Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data

Daniel J. Feller[1] Oliver J. Bear Don't Walk IV[1] Jason Zucker[2] Michael T. Yin[2] Peter Gordon[2] Noémie Elhadad[1]

[1] Department of Biomedical Informatics, Columbia University, New York, New York, United States
[2] Division of Infectious Diseases, Department of Internal Medicine, Columbia University Irving Medical Center, New York, New York, United States

Address for correspondence Noémie Elhadad, PhD, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, New York, NY 10032, United States (e-mail: ne60@cumc.columbia.edu).

Appl Clin Inform 2020;11:172–181.

Abstract

Background Social and behavioral determinants of health (SBDH) are environmental and behavioral factors that often impede disease management and result in sexually transmitted infections. Despite their importance, SBDH are inconsistently documented in electronic health records (EHRs) and typically collected only in an unstructured format. Evidence suggests that structured data elements present in EHRs can contribute further to identify SBDH in the patient record.

Objective Explore the automated inference of both the presence of SBDH documentation and individual SBDH risk factors in patient records. Compare the relative ability of clinical notes and structured EHR data, such as laboratory measurements and diagnoses, to support inference.

Methods We attempt to infer the presence of SBDH documentation in patient records, as well as patient status of 11 SBDH, including alcohol abuse, homelessness, and sexual orientation. We compare classification performance when considering clinical notes only, structured data only, and notes and structured data together. We perform an error analysis across several SBDH risk factors.

Results Classification models inferring the presence of SBDH documentation achieved good performance (F1 score: 92.7–78.7; F1 considered as the primary evaluation metric). Performance was variable for models inferring patient SBDH risk status; results ranged from F1 = 82.7 for LGBT (lesbian, gay, bisexual, and transgender) status to F1 = 28.5 for intravenous drug use. Error analysis demonstrated that lexical diversity and documentation of historical SBDH status challenge inference of patient SBDH status. Three of five classifiers inferring topic-specific SBDH documentation and 10 of 11 patient SBDH status classifiers achieved highest performance when trained using both clinical notes and structured data.

Conclusion Our findings suggest that combining clinical free-text notes and structured data provide the best approach in classifying patient SBDH status. Inferring patient SBDH status is most challenging among SBDH with low prevalence and high lexical diversity.

Keywords

- social determinants of health
- electronic health records
- machine learning
- natural language processing

received
October 14, 2019
accepted after revision
January 7, 2020

DOI https://doi.org/10.1055/s-0040-1702214.
ISSN 1869-0327.

Background and Significance

Social and behavioral determinants of health (SBDH) are environmental and behavioral factors that impede disease self-management and lead to or exacerbate existing comorbid conditions.[1] The impact of determinants, such as unstable housing and substance use disorders on medical, and cost outcomes has resulted in health systems being increasingly attuned to these determinants. In addition, many SBDH are strongly associated with the acquisition of sexually transmitted infections (STIs).[2–4] While knowledge of SBDH is clinically meaningful information and can lead to tailored care plans, evidence suggests that providers often struggle to retrieve information related to SBDH from electronic health records (EHRs), and that when SBDH are neglected the overall quality of care may suffer.[5,6]

Recent research in the informatics community has focused on integrating SBDH into the EHR.[7] However, these efforts have been impeded by both the infrequent documentation of SBDH in the patient record and the lack of national standards for collecting data related to SBDH.[8,9] While social and sexual history taking is a cornerstone of the patient interview, providers often fail to adequately document their patient's self-reported SBDH.[10–14] In addition, low adoption rates for clinical screening tools for SBDH in EHRs exist as a barrier to the collection of this information in a usable format.[9,15] While the majority of documentation related to SBDH exists in free-text notes, information on SBDH is also manifest in the structured data elements such as diagnosis codes and laboratory tests.[16–18] As a result, there is a requirement for approaches that can improve clinicians' documentation of SBDH, as well as leverage heterogeneous data within EHRs, to represent a patient's individual SBDH status.

This study has several goals: (1) attempt to automatically infer the presence of SBDH documentation to support initiatives designed to improve social history taking by clinicians, and (2) evaluate methods for inferring a patient's respective SBDH from EHR data. We hypothesize that modeling approaches that leverage both structured and unstructured data for this task will yield better performance than attempts based on either data source alone. In order to inform future research in this area, we also perform an error analysis to identify challenges to automated SBDH inference.

Introduction

Previous work on extracting SBDH from clinical data has employed Natural Language Processing (NLP) techniques, reflecting the fact that such information is most reliably documented in clinical notes.[16,19] NLP approaches, and in particular information extraction techniques, have been applied to different types of SBDH including smoking status,[20–22] substance abuse,[23–25] and homelessness.[18,26] Extraction techniques that have been used include regular expressions, named entity recognition, and more contemporary distributional semantic techniques. Efforts aimed at using NLP to infer SBDH have generally achieved modest performance, reflecting the inherent challenges associated with processing clinical notes (e.g., lexical and semantic ambiguity) and challenges specific to inferring SBDH.[27,28] Most significantly, the language used to express SBDH is often institution-specific, limiting the usefulness of lexicons contained in clinical terminologies like the Unified Medical Language System (UMLS).[29] Novel approaches are required to advance the science of inferring SBDH from patient records.

