Towards the Inference of Social and Behavioral Determinants of Sexual Health: Development of a Gold-Standard Corpus with Semi-Supervised Learning

Daniel J. Feller, MA, Jason Zucker, MD, Oliver Bear Don't Walk IV, MS, Bharat Srikishan MS, Roxana Martinez, Henry Evans, Michael T Yin, MD, Peter Gordon, MD, Noémie Elhadad, PhD
Columbia University, New York, NY, USA

Abstract

Social and behavioral determinants of health (SBDH) are environmental and behavioral factors that are increasingly recognized for their impact on health outcomes. We describe ongoing research to extract SBDH related to sexual health from clinical documentation. Our work addresses several challenges. First, there is no standard set of SBDHs for sexual health; we describe our curation of 38 such SBDHs. Second, it is unknown how SBDH related to sexual health are expressed in clinical notes; we detail the characteristics of an annotated corpus. Third, SBDH documentations are rare; we describe the use of semi-supervised learning to accelerate the annotation process by identifying notes likely to document SBDH. Fourth, we describe preliminary results to infer an array of SBDH from clinical documentation using supervised learning.

Introduction

Social and behavioral determinants of health (SBDH) are behavioral, environmental, and community factors that have been implicated in an array of adverse health outcomes. For instance, environmental factors such as food insecurity and neighborhood poverty are associated with adverse health outcomes, while housing instability and unmanaged substance use disorders can impede the delivery and efficacy of care and result in nonadherence to medication.[1-5] Because knowledge of these and other SBDH for specific patients is clinically meaningful and can lead to tailored care plans, accounting for SBDH has become increasingly recognized in healthcare delivery.[6] As a result, there has been increasing momentum for incorporating social determinants into the electronic health record (EHR). For example, the Meaningful Use program incentivized the collection of a range SBDH including tobacco and alcohol use, intimate partner violence, social isolation and support, and physical activity and has catalyzed the development of standardized item sets for some social determinants.

However, social and behavioral determinants of sexual health have received comparatively less attention than other SBDH. High-risk sexual activity like infrequent condom use and receptive anal intercourse is associated with increased risk for human immunodeficiency virus (HIV) and other sexually transmitted infections (STIs). Transgender persons are also at a higher risk of acquiring STIs and less likely to achieve desirable chronic-disease outcomes such as HIV viral load suppression.[7,8] Moreover, sexual orientation may play a role in STI risk. At present, sexual health SBDH are not collected in SBDH screening instruments such as the PRAPARE or the Accountable Health Communities Screening (AHCS) tool, nor routinely documented in the structured part of EHRs.[9,10] This challenges the development of clinical decision support systems and population-level interventions to reduce the prevalence of HIV and other STIs.[11,12]

While not documented in a structured fashion, sexual health-related SBDH are discussed in the narrative part of the EHR.[13,14] However, there has been limited research exploring the automated extraction of SBDH related to sexual health from clinical notes, for instance through natural language processing (NLP) methods. In this paper, we address several challenges related to this task; First, there is no standard set of determinants of sexual health; we describe an expert curation of 38 such SBDHs. Second, we detail the challenge of creating such a corpus and report its high-level characteristics. Third, SBDH documentation are rare; we describe how semi-supervised learning accelerated the annotation process by recognizing clinical notes likely to contain SBDH content. Fourth, we describe preliminary results to use supervised learning to infer an array of SBDH risk factors from clinical documentation.

Related Work

There is growing interest in integrating SBDH data into EHRs.[15,16] In 2015, the National Academy of Medicine recommended a set of 12 SBDH indicators, part of the Meaningful Use Program. The Office of the National Coordinator for Health Information Technology (ONC) recently drafted the 2018 guidelines for collection of 8 of the

12 aforementioned indicators, all of which are mapped to standard LOINC codes.[17] While the widespread adoption of these recommendations would achieve the important goal of moving critical elements of the social history into structured elements of the EHR, the selection of a concise set of SBDH indicators inevitably neglected an array of important social and behavioral factors.[15] To our knowledge, there has been no discussion of incorporating elements of sexual history into EHRs through standardized value sets.

