Clinical Suspicion for Kawasaki Disease from Emergency Department Notes. *Acad Emerg Med*. 2016;23(5):628-636. doi:10.1111/acem.12925

31. Hatef E, Rouhizadeh M, Tia I, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system. *JMIR Med Inform*. 2019;7(3):e13802. doi:10.2196/13802

32. Zeng Z, Espino S, Roy A, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinformatics*. 2018;19(Suppl 17):498. doi:10.1186/s12859-018-2466-x

33. Bettencourt-Silva JH, Mulligan N, Sbodio M, et al. Discovering New Social Determinants of Health Concepts from Unstructured Data: Framework and Evaluation. *Stud Health Technol Inform*. 2020;270:173-177. doi:10.3233/SHTI200145

34. Hegde H, Shimpi N, Glurich I, Acharya A. Tobacco use status from clinical notes using Natural Language Processing and rule based algorithm. *Technol Health Care*. March 2018. doi:10.3233/THC-171127

35. Carrell DS, Cronkite D, Palmer RE, et al. Using natural language processing to identify problem usage of prescription opioids. *Int J Med Inform*. 2015;84(12):1057-1064. doi:10.1016/j.ijmedinf.2015.09.002

36. MacRae J, Love T, Baker MG, et al. Identifying influenza-like illness presentation from unstructured general practice clinical narrative using a text classifier rule-based expert system versus a clinical expert. *BMC Med Inform Decis Mak*. 2015;15:78. doi:10.1186/s12911-015-0201-3

37. Karystianis G, Adily A, Schofield PW, et al. Automated analysis of domestic violence police reports to explore abuse types and victim injuries: text mining study. *J Med Internet Res*. 2019;21(3):e13067. doi:10.2196/13067

38. Wang Y, Sohn S, Liu S, et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak*. 2019;19(1):1. doi:10.1186/s12911-018-0723-6

39. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*. 2008;15(1):14-24. doi:10.1197/jamia.M2408

40. Carson NJ, Mullin B, Sanchez MJ, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PLoS One*. 2019;14(2):e0211116. doi:10.1371/journal.pone.0211116

41. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing. *Sci Rep*. 2018;8(1):7426. doi:10.1038/s41598-018-25773-2

42. Zhang Y, Zhang OR, Li R, et al. Psychiatric stressor recognition from clinical notes to reveal association with suicide. *Health Informatics J*. October 2018:1460458218796598. doi:10.1177/1460458218796598

43. 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.

44. n2c2 NLP Research Data Sets. https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/. Accessed July 10, 2020.

45. Blosnich JR, Marsiglio MC, Dichter ME, et al. Impact of Social Determinants of Health on Medical Conditions Among Transgender Veterans. *Am J Prev Med*. 2017;52(4):491-498. doi:10.1016/j.amepre.2016.12.019

46. Winden TJ, Chen ES, Melton GB. Representing residence, living situation, and living conditions: an evaluation of terminologies, standards, guidelines, and measures/surveys. *AMIA Annu Symp Proc*. 2016;2016:2072-2081.

47. UMLS - Metathesaurus. https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html. Accessed February 4, 2019.

48. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington (DC): National Academies Press (US); 2015. doi:10.17226/18951

49. Carrell DS, Cronkite DJ, Malin BA, Aberdeen JS, Hirschman L. Is the Juice Worth the Squeeze? Costs and Benefits of Multiple Human Annotators for Clinical Text De-identification. *Methods Inf Med*. 2016;55(4):356-364. doi:10.3414/ME15-01-0122

50. Natural Language Toolkit — NLTK 3.5 documentation. https://www.nltk.org/. Accessed May 12, 2020.

51. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*. 2013;14:10. doi:10.1186/1471-2105-14-10

52. Zhang R, Pakhomov SV, Lee JT, Melton GB. Using language models to identify relevant new information in inpatient clinical notes. *AMIA Annu Symp Proc*. 2014;2014:1268-1276.

53. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. *J Am Med Inform Assoc*. 2015;22(5):938-947. doi:10.1093/jamia/ocv032

54. Bazemore AW, Cottrell EK, Gold R, et al. Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. *J Am Med Inform Assoc*. 2016;23(2):407-412. doi:10.1093/jamia/ocv088

55. Navathe AS, Zhong F, Lei VJ, et al. Hospital Readmission and Social Risk Factors Identified from Physician Notes. *Health Serv Res*. 2018;53(2):1110-1136. doi:10.1111/1475-6773.12670

56. nltk.stem.porter — NLTK 3.5 documentation. https://www.nltk.org/_modules/nltk/stem/porter.html. Accessed August 22, 2020.

57. Wu S, Miller T, Masanz J, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PLoS One*. 2014;9(11):e112774. doi:10.1371/journal.pone.0112774

58. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*. 2011;18(5):540-543. doi:10.1136/amiajnl-2011-000465

59. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011.

60. Tf-idf weighting. https://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html. Accessed January 27, 2019.

61. Jiang M, Chen Y, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc*. 2011;18(5):601-606. doi:10.1136/amiajnl-2011-000163

62. Wei Q, Franklin A, Cohen T, Xu H. Clinical text annotation - what factors are associated with the cost of time? *AMIA Annu Symp Proc*. 2018;2018:1552-1560.

63. Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Int J Med Inform*. 2014;83(12):983-992. doi:10.1016/j.ijmedinf.2012.12.005

64. Conway M, Keyhani S, Christensen L, et al. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. *J Biomed Semantics*. 2019;10(1):6. doi:10.1186/s13326-019-0198-0

65. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195. doi:10.1186/s12916-019-1426-2

APPENDIX 1. SDH dictionary

		UNC data source				n2c2 data source			
SDH code	**Term or Phrase**	**TP**	**FN**	**Total**	**TPR**	**TP**	**FN**	**Total**	**TPR**
housing1	homeless	310	25	335	0.93	10	0	10	1.00
housing2	shelter	78	21	99	0.79	2	0	2	1.00
housing4	streets	9	5	14	0.64	0	2	2	0.00
housing5	motel	10	19	29	0.34	0	0	0	n/a
housing6	evict	19	3	22	0.86	2	0	2	1.00
housing8	foreclosed	1	0	1	1.00	0	0	0	n/a
housing9	landlord	1	17	18	0.06	0	0	0	n/a
housing10	banned	0	0	0	n/a	0	0	0	n/a
housing11	lack of satisfaction with housing	0	0	0	n/a	0	0	0	n/a
housing12	housing insecure	8	1	9	0.89	0	0	0	n/a
housing14	hoarder	1	0	1	1.00	0	0	0	n/a
housing15	transitional housing	1	0	1	1.00	0	0	0	n/a
housing16	cluttered	0	0	0	n/a	0	0	0	n/a
housing17	housing crisis	0	1	1	0.00	0	0	0	n/a
housing18	housing issue	0	0	0	n/a	0	0	0	n/a
housing19	hotel	3	0	3	1.00	0	0	0	n/a
housing20	infested	23	17	40	0.58	0	0	0	n/a
housing21	pay rent	0	0	0	n/a	0	0	0	n/a
housing22	unstable housing	2	1	3	0.67	0	0	0	n/a
housing23	pay mortgage	0	0	0	n/a	0	0	0	n/a
housing24	flooded	3	1	4	0.75	0	0	0	n/a
housing25	public housing	0	0	0	n/a	0	0	0	n/a
housing26	boarding house	0	0	0	n/a	2	0	2	1.00
housing27	rescue mission	4	2	6	0.67	0	0	0	n/a
housing28	lost her home	13	4	17	0.76	0	0	0	n/a
housing29	lost his home	0	0	0	n/a	0	0	0	n/a
housing30	lack of stable housing	0	0	0	n/a	0	0	0	n/a
housing31	mortgage assistance	0	0	0	n/a	0	0	0	n/a
general1	afford	74	224	298	0.25	1	2	3	0.33
general5	lack of resources	4	0	4	1.00	0	0	0	n/a
general6	financial stressors	5	0	5	1.00	0	0	0	n/a
general7	financial concerns	12	3	15	0.80	0	0	0	n/a
general8	financially	3	3	6	0.50	0	0	0	n/a
general9	subsidized	1	1	2	0.50	0	0	0	n/a
general12	lack of transportation	18	0	18	1.00	0	0	0	n/a
general13	finances	69	32	101	0.68	0	0	0	n/a