An open research question is whether structured data elements in the EHR can be utilized to infer SBDH status. There are more than 1,000 distinct codes in the four major medical vocabularies (Logical Observation Identifiers Names and Codes [LOINC], Systematized Nomenclature of Medicine–Clinical Terms [SNOMED CT], International Classification of Diseases, 10th Revision [ICD-10], and Current Procedural Terminology [CPT]) that are related to SBDH.[30] ICD codes for substance abuse and homelessness have previously served as indicators of patient SBDH status, but have been shown to exhibit high specificity but poor sensitivity for the determinants they represent. Vest and colleagues recently observed that structured EHR data alone could be used to infer a patient's need for social services.[31] Moreover, due to the comorbid nature of many SBDH, multiple studies have established that demographics and diagnosis codes related to behavioral health and substance abuse were predictors of homelessness in clinical text.[29,32] Despite the potential utility of structured EHR data in systems for extracting social and behavioral determinants, no previous study has leveraged such data for this purpose.

Dataset

We described the creation of a gold-standard corpus of patient records containing information on SBDH and derivation of an outcomes variable for classifying (1) the presence of SBDH documentation in the patient record and (2) 11 patient-specific SBDH risk factors.

Curation of Social and Behavioral Determinants of Health

Three clinicians identified an array of more than 30 distinct SBDH associated with adverse health outcomes such as hospital readmission and the acquisition of STIs. The clinicians classified each SBDH as belonging to one of five SBDH topics as follows: (1) alcohol use (social alcohol use or alcoholism), (2) substance abuse (amphetamine, opiates, cannabis, cocaine, or intravenous [IV] drugs), (3) sexual orientation (men who have sex with men, men who have sex with women, women who have sex with men, women who have sex with women, or bisexual), (4) sexual activity (history of STIs, condom usage, oral sex, vaginal sex, or receptive and insertive anal intercourse), and (5) housing status (homeless, unstable housing, or living with friends).

Gold-Standard Annotation Guidelines

We chose to obtain document-level annotations rather than mention-level annotations because we observed many notes where SBDH are not expressed as named entities. For example,

we observed mentions of SBDH, such as "he used occasional EtOH (scotch) at church functions" and "three-to-four lifetime male unprotected sexual partners," that would not be amenable to extraction using named-entity recognition.

Three annotators manually reviewed clinical notes for documentation of the six SBDH topics and 30 risk factors. First, annotators reviewed the entire length of each clinical note to assess the presence of SBDH documentation. Any confirmatory or negated mention of SBDH associated with a given topic was treated as documentation, for example, "patient denies sexual activity" or "patient has history of STIs" would result in a positive label for the "sexual activity" topic. If no confirmatory or negated mentions of any SBDH associated with a given topic was observed, annotators asserted an absence of documentation for that SBDH topic. Second, in notes with confirmed SBDH documentation, annotators recorded whether a patient was described as having an "active," "historical," or "never" status related to each SBDH risk factor within the topic. Statuses of "active" and "historical" were prompted by explicit statements, while the "never" status was designated by either negative documentation ("patient denies *X*") or the absence of information related to a specific SBDH risk factor. Detailed annotation guidelines are available at: *github.com/danieljfeller/SBDSH*.

Collection of Clinical Notes

Clinical notes were obtained from the clinical data warehouse at Columbia University Medical Center (CUMC), a large academic medical center in New York City. We initially isolated two corpora that were hypothesized to contain frequent documentation of social and behavioral determinants of health. First, we obtained all notes associated with HIV+ individuals within the EHR system at CUMC; this included all distinct note types such as "Admission," "Progress," and "Social Work" notes. Second, we obtained all notes written by social workers between 2005 and 2017 for the general patient population. Both corpora included notes generated during both inpatient and outpatient encounters, and we direct readers to a previous publication for more detail on the study cohort and clinical setting.[8] The study described herein was approved by the Institutional Review Board at CUMC.

Only a small proportion of all notes collected from the EHR system were used in the study, as semisupervised learning was used to identify a subset of notes likely to contain SBDH documentation. The output of the semisupervised learning approach was used by annotators to complete the annotation process and significantly increased the yield of annotation compared to unsupported manual annotation. The implementation details and evaluation results of our semisupervised approach are described at length in a previous publication from our research group.[8]

Outcome Variables

In these experiments we trained two sets of binary classification models: (1) classifiers to infer the presence of SBDH documentation, and (2) classifiers to infer a patient's SBDH status at a specific time *T*, where *T* represents the date that the patient's status was documented in a clinical note.

Presence of SBDH Documentation

We attempted to classify whether a given SBDH topic (sexual orientation, drug use, alcohol use, or housing status) was either documented or not documented in the patient record. Every annotated note received such a label, for example, the "drug use" topic was documented if there was (1) any mention of a recreational drug or (2) negative documentation (e.g., "patient denies drug use").

Patient SBDH Status

Inference of SBDH status was considered a binary classification task of a patient's SBDH status (i.e., active or inactive). Indication of historical use (e.g., "history of substance abuse but denies drugs use") was represented using the inactive label. As described in a previous section, only notes with confirmed SBDH documentation were assigned a positive or negative SBDH label at time *T*.