It is unclear how SBDH related to sexual health are expressed in longitudinal patient records. Walsh and Elhadad used topic modeling to characterize the content of social history sections and observed more language relevant to sexual history in outpatient notes compared to inpatient notes.[14] Chen and colleagues reviewed clinical notes across 3 health systems and found that sexual activity and sexual orientation were infrequently documented; combined these topics were mentioned with less frequency than caffeine use.[13] Simple textual searches for sexual orientation (e.g., 'LGBT', 'gay', 'lesbian') identified several thousand records at Vanderbilt University Medical Center.[18] While these studies suggest that information related to sexual health is sparsely documented in EHR data, they reported neither the prevalence nor lexical characteristics of specific SBDH related to sexual health in clinical notes.

Several studies have successfully used NLP to infer social and behavioral determinants of health from EHR data. These systems have modeled clinical language using a range of approaches from regular expressions to named entity recognition to topic modeling and contemporary distributional semantics.[19-22] It is worth noting that out-of-the-box NLP tools such as MetaMap and cTakes may be expected to have diminished performance on extracting SBDH because the language used to express these concepts is often regional and idiosyncratic.[19] Overall, there are still several research gaps when it comes to automated identification of SBDH related to sexual health: 1) only a few of the pertinent SBDHs have been examined in the literature; 2) they have diverse lexical realizations in clinical notes and for many, there might not even be good coverage in existing terminologies; and 3) documentation of SBDH are rare across patients (despite their acknowledged importance).

Methods

Social and Behavioral Determinants Relevant to Sexual Health

Three physicians (JZ, PG, MY) experienced in the prevention and treatment of STIs reviewed the biomedical literature and identified six high-level domains of behavioral risk factors for STIs (gender, sexual orientation, sexual activity, drug use, alcohol use, and homelessness), along with 32 individual-level SBDH (**Table 1**). These SBDH represent independent risk factors for the acquisition of STIs and have recently been the focus of targeted HIV prevention efforts.

The domain experts recommended that the six SBDH domain labels should indicate whether information related to a candidate risk domain was documented. For example, *"patient denies sexual activity"* and *"patient is sexually active"* would both result in a positive label for *the "Sexual Activity"* domain. It was hypothesized that these domain-level labels can be used to inform efforts directed at improving social and sexual history taking by clinicians.

Curation of a Gold-Standard Corpus of Annotated Clinical Notes for Sexual Health SBDH

Development of Annotation Guidelines for Sexual Health SBDH

We chose to obtain annotations at the document-level rather than mention-level because of how social and behavioral determinants are expressed in natural language. Unlike many other biomedical concepts, SBDH are infrequently expressed as named entities. For example, we observed the following examples in clinical notes at CUIMC:

> *"has continued to relapse on crack and beer since starting treatment 3 months ago"*
> *"noted that he used occasional social EtOH (scotch) at church functions"*
> *"Transmitted via heterosexual intercourse"*
> *"3-4 lifetime male unprotected sexual partners"*
> *"HIV/AIDS (Dx 1992 after unprotected sex, RF sex with both men and women"*

In addition, we also reasoned that obtaining document-level annotations would be less labor-intensive than obtaining mention-level annotations, an important consideration given how infrequently SBDH are mentioned in patient records.

We excluded from consideration relational modifiers such as amount (e.g., "*5 sexual partners"*), frequency of exposure (e.g., "*once a month"*), and status (e.g., "*current/past/none"*) because this information is not prioritized by clinical interventions addressing sexual health. However, document-level annotation does not preclude the capture of

relevant modifiers such as the frequency of condom use (e.g., *"condom never"*, *"condom sometimes"*, *"condom always"*). Annotators were instructed to review the entire length of each clinical note.

Table 1. Summary of annotation guideline for 6 domain-level and 32 individual-level SBDH indicators. Domain-level indicators are indicated in bold.