general15	payments	1	23	24	0.04	0	2	2	0.00
general16	financial strain	35	3	38	0.92	0	0	0	n/a
general17	bankruptcy	1	1	2	0.50	0	0	0	n/a
general18	transportation problems	0	0	0	0.00	0	0	0	n/a
general19	financial constraints	2	0	2	1.00	0	0	0	n/a
general20	financial issues	18	4	22	0.82	0	0	0	n/a
general22	money issues	0	0	0	0.00	0	0	0	n/a
general23	financial difficulties	18	5	23	0.78	0	0	0	n/a
general25	financial stress	15	2	17	0.88	0	0	0	n/a
insurance1	uninsured	34	10	44	0.77	0	0	0	n/a
insurance2	medicaid	191	38	229	0.83	2	0	2	1.00
insurance3	charity care	51	2	53	0.96	0	0	0	n/a
insurance4	pays out of pocket	4	0	4	1.00	0	0	0	n/a
insurance5	self pay	0	1	1	0.00	0	0	0	n/a
insurance6	no insurance	15	7	22	0.68	2	0	2	1.00
insurance7	lost insurance	0	1	1	0.00	0	0	0	n/a
insurance8	copay	27	30	57	0.47	4	0	4	1.00
insurance9	cheaper	0	1	1	0.00	0	0	0	n/a
insurance10	selfpay	0	0	0	n/a	0	0	0	n/a
insurance11	affordable medication	0	0	0	n/a	0	0	0	n/a
insurance12	co-pay	0	0	0	n/a	0	0	0	n/a
employment 2	disability income	11	0	11	1.00	0	0	0	n/a
employment 3	unemployment	38	0	38	1.00	0	0	0	n/a
employment 4	prison	37	11	48	0.77	5	0	5	1.00
employment 5	jail	54	15	69	0.78	2	0	2	1.00
employment 6	prostitute	0	0	0	n/a	0	1	1	0.00
employment 7	prostitution	6	1	7	0.86	1	0	1	1.00
employment 8	trespassing	13	0	13	1.00	0	0	0	n/a
employment 9	veteran	10	5	15	0.67	3	9	12	0.25
employment 10	forging	0	0	0	0.00	0	0	0	n/a
employment 11	shoplifting	3	0	3	1.00	0	0	0	n/a
employment 12	probation	31	2	33	0.94	0	0	0	n/a

employment 13	parole	6	1	7	0.86	0	0	0	n/a
employment 14	taken into custody	0	0	0	n/a	0	0	0	n/a
employment 15	court date	33	3	36	0.92	0	0	0	n/a
employment 17	job loss	3	0	3	1.00	0	0	0	n/a
employment 18	unemployed	175	2	177	0.99	26	0	26	1.00
employment 19	lost job	3	1	4	0.75	0	0	0	n/a
employment 20	receives disability	7	0	7	1.00	0	0	0	n/a
employment 21	receives ssi	6	1	7	0.86	0	0	0	n/a
employment 22	receives ssdi	1	1	2	0.50	0	0	0	n/a
employment 23	losing job	0	0	0	n/a	0	0	0	n/a
employment 24	loss of job	6	0	6	1.00	0	0	0	n/a
employment 25	not employed	9	0	9	1.00	0	0	0	n/a
employment 26	difficulty maintaining employment	0	0	0	n/a	0	0	0	n/a
employment 27	employment difficulties	0	0	0	n/a	0	0	0	n/a
employment 28	no income	12	1	13	0.92	0	0	0	n/a
employment 29	limited income	9	1	10	0.90	0	0	0	n/a
employment 30	unstable employment	1	0	1	1.00	0	0	0	n/a
employment 31	receiving disability	4	1	5	0.80	0	0	0	n/a
employment 32	on disability	33	0	33	1.00	18	0	18	1.00
employment 33	fixed income	0	0	0	n/a	0	0	0	n/a
food1	food pantries	53	5	58	0.91	0	0	0	n/a
food2	food stamps	209	10	219	0.95	0	0	0	n/a
food3	food insecure	1	1	2	0.50	0	0	0	n/a
food4	food insecurity	65	52	117	0.56	0	0	0	n/a
food5	food bank	2	7	9	0.22	0	0	0	n/a
social_support1	lives alone	201	14	215	0.93	130	4	134	0.97