We focus on 11 SBDH factors with greater than 40 positive cases in the gold-standard corpus; several SBDH classifications were composites of multiple lower level SBDH labels: "LGBT" (lesbian, gay, bisexual, and transgender) was combined from "men who have sex with men," "women who have sex with women," "bisexual," and "transgender." "Unsafe sex" was combined from "sometimes uses condoms" and "never uses condoms." "Unstable housing" was combined from labels "unstable housing" and "homeless."

Characteristics of Gold-Standard Corpus

A total of 4,663 notes associated with 1,501 patients treated at a large urban academic medical center were manually reviewed for mentions of SBDH; 3,273 notes were associated with HIV+ individuals and 1,390 social work notes associated with the general hospital population. Seventy-six notes were double annotated and a Kappa statistics of 0.736 was observed across all SBDH risk factors. The average age of persons in this cohort was 52.2 years old (standard deviation [SD] = 12.9 years) with 916 males and 585 females. The number of patients with explicit mentions of specific SBDH ranged from 274 for cocaine abuse (most prevalent) to 36 for amphetamine abuse (least prevalent, ►Table 1).

Methods

Experimental Design

For each of the five SBDH topics and 11 SBDH risk factors, a discrete classifier was trained. To avoid any potential data leakage between training and testing stages, each patient's data were included only once in the entire dataset; inputs to the classification models included a single free-text note and structured EHR data aggregated over a 30-day period preceding the free-text note (►Fig. 1).

Supervised Machine Learning to Classify SBDH Documentation and SBDH Status

Unstructured Input

Clinical documents were represented as a bag-of-words using term frequency–inverse document frequency (TF-IDF) weights.

Table 1 Performance of models inferring presence of SBDH documentation among 1,501 patients using five-fold cross validation

	Documented (missing)	Structured EHR features			Text only			Structured EHR + text		
		F1	P	R	F1	P	R	F1	P	R
Sexual history	807 (694)	64.6 ± 1.7	60.4 ± 2.1	71.6 ± 2.2	78.6 ± 1.4	80.3 ± 1.9	77.1 ± 2.0	78.7 ± 1.8	80.0 ± 2.3	77.4 ± 1.9
Sexual orientation	1,059 (442)	65.3 ± 1.9	74.8 ± 2.7	73.4 ± 2.8	85.3 ± 1.7	86.0 ± 2.6	84.9 ± 2.9	86.1 ± 1.8	86.0 ± 2.6	86.4 ± 2.5
Alcohol use	1,192 (309)	88.0 ± 1.4	80.5 ± 2.2	96.7 ± 2.0	91.3 ± 1.2	88.7 ± 1.9	94.1 ± 1.7	90.7 ± 1.3	89.0 ± 2.0	92.5 ± 1.8
Substance use	1,262 (239)	90.8 ± 1.3	90.5 ± 2.0	94.6 ± 1.7	92.5 ± 1.0	90.5 ± 1.7	94.6 ± 1.5	92.7 ± 1.1	90.7 ± 1.8	94.8 ± 1.6
Housing status	1,240 (261)	88.7 ± 1.3	83.0 ± 2.0	95.5 ± 2.1	92.6 ± 1.1	90.0 ± 1.8	95.5 ± 1.6	92.2 ± 1.1	90.0 ± 1.8	94.5 ± 1.6

Abbreviations: EHR, electronic health record; P = precision, R = recall, ± standard deviation estimated using bootstrap method; SBDH, social and behavioral determinants of health.

Preprocessing of documents comprised the following steps: (1) tokenize all documents, (2) remove all nonalphabetical characters, (3) remove general-domain stop words, and (4) remove words that were observed fewer than thrice. These steps yielded a vocabulary size of approximately 14,000 words.

Structured Input

For each patient, structured EHR data were aggregated from the 30-day period preceding the date of the patient's recorded SBDH status. We hypothesized such an extended observation period was necessary to collect enough structured EHR data to impact model performance. However, a longer observation period was not used due to the fact that a patient's SBDH status is liable to change over time.

Diagnoses, procedures, laboratory tests, and demographics were collected from the institutional data warehouse, which has been mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model, a standard for storing health care data.[33] We reduced each data table to contain only those observations that were collected between 0 and 30 days prior to the date the patient's SBDH status was observed (►Fig. 1). Structured features were represented as a vector of counts for each vocabulary item (e.g., two occurrences of ICD-9 code V08) associated with the patient in the 30 days prior to the index date.

Heterogeneous Input

When structured and unstructured data were combined, we simply concatenated the features obtained from the notes and the structured EHR data described above. This process is presented in ►Fig. 2. Hereafter, we refer to the combination of structured and unstructured EHR data as "heterogeneous" data.

Training Classifiers

We used Scikit Learn[34] to train a classifier for each of the five SBDH topics and 11 SBDH labels under three conditions (clinical notes alone, clinical notes + structured data, and structured data alone). We experimented with a variety of classifiers including L2-penalized logistic regression, support vector machines, Random Forests, CaRT, and AdaBoost. Each model training leveraged Chi-square feature selection, using 2,000 features with the strongest univariate association with the classification target; this step improved performance and reduced model training time.