SBDH Indicator	Example	SBDH Indicator	Example
Gender Documented	*56 yr. old male*	Transgender Male	*trans male, trans FtoM*
Male	*male, man, boy*	Transgender Female	*trans female, trans MtoF*
Female	*female, woman, girl*	Non-Conforming	*non-conforming*
Sexual Orientation Documented	*pt. is heterosexual*	Bisexual	*male and female partners*
MSW	*has a wife*	WSM	*heterosexual female*
MSM	*gay male, LGBT male*	WSW	*LGBT female*
Sexual History Documented	*pt. not sexually active*		
History of STIs	*history of GC/CT*	Condom Always	*consistent condom use*
Oral Sex	*reports oral sex*	Condom Sometimes	*infrequent, occasional*
Vaginal Sex	reports vaginal sex	Condom Never	*pt. doesn't use condoms*
Insertive Anal Intercourse	*insertive anal intercourse*	Receptive Anal Intercourse	*receptive anal intercourse*
Alcohol Use Documented	*pt. denies alcohol use*	Social alcohol use	*pt. drinks occasionally*
Active alcohol use	*currently uses alcohol*	Alcoholism	*pt. drinks frequently*
Substance Use Documented	*pt. denies substance use*	Cocaine	*pt. reports cocaine use*
History of Drug Use	*used cocaine in the past*	Methamphetamine	*reports meth use*
Active Drug Use	*pt. reports cocaine use*	Intravenous Drug Use	*uses intravenous drugs*
Marijuana	*pt. uses cannabis*	Cocaine	*pt. reports cocaine use*
Housing Status Documented	*pt. lives alone*		
Homeless	*pt. lives on the street*	Stable housing	*lives in apartment*
Living with family/friends	*lives on friend's couch*	Unstable housing	*reports unstable housing*

Collection of Clinical Notes for Manual Annotation

A corpus of clinical notes was obtained from the clinical data warehouse at Columbia University Irving Medical Center (CUIMC), a large academic medical center in New York City. For this study, we obtained all individual notes types associated with 4,000 HIV+ individuals within the commercial EHR system at CUIMC (e.g., Admission Note, Progress Note). Additional details on this cohort are described elsewhere.[11] The study described herein was approved by the Institutional Review Board at CUIMC.

Systematic Annotation of Clinical Notes with Curated SBDH Labels

We recruited one Infectious Disease fellow and two medical students to manually review clinical notes for the presence of the 38 SBDH labels. To train for annotation, an initial set of three longitudinal patient records were coded by all three annotators; all discrepant labels were discussed and resolved by consensus to create a shared understanding of the SBDH concepts. The annotators utilized the annotation guidelines and iteratively improved on them during that phase.

Subsequent to the initial training, annotators were instructed to systematically review all notes associated with each patient's record. After annotating >1,000 notes, we observed that many notes rarely contained SBDH mentions (e.g. "Nursing Progress Note"). The annotators then isolated their review to select note types written by healthcare providers related to hospital admission and discharge, outpatient care, and psychiatry consultations.

Semi-Supervised Learning for Expansion of Annotated Corpus

The scarcity of clinical notes containing explicit SBDH mentions rendered the annotation process described above extremely labor-intensive. The excessive amount of human effort required to compile a sizeable annotated corpus exists as a barrier to the use of deep neural networks, which achieve state-of-the-art performance on a range of NLP

tasks.[23] As a result, we employed a technique leveraging modern distributional semantic techniques to accelerate the manual annotation process by identifying clinical documents likely to mention SBDH.

Training Word Embeddings for SBDH

We used the popular GloVe software package to train vectors representations of words. We isolated the social history section of 343,322 randomly selected notes from within our overall clinical data warehouse (343,322 patients). We extracted Admission, Nursing Adult Admission History, and Ambulatory Aim Primary Provider notes because these notes often contained relevant SBDH information and had a Social History section. Prior to training, clinical notes were preprocessed by removing non-alphanumeric characters, replacing numbers with a special token, and converting all tokens to lowercase.