social_support2	no family support	2	0	2	1.00	0	0	0	n/a
social_support3	no social support	3	0	3	1.00	0	0	0	n/a
social_support4	lack of assistance	0	0	0	0.00	0	0	0	n/a
social_support5	limited social support	138	1	139	0.99	0	0	0	n/a
social_support6	lack of caregivers	7	0	7	1.00	0	0	0	n/a
social_support7	lack of support	103	4	107	0.96	0	0	0	n/a
social_support8	living alone	0	1	1	0.00	2	0	2	1.00
social_support9	lack of social support	6	0	6	1.00	0	0	0	n/a
social_support10	poor social support	9	2	11	0.82	0	0	0	n/a
social_support11	social needs	4	29	33	0.12	0	0	0	n/a
social_support12	lack of adequate family	0	0	0	n/a	0	0	0	n/a
social_support13	care giver support	0	0	0	n/a	0	0	0	n/a
social_support14	without a support system	0	0	0	n/a	0	0	0	n/a
social_support15	social crisis	1	0	1	1.00	0	0	0	n/a
social_support16	social isolaiton	9	1	10	0.90	0	0	0	n/a
social_support17	limited support	7	0	7	1.00	0	0	0	n/a

CHAPTER 5: CONCLUSION

Our findings demonstrate the feasibility of identifying SDH characteristics financial resource strain and poor social support from electronic health record clinical notes. Implementation could increase both the quality of provider documentation related to these determinants as well as the availability of this information at the point of care. This study was guided by the 2014 the Institute of Medicine published two reports that made recommendations on which social and behavioral-related measures to use for data collection in EHRs[1,2]. Despite this level of interest, no uniform, accepted data model exists for identifying or documenting SDH in EHRs[3,4]. Data standardization is important for implementing appropriate clinical decision support interventions to address SDH within an EHR system and across different systems. SDH can be represented in demographic data elements (such as housing status), diagnoses (homelessness), or procedures (referral to supportive housing). However, no single current biomedical standard captures the breadth of information necessary for documenting SDH in a manner appropriate for clinical care, quality improvement, and research[4–6]. Current research suggests when SDH are documented they appear unstructured data or free-text clinical notes[6–8]. An efficient and effective methodology for systematically capturing SDH from clinical notes could lead to improved patient and population health outcomes.

In Chapter 2, we demonstrated the feasibility of developing an SDH dictionary (111 terms) using published, validated terminology and expanding on those terms through a word embedding model. The word embedding model approach, supported by subject matter experts (SMEs), produced a high quality SDH rich corpus that found SDH terminology among 79.3% of corpus patients (N=38,971). This methodology supports similar applications of applying word embedding for terminology expansion in clinical notes[9,10] suggesting this may be an efficient method of developing rich data sources for rare

events. The highly positive SDH dataset was used to train and test multiple NLP and machine learning models for classifying SDH in Chapters 3 and 4. We created high performing models to identify the SDH characteristics financial resource strain and poor social support were developed by being trained on an SDH rich corpus. The qualitative analysis of a large SDH rich corpus revealed a difference between a topic type and topic pattern. For example, a topic type description of food insecurity used natural language (i.e. "the patient stated he does not have enough money for regular meals") and did not conform to the pattern rules created by a screening survey or auto-generated narrative (i.e. "food insecurity: yes"). This is likely due to an organizational change with certain departments and clinics within the larger health system integrating SDH screens and questions within their patient assessments. These changes may be consequential, such as that the predictions made by a model trained on older historical data are no longer correct or as correct as they could be if the model was trained on more recent historical data. Many data mining methods assume that discovered patterns are static; however, in practice patterns in the data evolve over time[11]. This poses two important challenges. The first challenge is to detect when concept drift occurs and the second is to keep the patterns up-to-date without inducing the patterns from scratch.