We found that in all cases either AdaBoost or CaRT yielded the best performance. In order to optimize performance of the AdaBoost classifiers, we empirically identified the optimal number of weak learners (AdaBoost's primary hyperparameter) and number of features retained using Chi-square feature selection. We evaluated 30, 50, and 100 weak learners with 2,000 and 4,000 features selected using Chi-square on a development set.

We also experimented with several deep learning models that have been previously shown to successfully leverage both structured and unstructured clinical data for classification tasks.[35] We fit both feedforward and convolutional neural networks and performed hyperparameter search over learning rate, number of layers, and batch size. However,

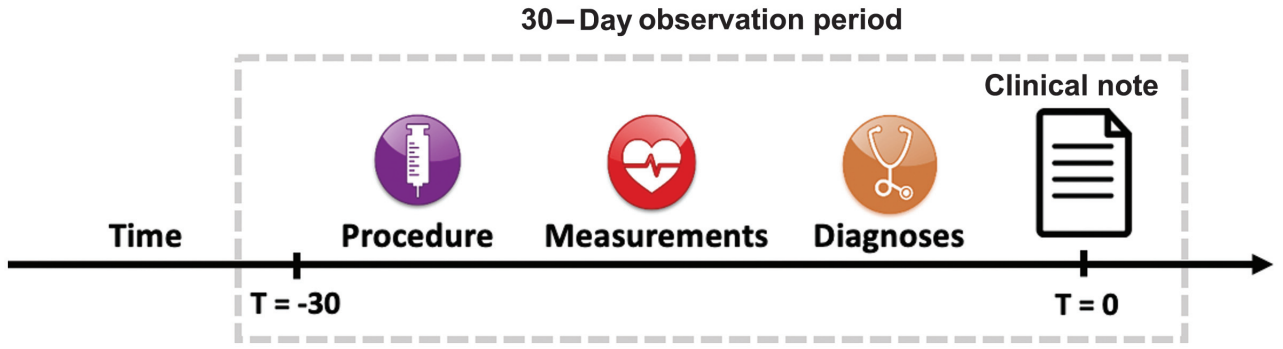

Fig. 1 Prediction Task for SBDH labels. SBDH, social and behavioral determinants of health.

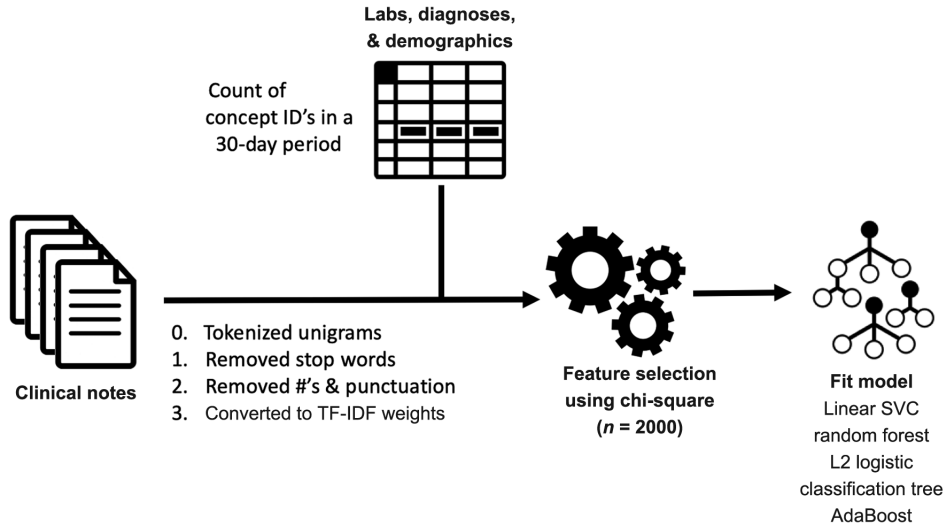

Fig. 2 Overview of methods for machine learning. TF-IDF, term frequency–inverse document frequency; SVC, support vector classifier.

all neural networks yielded worse performance compared to the machine learning approach described above due to the small size of our training dataset.

Evaluation

Precision, recall, and F1 scores were computed across the SBDH models using five-fold cross validation. F1 was considered the primary evaluation metric because the authors had no grounds to assign precedence to either precision or recall, as the determination of whether to optimize precision or recall will depend on the specific clinical application used. F1 is the harmonic mean of precision and recall and takes both metrics into account.

We estimated the SD associated with each metric by bootstrapping 200 classifiers for each SBDH on different samples and calculating the SD of resultant scores. We also estimated feature importance by using the total decrease in node impurity attributed to a single feature, averaged over all trees in an AdaBoost ensemble classifier. The effect of label frequency on classification performance was estimated using the Spearman rank correlation between a classifier's F1 score and the number of positive labels available for model training.

In addition, we conducted an error analysis to gain insight into model performance for SBDH risk factors. This was performed by reviewing a random sample of 100 incorrectly labeled patients. First, incorrectly classified patients were labeled as either a true or false error, the latter being attributable to incorrect annotation. Second, we associated each true error with one or more of the following characteristics: (1) idiosyncratic language used to express SBDH, (2) unrecognized negation, (3) attribution (e.g., "her mother is homeless"), (4) historical phrases ("he stopped drinking heavily in 2007 and now drinks approximately once per month"), (5) syntactic dependencies, (6) conflicting information, and (7) misspelling.