We evaluated several different configurations of GloVe by using word vectors of 50, 150, and 300 dimensions while using a window size of 8 words and 50 iterations of training. 50-dimensional vectors were chosen because they had similar performance compared to 150 and 300-dimensional vectors with lower complexity; performance was assessed by evaluation precision at 10 on a held-out testing set of 20% of the notes with 0 patients overlapping between the testing and training sets. The isolated corpus yielded 33,206,266 tokens and 2,37,072 unique words. Removal of words with fewer than 5 mentions in the corpus were excluded to generate a vocabulary size of 47,479.

Creation of Centroid Document for each SBDH Label

We then sought to retrieve and rank unannotated clinical notes with regard to similarity to annotated notes with explicit mentions of our target SBDH. This required the generation of a single centroid document for each SBDH label. For each SBDH label, we isolated the social history sections of all notes with valid mentions of the specific SBDH and represented each document itself as a 50-dimensional vector by averaging the word embeddings associated with the document. The singular centroid for each SBDH was generated by again averaging the document vectors associated with the SBDH. We did not construct a centroid for gender-related documentation.

Identification of Unannotated Notes Likely to Contain SBDH Documentation

We then isolated 144,432 clinical notes associated with HIV+ individuals obtained from the enterprise data warehouse at Columbia University Irving Medical Center. Each note was represented as a 50-dimensional vector by averaging the individual word embeddings of each note as described above. For each SBDH label, cosine similarity was used to rank unannotated notes in order of descending cosine distance to the corresponding SBDH centroid document. This resulted in a ranking of notes in regards to cosine distance from one or more document centroids. Notes identified using this methodology are hereafter referred to as candidate SBDH notes.

Annotation of Candidate Clinical Notes

Following the annotation guidelines, two annotators subsequently annotated candidate SBDH notes corresponding to 5 domain-level SBDH and observed high interrater reliability; a Kappa statistic of 0.598 was observed across all SBDH labels.

Supervised Learning of SBDH Labels

Feature Engineering

Lexical Features: Each clinical note was represented as a vector of term frequency-inverse document frequency (tf-idf) weights. Notes were preprocessed by removing numbers, punctuation, and stop words. This yielded approximately 13,000 unique lexical features for use in classification.

Ontological Features: We compiled a list of 216 concepts from the UMLS such as 'homosexual male' (C0870597) and 'intravenous drug abuse' (C0086181). Each clinical note was analyzed using our in-house Named Entity Recognition system which performs shallow syntactic analysis to detect biomedical concepts within the Unified Medical Language System (UMLS).[24,25] The system identified 2,886 SBDH concepts in 805 annotated notes (20% of total). All identified UMLS concepts were analyzed for negation; we used the NegEx algorithm which implements regular expressions to identify negation phrases.[26] A large proportion of the notes in our corpus were 'semi-structured' and prompted physicians to respond to templates with yes or no answers (*"Does patient have history of unsafe sex: "*) As a result, we enriched the standard set rules available with the publicly available NegEx.

Multi-label Classification of Clinical Notes with SBDH

We trained one classifier per SBDH label. This approach has been previously used within the informatics community for clinical diagnosis, ICD code assignment, and semantic indexing.[27-29] We used chi-squared goodness of fit tests to perform feature selection. 200 features that had the strongest univariate association with each SBDH label were utilized in training a Support Vector Machine classifier (SVM) for each label. Each SBDH classifier consisted of a single SVM with a radial basis function kernel with a coefficient of 0.002. The Python library *scikit-learn* v0.19.1 was used for model development and evaluation.[30] Note that we did not train a classifier on the presence of gender-related documentation.

To derive an unbiased estimate of model performance, we ensured that no individual patient contributed notes to both training and testing data for each SBDH classifier. The frequency of redundant text in clinical notes presents a number of challenges for training and evaluating text mining models.[31] Any evaluation that did not stratify the training and testing data by MRN would likely artificially inflate performance estimates, due to the frequency of copy-and-pasted text in longitudinal patient records.[31]

Results

Systematic Annotation of Clinical Notes for Sexual Health SBDH

Three annotators reviewed every clinical note associated with 32 randomly selected HIV+ individuals to detect the presence of all SBDH labels. 76 notes were double annotated (Kappa 0.736 across all SBDH). 3883 clinical notes were manually annotated and 17.9% (695) had one or more SBDH label. On average, there were 0.83 SBDH mentions per annotated note. In addition, we observed a high frequency of redundant text, reflecting the widespread usage of copy-and-paste at CUIMC.