A large number of medical codes exist for SDH; however, there are persistent gaps in the capacity of current medical vocabularies to identify SDH[5,12]. Our work makes significant progress towards a comprehensive SDH dictionary from which to build a variety of NLP and machine learning classification tasks. Current text classification tasks in clinical notes suffer without a standard terminology and may lead to a learning bias[9,13]. In the study presented in Chapter 2, we applied word embedding models to overcome this limitation[9] and attempted to more comprehensively generate terms that characterize financial resource strain and poor social support. Our results support the current hypothesis that semantic variants of specific topics appear in similar context in EHR clinical notes and that applying word embedding models based on distributional semantics improves detection of these syntactic and semantic variants[9,14,15]. Unlike previous studies[9,16,17], our approach integrated a SME supported by

Electronic Medical Record Search Engine[18] (EMERSE) to search clinical notes using terms and phrases of interest. EMERSE aided the identification and validation of SDH terminology in the context of the patient's entire clinical notes appearing in its clinical format, a structure that was lost during data export of the clinical notes. Future research should further leverage EMERSE for advanced text mining and machine learning classification tasks. A significant limitation of this study was the use of a single institutions' data to develop the SDH dictionary. Future research should explore using this approach to develop an SDH dictionary on multiple large clinical note repositories. To our knowledge, this is the first SDH dictionary developed to characterize financial resource strain and poor social support for information extraction in clinical notes. Our work did not explore the prevalence of SDH in the mental health and substance use disorder patient population; however, future research should explore the prevalence of SDH in EHR documentation.

In Chapter 3, we describe development of a neural network model, bidirectional-LSTM, that performed the highest across all evaluation metrics in the task of classifying clinical note sentences for SDH (average precision-recall= 0.76). Lack of neural network transparency or interpretability continues undermine health professionals' willingness to accept machine recommendations without clarity regarding the underlying rationale[21]. Gradient decision tree algorithms provide more transparency[22,23] and performed well across traditional evaluation metrics precision (0.85), recall (0.78), and micro-averaged F1 (0.82) across al SDH labels. Our model outperformed (F1; 0.89-0.43) a similar model published by Feller and colleagues[8]. They developed a multi-class gradient boosting tree to classify SDH sexual risk factors with a range of F1 scores (e.g. F1= 79.2 for LGBT status; F1= 27.3 for intravenous drug abuse). A more accurate measure of our classification models may be precision because our approach of building an SDH rich corpus. Unlike previous studies, we developed a multi-label learning (MLL) model that differs from classical machine learning by tackling the learning problem from a different perspective. In contrast to the classical classification tasks where each observation belongs to only one mutually exclusive class,

in MLL decision areas of labels (i.e. classes) overlap[24]. The SDH classification tasks appears to be similar to large-scale multiple phenotyping, such as diagnosis code assignment, that are cast as a multi-label classification over a large label set[25,26]. In our study we attempted to classify financial resource strain and poor social support, creating 6 labels (Chapter 2). However, the Institute of Medicine recommends (Table 1) an additional fifteen (i.e. education status) SDH domains. Future work should explore development of models that classify all recommended SDH domains, potentially representing an extreme multi-label classification problem[26,27]. A limitation of MLL models is the lack of transparency and interpretability of current open source machine learning libraries that could hinder acceptance by health professionals. We believe this is computational problem is addressable in future work. Our work did not address changes in a patients' status, an advantage of classifying on the sentence-level. For example, housing insecurity can be a fluid domain[19,20] for a patient who may gain or lose housing over a period of time. Future research should explore the longitudinal documentation of SDH to determine long-term trends and if there are SDH domains where predictions are ordered by time, such as forecasting SDH.