Results

Classifier Performance

Classification results inferring the presence of topic-specific SBDH documentation are presented in ►Table 1 and ranged from F1 = 92.7 for substance use to F1 = 78.7 for sexual history. While in three of five cases, models with text and structured data yielded the best results, these differences were not statistically significant.

The highest performing SBDH risk factor model was the classifier of LGBT status trained using heterogeneous data (F1 = 82.7; ►Table 2), while the lowest performing model was the classifier for "IV drug abuse" using structured data

Table 2 Performance of models inferring SBDH labels using five-fold cross validation

	+/−	Structured EHR features			Text only			Structured EHR + text		
		F1	P	R	F1	P	R	F1	P	R
Sexual history										
LGBT status	263/796	54.4 ± 5.1	55.9 ± 4.9	58.1 ± 5.7	79.2 ± 4.3	84.8 ± 5.3	74.7 ± 5.7	82.7 ± 4.0	**86.1 ± 4.9**	80.0 ± 5.7
History of STIs	204/603	32.3 ± 6.4	30.1 ± 7.5	48.2 ± 8.0	48.9 ± 6.3	50.1 ± 7.7	56.7 ± 7.7	54.0 ± 6.7	54.2 ± 7.9	53.7 ± 8.0
Unsafe sex	160/647	21.1 ± 6.4	21.3 ± 7.5	35.0 ± 7.4	43.8 ± 6.3	52.1 ± 7.7	38.9 ± 7.7	38.5 ± 6.5	46.0 ± 7.5	35.8 ± 8.0
Alcohol use										
Social alcohol use	252/940	27.9 ± 5.6	35.0 ± 7.3	23.8 ± 5.3	39.2 ± 6.7	49.4 ± 8.8	32.7 ± 6.7	40.1 ± 6.5	51.6 ± 8.6	33.2 ± 6.7
Alcoholism	165/1,027	33.4 ± 8.6	49.9 ± 11.5	42.4 ± 8.3	50.0 ± 7.9	61.2 ± 10.3	42.4 ± 8.3	52.0 ± 7.9	**62.8 ± 10.3**	44.8 ± 8.5
Substance use										
Marijuana use	210/1,052	29.0 ± 7.4	52.5 ± 11.1	21.4 ± 6.4	49.8 ± 6.8	51.7 ± 7.8	49.0 ± 8.3	56.4 ± 6.8	57.8 ± 7.8	55.7 ± 8.6
Cocaine abuse	274/988	56.2 ± 5.6	70.2 ± 7.3	47.0 ± 6.3	62.1 ± 5.5	67.2 ± 7.3	58.4 ± 6.3	65.1 ± 5.1	**66.0 ± 6.2**	64.6 ± 7.0
Opioid abuse	99/1,163	30.9 ± 9.9	48.8 ± 16.6	23.2 ± 8.5	37.9 ± 10.7	48.7 ± 15.1	23.2 ± 10.3	40.0 ± 11.8	48.3 ± 14.7	34.4 ± 12.0
Intravenous drug abuse	65/1,197	13.8 ± 9.6	19.9 ± 14.2	10.8 ± 10.0	27.3 ± 11.5	43.4 ± 19.6	21.5 ± 10.2	28.5 ± 12.3	38.3 ± 22.0	23.1 ± 10.1
Amphetamine abuse	36/1,226	33.6 ± 16.3	55.4 ± 36.7	27.5 ± 17.8	47.0 ± 19.5	68.4 ± 31.1	42.5 ± 18.4	51.1 ± 17.1	51.4 ± 19.7	53.5 ± 21.9
Housing status										
Unstable housing	262/978	27.4 ± 5.6	35.0 ± 6.0	23.6 ± 6.4	49.3 ± 6.4	59.4 ± 7.8	42.3 ± 7.5	53.1 ± 6.4	**62.2 ± 5.8**	46.9 ± 7.2

Abbreviations: LGBT, lesbian, gay, bisexual, transgender; P = precision, R = recall, ± standard deviation estimated using bootstrap method; SBDH, social and behavioral determinants of health; STI, sexually transmitted disease.

(F1 = 28.5). In 10 of 11 cases, training models with both text and structured data yielded better results than models trained with either of those data sources alone. In contrast to other models, the classifier for "unsafe sex" achieved the best results when trained using only text data. The best performing algorithms and their respective hyperparameters are presented in ►**Supplementary Material Appendix A** (available in the online version).

Features Used for SBDH Risk Factor Classification

The top features for the heterogeneous, text, and structured data SBDH classifiers are presented in **Supplementary Material Appendices B-D** (available in the online version). Textual features used by the classifiers included explicit indicators of SBDH, as well as cooccurring determinants. For example, top features for the cocaine abuse classifiers included "cocaine," "PSA" (for polysubstance abuse) and "heroin."

While the majority of top features utilized by the heterogeneous models were derived from text, models also included structured features. The top feature for the alcohol abuse classifier was SNOMED code 191811004 ("continuous chronic alcoholism"), while codes 191918009 ("nondependent cocaine abuse") and 78267003 ("cocaine abuse") were among the top 10 features used by the cocaine abuse classifier. LOINC code 5393-4 ("treponema pallidum antibody test") is a test for syphilis infection and was a leading indicator for having a "history of STIs."