Semi-supervised learning to Identify Notes Likely to Contain Sexual Health SBDH Documentation

We isolated 10 clinical notes within the lowest cosine distance compared to 5 'domain-level' centroids and computed precision-at-10. Our approached yielded the following precision-at-10 results: "*housing status documented" (100%), "alcohol documented" (90%), "substance use documented" (90%), "sexuality activity documented" (60%),* and *"sexual orientation documented" (60%)*. Averaged across these, we observed a precision-at-10 of 80%.

118 notes with high similarity to 1 or more SBDH centroids were annotated and 113 notes (95.7%) contained 1 or more SBDH mentions. On average, there were 8.26 SBDH mentions per note. This represents an astounding 10-fold increase in the yield of positive SBDH mentions compared to the systematic review (0.83 vs. 8.26 SBDH per note). In addition, the 118 notes were associated with 80 HIV+ patients, achieving our goal of diversifying the annotated corpus and reducing the frequency of redundant text.

Characteristics of Annotated Clinical Notes

While our annotation process is underway, we thus far have amassed a corpus of 4,045 annotated clinical notes, overall associated with 105 HIV+ individuals. 19.0% of notes (770) contained documentation of 1 or more of the six SBDH domains. Among domains, alcohol use was documented most frequently (439 notes), followed by substance use (422), housing status (335), sexual activity (326), and sexual orientation (259). No mention of patients who were transgender or gender non-conforming was observed.

In the annotated corpus, we observed 2,466 mentions of individual SBDH. *"Marijuana use"* was the most frequent (188 notes), followed by "living with friends" (171 notes), and "alcohol abuse" (143 notes). 99 notes documented patients as LGBT, with 73, 22, and 9 "MSM", "Bisexual", and "WSW", respectively. 12 SBDH labels were each documented in fewer than 25 notes.

We observed a significant among of dependence among individual SBDH labels, displayed in **Figure 1**. SBDH related to active substance use such as methamphetamine and cocaine use (correlation coefficient = 0.63) often displayed a strong correlation, as did active alcohol use and a history of substance

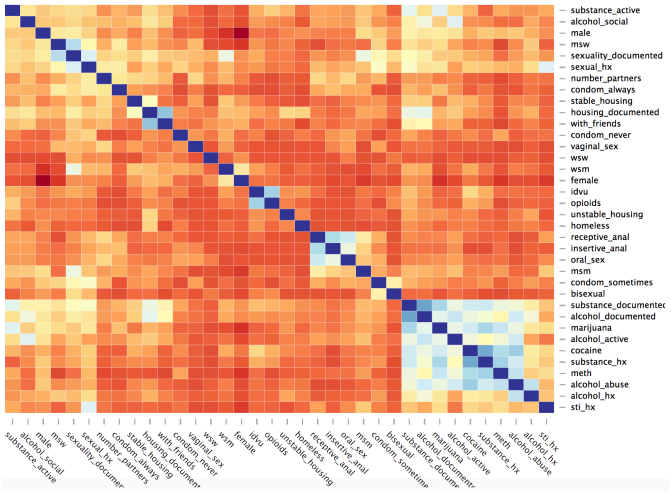

Figure 1. *Label Dependence among SBDH indicators.*

abuse (0.39). MSM and receptive anal intercourse (0.35) also displayed a correlation. However, the correlation matrix suggests that a considerable number of SBDH exhibit little association with other labels.

Supervised Learning of SBDH Domains

The annotated corpus (4,045 notes) was split into a training and testing set stratified by patient. The documents were represented as a vector of features which included 41 ontological and 29,284 lexical features.