In Chapter 4, we demonstrated that simpler, lower cost text mining techniques (i.e. auto-encoding) can perform as well as more complex NLP approaches (i.e. rule and machine learning based) when classifying SDH characteristics in clinical notes. We applied these approaches on a single institution data set provided by UNC's Clinical Data Warehouse and open source clinical notes provided by the n2c2 NLP challenges[28]. To our knowledge, this is the first study to attempt to identify SDH characteristics in open source clinical notes. High performance was achieved using auto-encoding (F1=0.83) and the machine learning based algorithms SVM (F1=0.86) and XGBoost (F1=0.86). These results may be due to the small sample size; we identified only 256 positive SDH characteristics among 2,797 clinical notes representing 814 patients. For example, we found only one food insecurity representation. Low occurrence of SDH characteristics in the n2c2 data set reduced our ability to fully compare results from models derived using the UNC data source. These findings highlight the importance

of unstructured data in health care organizational activities and population health services research. Vest and colleagues found that organizations relying solely on structured data (e.g. procedure codes and appointments), will likely underestimate needed services that directly address SDH[7]. Future work would benefit from a secondary larger corpus for generalizability comparison. Our work suggests that NLP model develop may not require traditional, expensive gold standard annotation as our SDH dictionary auto-encoding performed well for both UNC (F1=0.76) and n2c2 (F1=0.83) data sources. This would greatly reduce the cost, time, and NLP experts required for customized algorithm development as opposed to SDH dictionary auto-encoding. While the auto-encoding approach had slightly lower performance, Lundberg and colleagues[29] found that physicians using machine learning models in clinical practice preferred simpler models that were more interpretable at the expense of lower accuracy. Leveraging our existing SDH dictionary would be far cheaper than investing the estimated $2,760 to develop a custom dictionary and addresses health providers' transparency and interpretability concerns. Future work should explore acceptance tested among health professionals based on interpretability and performance.

We recognize that all three studies described in this dissertation suffered from the use of a single data source: clinical note text. Feller and colleagues[8] found that using both structured (i.e diagnosis, procedures, and laboratory tests) and unstructured (i.e. clinical notes) EHR data yielded better performance than either data source alone when identifying social sexual risk factors. However, the higher performance in models using both data structures was not statistically significant. Further research should explore structured EHR data and integration of population-level socioeconomic data variables provided by the U.S. Census into SDH model development or applications[30]. Current research in SDH identification using machine learning is limited due to the unknown quality of SDH documentation. Future research should explore the consistency between a patients SDH clinical note documentation and NIH approved SDH screeners. A significant hurdle to any future research using EHR clinical notes is the lack of standardization surrounding the collection and storing of clinical notes. Best practices should be

developed on how to store clinical notes for translational science efforts as a large source of our time and labor was attributed to cleaning excess noise and deciphering blank auto-generated text. Additionally, the current storage system did not allow us to track the type of note or health professional documenting SDH characteristics. Cumbersome storage systems will continue to hinder efforts that aim to leverage clinical notes.

We found very little systematic documentation of patients' SDH data in the EHRs clinical notes including a large amount of incomplete or empty SDH screening surveys. Experts recommend limiting SDH screening to a subset of patients and enabling EHR-based SDH data tools to target this subset to avoid overwhelming or burdening health professionals[3,31]. The SDH classification models we developed may be best applied as a tool to identify patients requiring standardized screening for SDH such as food and housing insecurity. EDs are beginning to take ownership of SDH for their patients with recent examples of successful SDH interventions focused on the development of coordinated care models and partnerships with local resources[32,33]. For example, emergency medicine researchers worked with the Housing First partnership between the Center for Medicare and Medicaid Services and New York City, which provided housing for high-risk homeless patients, resulting in improved health and cost savings for the city[34]. The SDH identification approaches evaluated in our studies are a step towards improving SDH data accessibility and standardization for linking SDH interventions through the ED. Better SDH data collection may also be a useful tool when approaching policy makers for decisions regarding social services resources, affordable housing, and access to affordable health care[4,35,36].

Our work shows many implications for SDH collection in the EHR. Future research from this project may lead to an SDH screening alert within the EHR to increase adoption rates. In this task recall may be prioritized to limit alert fatigue and assure that all those who need a screening, receive one. Gold and colleagues found that health professionals did not want to collect SDH data themselves, preferring to

transfer the responsibility to another team member[3]. With SDH data collected via multiple routes and certain SDH data are already collected regularly by specific health professionals (e.g. social workers), future research should explore a need for an EHR-based summary that contains all of a patient's SDH data. When adequately leveraged, electronic platforms improve integration between medical and social service delivery. EHRs could provide opportunities to improve the evidence by improving data accessibility and standardization, linking SDH interventions with health outcomes, and supporting the examination of individual and population-level data.