Several textual features were institution-specific or regional in nature. For example, the word "nicholas" used in the homelessness classifier likely refers to a homeless shelter in New York City. In addition, "HASA" represents the HIV/AIDS Services Administration, a governmental organization in New York that provides housing for persons with HIV.

Label Frequency

We also tested the impact of the prevalence of each SBDH on the performance of the classification models (►**Fig. 3**). A comparison between the number of positive cases used to train each classifier and the resulting performance of that classifier yielded a correlation coefficient of 0.762 ($p = 0.0059$). Amphetamine abuse seemingly invalidated the trend, as the classifier only had 36 positive cases available but achieved a modest F1 score of 51.1.

Error Analysis

Among 100 randomly sampled incorrect classifications, 18 errors were attributed to historical phrases such as "history of cocaine snorting quit 18 years ago." Seventeen were attributed to short- and long-term syntactic dependencies such as "reports very large amounts of alcohol consumption, IV heroin, and cocaine use." Unrecognized negation was associated with 15 errors (e.g., "has not had EtOH intake in over 20 years"), and lexical diversity accounted for 13 errors (e.g., "actively smokes crack") that reflected use of idiosyncratic language by clinicians. Three errors reflected misspellings (e.g., "former IV cocaine and heroin use"), and two errors reflected conflicting information in the note (e.g., "his wife was present during the interview … in private [patient] reported sex with men"). Four errors were in fact correct and attributed to inaccurate annotations.

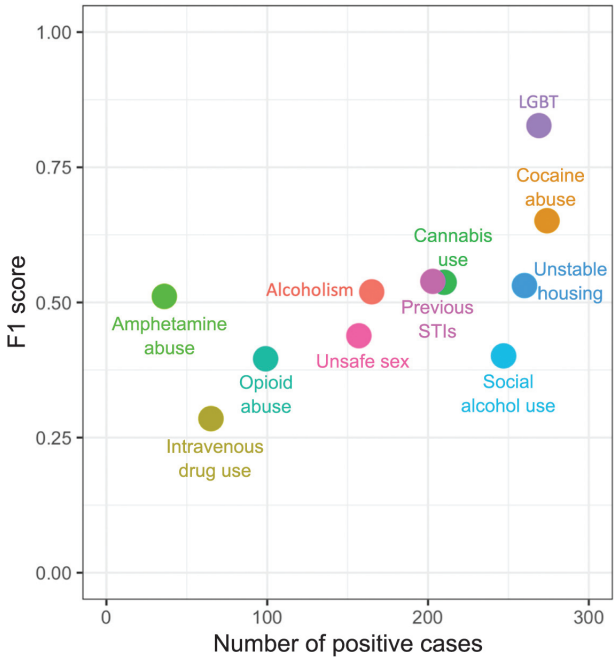

Fig. 3 Relationship between SBDH prevalence and classifier performance. LGBT, lesbian, gay, bisexual, transgender; SBDH, social and behavioral determinants of health; STI, sexually transmitted disease.

Discussion

Our findings suggest that the identification of topic-specific SBDH documentation and individual SBDH risk factors can be improved by leveraging both structured EHR data and clinical notes. We also provide evidence that model performance is correlated with the lexical diversity used by clinicians to document a given SBDH and the prevalence of a given SBDH within a patient population.

The presence of topic-specific SBDH documentation in the patient record was inferred using classification models. A 2014 report published by the Institute of Medicine brought attention to the importance of collecting SBDH information in EHRs, as well as the fact that such information is sporadically collected in patient records.[36,37] The acceptable performance of classification models trained to infer presence of SBDH documentation suggests that information and technology (IT) systems could alert providers when certain SBDH topics are undocumented in the patient record, thereby supporting the development of quality initiatives to improve provider's documentation of SBDH. Such an approach could increase the specificity of EHR prompts alerting clinicians to collect SBDH information, which have been previously deployed in clinic settings.[38,39]

The combination of free-text and structured data yielded better performance than either data source alone when inferring SBDH risk factors. These findings are corroborated by recent studies that combined textual features with diagnoses and laboratory data and observed improved phenotyping and prediction compared to using those sources alone. Several of these techniques have found improved performance by preprocessing textual data with topic modeling[40,41] and structured data with autoencoders.[42] More recently, deep neural networks have been used to leverage heterogeneous clinical data for prediction, although our findings demonstrate that these methods require much larger datasets those are currently available in the SBDH domain.[43,44]

We observed a positive correlation between model performance and the prevalence of each specific SBDH. This demonstrates the necessity of building gold-standard corpora of adequate size, especially for infrequently documented SBDH such as those related to sexual activity.[8] However, the "amphetamine use" and "LGBT" classifiers outperformed SBDH models for labels with similar prevalence, likely reflecting the limited lexical diversity used to express these SBDH. For example, amphetamine use was often referenced by "meth" or "methamphetamine" and most LGBT patients in our cohort were gay who were characterized as "MSM" or "men who have sex with men."