Table 2. Performance of SVC classifiers on documentation. 75/25 training/testing split, stratified by individual patients.

Topic Documented	# Notes	Precision	Recall	F1
Sexual Orientation	270	0.66	0.11	0.19
Sexual History	349	0.72	0.26	0.38
Alcohol Use	463	0.80	0.34	0.48
Substance Use	450	0.84	0.29	0.44
Housing Status	345	0.83	0.35	0.49

The document classifiers trained on domain-level labels such as *"housing documented"* or *"sexuality documented"* demonstrated acceptable performance (**Table 2**). In general, the domain-labels with the most annotated instances yielded the best result; the *"alcohol use documented"* classifier achieved an F1-score of 0.89 with 463 instances while *"sexual orientation documented"* had 270 instances and achieved an F1-score of just 0.19.

We experimented with Support Vector classifiers on the 15 most common individual-level SBDH labels. Our naïve approach to multi-label classification (binary relevance) yielded poor results. While classifiers trained on high-prevalence SBDH labels like *"history of STIs"* (139 mentions) demonstrated modest discriminative ability (Precision 0.5, Recall 0.14), those trained on less prevalent SBDH labels such as *"intravenous drug use"* had little or no prognostic ability.

Discussion

We posit that our schema of SBDH related to sexual health can serve as a common foundation on which to build data collection and analysis efforts. Our set of 32 individual-level SBDH constitute an array of behaviors that have been associated with an increased likelihood of acquiring syphilis, HIV, and other STIs.[32] In addition, our six domain-level SBDH indicators may be used to assess the quality of clinical documentation. The annotation guidelines can inform future efforts in corpus curation and computational methods to infer determinants of sexual health from EHRs (available at github.com/danieljfeller/SBDSH).

We describe the first annotation of clinical notes for social and behavioral determinants of sexual health. Overall, the SBDH domains were observed infrequently in clinical notes; alcohol and substance use were the most prevalent individual-level SBDH in our corpus but were observed in only 4% of annotated notes. Sexual orientation was documented in less than 1% of notes. SBDH documentation occurred mostly in outpatient notes, admission notes, and discharge summaries. We are also the first to observe inter-SBDH correlation; several individual-level SBDH displayed moderate associations with other labels (e.g., *"cocaine use"* and *"alcohol use"*). This label dependence could potentially be leveraged to improve their automated extraction.

Because SBDH documentation is so rare and the requirement of a large gold-standard corpus to learn from, our approach necessitated the use of computational methods to identify clinical notes likely to contain SBDH content. Our semi-supervised approach using similarity based on section-embeddings successfully increased the yield of manual annotation. Annotators observed 8.26 distinct SBDH mentions per note for the 60 notes closest to the 6 SBDH domain centroids, compared to 0.83 mentions per note randomly sampled from a cohort of HIV+ individuals. The utility of distributional semantics techniques for modeling the diverse lexical realizations of SBDH in notes has been established.[19] The success of our approach will allow our research team to increase the size and diversity of our annotated corpus. Further, this approach also enables the diversity of our corpus in patients and types of lexical realizations of SBDH, thereby ensuring the extensibility of future SBDH models to various patients and healthcare settings.

The manual annotation confirmed our hypothesis that there is a wide variation in lexical realizations of SBDH, ranging from word to multi-word expressions to whole sentences. As such our approach to treating annotation as document-level labeling circumvented this phenomenon. We experimented with using supervised learning to infer SBDH labels for given clinical notes. The inability to infer the presence of individual-level SBDH likely reflects the limited size of our annotated corpus, as compared to the training size typical of other document classification systems for medical concept recognition.[33,34] Moreover, the poor classifier performance may reflect the UMLS's lack of coverage for SBDH;