The body of evidence demonstrating the link between SDH and health and illness grows, health provider access to SDH data becomes critical to both in-the-moment care decision making and broader policy and resource planning. Our findings demonstrate the feasibility of identifying the SDH characteristics financial resource strain and poor social support from EHR clinical notes. Our approach of developing an SDH dictionary using a SME-driven word embedding expansion model derived on a rich SDH population produced a high quality dataset used for gold-standard corpus curation and ML model development. Unlike previous work, we evaluated our models using sentence-level data that contained multiple instances of SDH documentation thus ensuring our models could be used for real-world SDH clinical decision support tasks. This work is a step towards developing a clear process for identifying and classifying data on SDH that are important for next step toward transforming health care decision making, refining value-based payments, and ultimately influencing healthcare and policy makers to improve population health. Using these recommendations, healthcare settings may create opportunities to integrate evidence-based SDH metrics systematically into clinical care processes, including functions related to social screening; triaging social needs; making referrals; tracking individual and population-level data; and sharing tracked data. Enhancing the integration of social services and medical care, and improve health outcomes for individuals and communities.

TABLE 1: Institute of Medicine recommendations for inclusion of SDH

Individual Factors	**Sociodemographic**	**Sexual orientation**
		Racial identity
		Ethnic identity
		Country of origin/migration history
		Education
		Employment
		Financial resource strain food and housing insecurity
		Health literacy
	Psychological	Depression/anxiety
		Stress
		Optimism/Self-efficacy/Patient empowerment/activation/engagement
	Behavioral	Dietary patterns
		Physical activity
		Tobacco use and exposure
		Alcohol use
	Individual-level social relationships and living conditions	Social isolation and social connections
		Exposure to violence
Neighborhoods/ Communities	Compositional characteristics	Neighborhood and community compositional characteristics

REFERENCES

1. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing social and behavioral domains in electronic health records: phase 1*. Washington (DC): National Academies Press (US); 2014. doi:10.17226/18709

2. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing social and behavioral domains and measures in electronic health records: phase 2*. Washington (DC): National Academies Press (US); 2015. doi:10.17226/18951

3. Gold R, Cottrell E, Bunce A, et al. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *J Am Board Fam Med*. 2017;30(4):428-447. doi:10.3122/jabfm.2017.04.170046

4. Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. *Health Aff (Millwood)*. 2018;37(4):585-590. doi:10.1377/hlthaff.2017.1252

5. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. *JAMIA Open*. 2019;2(1):81-88. doi:10.1093/jamiaopen/ooy051

6. Hatef E, Rouhizadeh M, Tia I, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system. *JMIR Med Inform*. 2019;7(3):e13802. doi:10.2196/13802

7. Vest JR, Grannis SJ, Haut DP, Halverson PK, Menachemi N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int J Med Inform*. 2017;107:101-106. doi:10.1016/j.ijmedinf.2017.09.008

8. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data. *Appl Clin Inform*. 2020;11(1):172-181. doi:10.1055/s-0040-1702214

9. Fan Y, Pakhomov S, McEwan R, Zhao W, Lindemann E, Zhang R. Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA Open*. 2019;2(2):246-253. doi:10.1093/jamiaopen/ooz007

10. Ye C, Fabbri D. Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. *J Biomed Inform*. 2018;83:63-72. doi:10.1016/j.jbi.2018.05.014

11. Maimon O, Rokach L, eds. *Data Mining and Knowledge Discovery Handbook (Springer series in solid-state sciences)*.

12. Monsen KA, Rudenick JM, Kapinos N, Warmbold K, McMahon SK, Schorr EN. Documentation of social determinants in electronic health records with and without standardized terminologies: A comparative study. *Proceedings of Singapore Healthcare*. 2018;28(1):201010581878564. doi:10.1177/2010105818785641

13. Shivade C, Malewadkar P, Fosler-Lussier E, Lai AM. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. *J Biomed Inform*. 2015;58 Suppl:S103-10. doi:10.1016/j.jbi.2015.08.025

14. Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform*. 2018;87:12-20. doi:10.1016/j.jbi.2018.09.008

15. Zhang Y, Li H-J, Wang J, Cohen T, Roberts K, Xu H. Adapting Word Embeddings from Multiple Domains to Symptom Recognition from Psychiatric Notes. *AMIA Jt Summits Transl Sci Proc*. 2018;2017:281-289.