The results of our error analysis suggest several areas for improvement in automated SBDH inference. The inability of the SBDH classifiers to detect syntactic dependencies and historical phrases is unsurprising, given our use of a simple bag-of-words approach to extracting information from clinical text. In addition, several of the SBDH were associated with high lexical diversity, suggesting that clinicians lack a standardized way for expressing those SBDH. While contemporary methods that leverage distributional semantics and use neural networks to model temporal sequences can overcome such challenges, these methods require very large datasets that are difficult to curate in the medical domain.[8,45] Transfer learning, which entails pretraining a neural network on a large, related dataset, and subsequently "fine tuning" the network on a smaller dataset, has the potential to overcome the aforementioned barriers posed by the challenge of collecting large annotated corpora.[46,47]

Limitations

First, our SBDH classifiers were trained using data from a single institution.[48] However, our use of the OMOP Common Data Model enables generalizability of our trained models to other institutions. Second, our overall modest results may have resulted from data quality issues in the documentation of SBDH and/or inaccurate annotation. Third, most approaches cast this problem as a named-entity recognition task but because we approach the problem as a document labeling task, our experimental setup does not allow for direct comparison to previous work. With larger training datasets, future studies may be able to experiment with using machine learning for information extraction. Fourth, our model performance may have been improved by considering negation or by correcting misspellings in text; we did not consider negation due to the fact that not all SBDH studied herein would have benefitted from this addition. Fifth, a considerable number of the patients whose records were used to train the classification models have HIV; the records of these patients likely differ from patients who are currently HIV negative, potentially compromising the generalizability of the classification models. Sixth, we did not use a "holdout" dataset that was never used in model training;

we simply did not have the requisite volume of data to create training, validation, and test sets and thus the observed model performance may be inflated. Future studies should use a holdout set to estimate unbiased model performance.

Conclusion

We observed that the combination of structured and unstructured data improves automated inference of SBDH documentation from the patient record, motivating the development of EHR prompts to improve the quality of provider documentation of SBDH. In addition, our findings suggest that while automated inference of patient SBDH status is challenging, the combination of text and structured EHR data improves performance compared to either data source alone. The study findings also suggest that SBDH prevalence and the lexical diversity used to express a given determinant have an impact on the performance of classification algorithms for this purpose. Future studies should explore computational methods that can effectively learn models using datasets of limited size.

Clinical Relevance Statement

Our findings demonstrate the feasibility of inferring social and behavioral determinants of sexual health from EHRs. 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.

Multiple Choice Questions

1. What characteristics of SBDH documentation impacted the ability of machine learning models to automatically infer SBDH risk factors?
   a. Unrecognized negation
   b. Lexical diversity
   c. Label prevalence
   d. (a, b, and c)
   e. (b and c)

   Correct Answer: The correct answer is option e, as both lexical diversity and label prevalence impact the author's ability to automatically infer SBDH risk factors. The lexical diversity used to express SBDH was negatively associated with model performance. For example, amphetamine use was often referenced in one of two ways; "meth" or "methamphetamine." The majority of LGBT patients in our cohort were gay men who were characterized as "MSM" or "men who have sex with men." In contrast, unstable housing had a similar cohort prevalence but was more challenging to infer, reflecting the myriad ways this SBDH was documented (e.g., "shelter," "homeless," and "hotel"). Less prevalent labels were also more challenging to infer; a comparison between the number of positive cases used to train each classifier and the resulting performance of that classifier yielded a correlation coefficient of 0.762 ($p = 0.0059$).

2. Why did the authors use a document classification approach to inference SBDH, instead of the more traditional named-entity-recognition?
   a. SBDH mentions are not typically expressed as named entities.
   b. easier to train machine learning models when using document classification.
   c. the annotation process required to train named-entity-recognition models is more labor intensive.
   d. the authors were interested in developing a novel approach to SBDH classification.

   Correct Answer: The correct answer is option a. The authors chose to obtain document-level annotations rather than mention-level annotations because of evidence that SBDH are not typically recorded as named entities. For example, we observed mentions of SBDH such as "he used occasional EtOH (scotch) at church functions" and "three-to-four lifetime male unprotected sexual partners" that would not be amenable to extraction using named-entity recognition.

Protection of Human and Anmial Subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and was reviewed by the Institutional Review Board at Columbia University Medical Center.

Funding

This study was funded by the following sources:
National Library of Medicine—T15 LM007079: "Training in Biomedical Informatics at Columbia University."
National Institute of Allergy and Infectious Diseases—T32AI007531 "Training in Pediatric Infectious Diseases."
National Institute of General Medical Sciences—R01 GM114355.

Conflict of Interest

None declared.

Acknowledgements

The authors would like to thank Henry Evans and Roxanna Martinez for their help in preparing the gold-standard corpus.