62% of annotated notes with explicit mentions of SBDH were tagged with no relevant UMLS concepts by our in-house named-entity recognition system. Only 38% of annotated notes with explicit mentions of SBDH were tagged with 1 or more relevant UMLS concepts. There are a number of ways to improve on our experiments: **(1)** multi-label classification may be improved by accounting for the observed structure of SBDH labels. Hierarchically structured sets of SVM have demonstrated improved performance compared to binary relevance for multi-label classification of clinical documents.[27,29,35,36] In addition, it may be possible to leverage the SBDH label dependence exhibited in **Figure 1**;[37] **(2)** document-level SBDH labeling may benefit from document zoning.[38] Long documents like clinical notes typically contain many words unrelated to the modeling task; in clinical documentation this is manifested by sections (ie. 'Review of Systems') potentially irrelevant to SBDH; **(3)** structured elements of the EHR such as laboratory tests and diagnosis codes can improve the inference of social determinants compared to using notes alone. Future studies should examine whether laboratory tests for STIs hold prognostic value; and **(4)** with a larger annotated corpus, a neural network with attention layer that could provide transparency for classification decisions, may improve results.

Our study has several limitations. First, document-level annotations lack the granularity of mention-level annotations and thus systems trained on such data may be inappropriate for some informatics interventions.[20] Second, our semi-supervised learning approach relied on notes which contained a social history section; not all notes do so. Third, a relatively small sample size was used for evaluation of the semi-supervised learning approach.

Conclusion

We describe a set of social and behavioral determinants related to sexual health and report on the curation of a gold-standard corpus of clinical notes documenting such determinants. Our findings demonstrate that while these SBDH are infrequently documented in clinical notes, semi-supervised learning can reduce the burden of manual annotation. In addition, our experiments with supervised learning suggest that existing lexical resources may be inadequate for extracting SBDH.

References

1. Gundersen, C. & Ziliak, J. P. Food Insecurity And Health Outcomes. Health Aff. Proj. Hope 34, 1830–1839 (2015).
2. Leigh-Hunt, N. et al. An overview of systematic reviews on the public health consequences of social isolation and loneliness. Public Health 152, 157–171 (2017).
3. Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1. Institute of Medicine Available at: http://www.nationalacademies.org/hmd/Reports/2014/Capturing-Social-and-Behavioral-Domains-in-Electronic-Health-Records-Phase-1.aspx. (Accessed: 18th October 2016)
4. Feller, D. J., Akiyama, M. J., Gordon, P. & Agins, B. D. Readmissions in HIV-Infected Inpatients: A Large Cohort Analysis. J. Acquir. Immune Defic. Syndr. 1999 71, 407–412 (2016).
5. Doran, K. M., Misa, E. J. & Shah, N. R. Housing as Health Care — New York's Boundary-Crossing Experiment. N. Engl. J. Med. 369, 2374–2377 (2013).
6. Behforouz, H. L., Drain, P. K. & Rhatigan, J. J. Rethinking the Social History. N. Engl. J. Med. 371, 1277–1279 (2014).
7. Reisner, S. L. et al. Characterizing the HIV Prevention and Care Continua in a Sample of Transgender Youth in the U.S. AIDS Behav. 21, 3312–3327 (2017).
8. Clark, H., Babu, A. S., Wiewel, E. W., Opoku, J. & Crepaz, N. Diagnosed HIV Infection in Transgender Adults and Adolescents: Results from the National HIV Surveillance System, 2009-2014. AIDS Behav. 21, 2774–2783 (2017).
9. PRAPARE. NACHC
10. Billioux, A., Verlander, K., Anthony, S. & Alley, D. The Accountable Health Communities Screening Tool. Discuss. Pap. 9 (2017).
11. Feller, D. J., Zucker, J., Yin, M. T., Gordon, P. & Elhadad, N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. J. Acquir. Immune Defic. Syndr. 1999 (2017). doi:10.1097/QAI.0000000000001580
12. Feller, D. & El Hadad, N. HIV Risk Assessment using Longitudinal Electronic Health Records. in ID Week (2017).
13. Chen, E. S., Manaktala, S., Sarkar, I. N. & Melton, G. B. A Multi-Site Content Analysis of Social History Information in Clinical Notes. AMIA. Annu. Symp. Proc. 2011, 227–236 (2011).
14. Modeling clinical context: rediscovering the social history and evaluating language from the clinic to the wards. (2014).