16. Bejan CA, Angiolillo J, Conway D, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc*. 2018;25(1):61-71. doi:10.1093/jamia/ocx059

17. Carrell DS, Cronkite D, Palmer RE, et al. Using natural language processing to identify problem usage of prescription opioids. *Int J Med Inform*. 2015;84(12):1057-1064. doi:10.1016/j.ijmedinf.2015.09.002

18. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *J Biomed Inform*. 2015;55:290-300. doi:10.1016/j.jbi.2015.05.003

19. Doran KM, Kunzler NM, Mijanovich T, et al. Homelessness and other social determinants of health among emergency department patients. *J Soc Distress Homeless*. 2016;25(2):71-77. doi:10.1080/10530789.2016.1237699

20. Ku BS, Fields JM, Santana A, Wasserman D, Borman L, Scott KC. The urban homeless: super-users of the emergency department. *Popul Health Manag*. 2014;17(6):366-371. doi:10.1089/pop.2013.0118

21. Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLoS One*. 2019;14(2):e0212356. doi:10.1371/journal.pone.0212356

22. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD' ' '16*. New York, New York, USA: ACM Press; 2016:785-794. doi:10.1145/2939672.2939785

23. Liu Y, Gu Y, Nguyen JC, et al. Symptom severity classification with gradient tree boosting. *J Biomed Inform*. 2017;75S:S105-S111. doi:10.1016/j.jbi.2017.05.015

24. Zufferey D, Hofer T, Hennebert J, Schumacher M, Ingold R, Bromuri S. Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Comput Biol Med*. 2015;65:34-43. doi:10.1016/j.compbiomed.2015.07.017

25. Gangavarapu T, Jayasimha A, Krishnan GS, S. SK. Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowledge-Based Systems*. 2020;190:105321. doi:10.1016/j.knosys.2019.105321

26. Baumel T. Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment.

27. Zhang M-L, Zhou Z-H. A Review on Multi-Label Learning Algorithms. *IEEE Trans Knowl Data Eng*. 2014;26(8):1819-1837. doi:10.1109/TKDE.2013.39

28. n2c2 NLP Research Data Sets. https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/. Accessed July 10, 2020.

29. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-760. doi:10.1038/s41551-018-0304-0

30. Bazemore AW, Cottrell EK, Gold R, et al. Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. *J Am Med Inform Assoc*. 2016;23(2):407-412. doi:10.1093/jamia/ocv088

31. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Med*. 2015;48(2):215-218. doi:10.1016/j.amepre.2014.07.009

32. Kahan D, Leszcz M, O'Campo P, et al. Integrating care for frequent users of emergency departments: implementation evaluation of a brief multi-organizational intensive case management intervention. *BMC Health Serv Res*. 2016;16:156. doi:10.1186/s12913-016-1407-5

33. Soril LJJ, Leggett LE, Lorenzetti DL, Noseworthy TW, Clement FM. Reducing frequent visits to the emergency department: a systematic review of interventions. *PLoS One*. 2015;10(4):e0123660. doi:10.1371/journal.pone.0123660

34. Anderson ES, Lippert S, Newberry J, Bernstein E, Alter HJ, Wang NE. Addressing Social Determinants of Health from the Emergency Department through Social Emergency Medicine. *West J Emerg Med*. 2016;17(4):487-489. doi:10.5811/westjem.2016.5.30240

35. Chen C, Weider K, Konopka K, Danis M. Incorporation of socioeconomic status indicators into policies for the meaningful use of electronic health records. *J Health Care Poor Underserved*. 2014;25(1):1-16. doi:10.1353/hpu.2014.0040

36. Gottlieb L, Sandel M, Adler NE. Collecting and applying data on social determinants of health in health care settings. *JAMA Intern Med*. 2013;173(11):1017-1020. doi:10.1001/jamainternmed.2013.560