References

1 Cohen SM, Hu X, Sweeney P, Johnson AS, Hall HI. HIV viral suppression among persons with varying levels of engagement in HIV medical care, 19 US jurisdictions. J Acquir Immune Defic Syndr 2014;67(05):519–527
2 Facente SN, Pilcher CD, Hartogensis WE, et al. Performance of risk-based criteria for targeting acute HIV screening in San Francisco. PLoS One 2011;6(07):e21813–e21813
3 Haukoos JS, Hopkins E, Bender B, Sasson C, Al-Tayyib AA, Thrun MW; Denver Emergency Department HIV Testing Research Consortium. Comparison of enhanced targeted rapid HIV screening using the Denver HIV risk score to nontargeted rapid HIV screening in the emergency department. Ann Emerg Med 2013;61(03):353–361
4 Lauby J, Zhu L, Milnamow M, et al. Get real: evaluation of a community-level HIV Prevention intervention for young MSM

who engage in episodic substance use. AIDS Educ Prev 2017;29 (03):191–204
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(02):215–218
6 Weir CR, Staggers N, Gibson B, Doing-Harris K, Barrus R, Dunlea R. A qualitative evaluation of the crucial attributes of contextual information necessary in EHR design to support patient-centered medical home care. BMC Med Inform Decis Mak 2015;15:30
7 Cantor MN, Chandras R, Pulgarin C. FACETS: using open data to measure community social determinants of health. J Am Med Inform Assoc 2018;25(04):419–422
8 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 AMIA Symp 2018;2018:422–429
9 Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. Health Aff (Millwood) 2018; 37(04):585–590
10 Riese A, Tarr EE, Baird J, Alverson B. Documentation of sexual history in hospitalized adolescents on the general pediatrics service. Hosp Pediatr 2018;8(04):179–186
11 Siegel J, Coleman DL, James T. Integrating social determinants of health into graduate medical education: a call for action. Acad Med 2018;93(02):159–162
12 Andermann A. Screening for social determinants of health in clinical care: moving from the margins to the mainstream. Public Health Rev 2018;39(01):19
13 Dubin SN, Nolan IT, Streed CG Jr., Greene RE, Radix AE, Morrison SD. Transgender health care: improving medical students' and residents' training and awareness. Adv Med Educ Pract 2018; 9:377–391
14 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(03):e13802
15 Gottlieb L, Ackerman S, Wing H, Adler N. Evaluation activities and influences at the intersection of medical and social services. J Health Care Poor Underserved 2017;28(03):931–951
16 Chen ES, Manaktala S, Sarkar IN, Melton GB. A multi-site content analysis of social history information in clinical notes. AMIA Annu Symp Proc 2011;2011:227–236
17 Walsh C, Elhadad N. Modeling clinical context: rediscovering the social history and evaluating language from the clinic to the wards. AMIA Jt Summits Transl Sci Proc 2014; 2014:224–231
18 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 JAMIA 2018; 25(01):61–71
19 Navathe AS, Zhong F, Lei VJ, et al. Hospital readmission and social risk factors identified from physician notes. Health Serv Res 2018; 53(02):1110–1136
20 McCormick PJ, Elhadad N, Stetson PD. Use of semantic features to classify patient smoking status. AMIA Annu Symp Proc 2008; 2008:450–454
21 Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(01):14–24
22 Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc 2008;15(01):25–28
23 Yetisgen M, Vanderwende L. Automatic identification of substance abuse from social history in clinical text. In: Teije AT, Popow C, Holmes JH, Sacchi L. Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017. Vienna, Austria: Springer Cham; 2017:171–181
24 Carter EW, Sarkar IN, Melton GB, Chen ES. Representation of drug use in biomedical standards, clinical text, and research measures. AMIA Annu Symp Proc 2015;2015:376–385
25 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
26 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
27 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
28 Demner-Fushman D, Elhadad N. Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. IMIA Yearb 2016;(01):224–233
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(01):61–71
30 Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. JAMIA Open 2019;2(01):81–88
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
32 Erickson J, Abbott K, Susienka L. Automatic address validation and health record review to identify homeless social security disability applicants. J Biomed Inform 2018;82:41–46
33 Hripcsak G, Shang N, Peissig PL, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform 2019; 96:103253
34 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–2830
35 Liu J, Zhang Z, Razavian N. Deep EHR: chronic disease prediction using medical notes. Available at: http://arxiv.org/abs/1808.04928. Accessed September 19, 2018
36 Institute of Medicine. Clinical rationale for collecting sexual orientation and gender identity data. In: Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records: Workshop Summary. Washington (DC): National Academies Press (U.S.); 2013
37 Populations I of M (US) B on the H of S. Existing Data Collection Practices in Clinical Settings. National Academies Press (United States). Available at: https://www-ncbi-nlm-nih-gov.ezproxy.cul.columbia.edu/books/NBK154082/ Accessed January 30, 2020
38 Friedman NL, Banegas MP. Toward addressing social determinants of health: a health care system strategy. Perm J 2018; 22:18–95
39 Taylor LA, Tan AX, Coyle CE, et al. Leveraging the social determinants of health: what works? PLoS One 2016;11(08):e0160217
40 Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc 2015;22(04):872–880
41 Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assoc 2016;23(e1):e11–e19
42 Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6:26094–26094
43 Liu H. Automatic argumentative-zoning using word2vec. Available at: http://arxiv.org/abs/1703.10152. Accessed October 12, 2018
44 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1(01):18

45 Halpern Y, Choi Y, Horng S, Sontag D. Using anchors to estimate clinical state without labeled data. AMIA Annu Symp Proc 2014; 2014:606–615
46 Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. Available at: http://arxiv.org/abs/1904.05342. Accessed May 9, 2019
47 Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. Available at: http://arxiv.org/abs/1904.03323. Accessed May 9, 2019
48 Ferraro JP, Ye Y, Gesteland PH, et al. The effects of natural language processing on cross-institutional portability of influenza case detection for disease surveillance. Appl Clin Inform 2017;8(02): 560–580