15. Cantor, M. N. & Thorpe, L. Integrating Data On Social Determinants Of Health Into Electronic Health Records. Health Aff. (Millwood) 37, 585–590 (2018).
16. Gold, R. et al. Developing Electronic Health Record (EHR) Strategies Related to Health Center Patients' Social Determinants of Health. J. Am. Board Fam. Med. 30, 428–447 (2017).
17. 2018 Interoperability Standards Advisory. 113
18. Populations, I. of M. (US) B. on the H. of S. Existing Data Collection Practices in Clinical Settings. (National Academies Press (US), 2013).
19. Bejan, C. A. 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. 25, 61–71 (2018).
20. Yetisgen, M., Pellicer, E., Crosslin, D. & Vanderwende, L. Automatic Identification of Lifestyle and Environmental Factors from Social History in Clinical Text. Microsoft Res. (2016).
21. Gundlapalli, A. V. 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. AMIA Symp. 2013, 537–546 (2013).
22. Oreskovic, N. M., Maniates, J., Weilburg, J. & Choy, G. Optimizing the Use of Electronic Health Records to Identify High-Risk Psychosocial Determinants of Health. JMIR Med. Inform. 5, e25–e25 (2017).
23. Zhang, Y. & Wallace, B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. ArXiv151003820 Cs (2015).
24. Lipsky-Gorman, S. & Elhadad, N. Clin Note and Health Term Finder: A pipeline for processing clinical notes. (Columbia University Technical Report, 2011).
25. Hirsch, J. S. et al. HARVEST, a longitudinal patient record summarizer. J. Am. Med. Inform. Assoc. 22, 263–274 (2015).
26. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Inform. 34, 301–310 (2001).
27. Perotte, A. et al. Diagnosis code assignment: models and evaluation metrics. J. Am. Med. Inform. Assoc. JAMIA 21, 231–237 (2014).
28. Papanikolaou, Y., Tsoumakas, G., Laliotis, M., Markantonatos, N. & Vlahavas, I. Large-scale online semantic indexing of biomedical articles via an ensemble of multi-label classification models. J. Biomed. Semant. 8, 43 (2017).
29. Zhang, K., Ma, H., Zhao, Y., Zan, H. & Zhuang, L. The Comparative Experimental Study of Multilabel Classification for Diagnosis Assistant Based on Chinese Obstetric EMRs. Journal of Healthcare Engineering (2018). doi:10.1155/2018/7273451
30. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
31. Cohen, R., Elhadad, M. & Elhadad, N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC Bioinformatics 14, 10 (2013).
32. Detailed STD Facts - HIV/AIDS & STDs. (2017). Available at: https://www.cdc.gov/std/hiv/stdfact-std-hiv-detailed.htm. (Accessed: 8th March 2018)
33. Bates, J., Fodeh, S. J., Brandt, C. A. & Womack, J. A. Classification of radiology reports for falls in an HIV study cohort. J. Am. Med. Inform. Assoc. 23, e113–e117 (2016).
34. Garla, V. et al. The Yale cTAKES extensions for document classification: architecture and application. J. Am. Med. Inform. Assoc. 18, 614–620 (2011).
35. Pestian, J. P. et al. A Shared Task Involving Multi-label Classification of Clinical Free Text. in Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing 97–104 (Association for Computational Linguistics, 2007).
36. Zhang, Y. A Hierarchical Approach to Encoding Medical Concepts for Clinical Notes. in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop 67–72 (Association for Computational Linguistics, 2008).
37. Gibaja, E. & Ventura, S. A Tutorial on Multilabel Learning. ACM Comput Surv 47, 52:1–52:38 (2015).
38. Li, Y., Lipsky Gorman, S. & Elhadad, N. Section Classification in Clinical Notes Using Supervised Hidden Markov Model. in Proceedings of the 1st ACM International Health Informatics Symposium 744–750 (ACM, 2010). doi:10.1145/1882992.1883105