Journal of the American Medical Informatics Association, 25(1), 2018, 61–71
doi: 10.1093/jamia/ocx059
Advance Access Publication Date: 8 July 2017
Research and Applications

Research and Applications

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

Cosmin A Bejan,[1] John Angiolillo,[2] Douglas Conway,[3] Robertson Nash,[2] Jana K Shirey-Rice,[3] Loren Lipworth,[2] Robert M Cronin,[1,2,4] Jill Pulley,[3] Sunil Kripalani,[2] Shari Barkin,[4] Kevin B Johnson,[1,4] and Joshua C Denny[1,2]

[1]Department of Biomedical Informatics; [2]Department of Medicine; [3]Institute for Clinical and Translational Research; [4]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

Corresponding Author: Cosmin Adrian Bejan, Department of Biomedical Informatics, Vanderbilt University, 2525 West End Avenue, Suite 1450, Nashville, TN 37203, USA. Phone: (615) 875-2422, Fax: (615) 322-0502, E-mail: adi.bejan@vanderbilt.edu

Received 3 November 2016; Revised 22 April 2017; Accepted 10 May 2017

ABSTRACT

Objective: Understanding how to identify the social determinants of health from electronic health records (EHRs) could provide important insights to understand health or disease outcomes. We developed a methodology to capture 2 rare and severe social determinants of health, homelessness and adverse childhood experiences (ACEs), from a large EHR repository.

Materials and Methods: We first constructed lexicons to capture homelessness and ACE phenotypic profiles. We employed word2vec and lexical associations to mine homelessness-related words. Next, using relevance feedback, we refined the 2 profiles with iterative searches over 100 million notes from the Vanderbilt EHR. Seven assessors manually reviewed the top-ranked results of 2544 patient visits relevant for homelessness and 1000 patients relevant for ACE.

Results: word2vec yielded better performance (area under the precision-recall curve [AUPRC] of 0.94) than lexical associations (AUPRC = 0.83) for extracting homelessness-related words. A comparative study of searches for the 2 phenotypes revealed a higher performance achieved for homelessness (AUPRC = 0.95) than ACE (AUPRC = 0.79). A temporal analysis of the homeless population showed that the majority experienced chronic homelessness. Most ACE patients suffered sexual (70%) and/or physical (50.6%) abuse, with the top-ranked abuser keywords being “father” (21.8%) and “mother” (15.4%). Top prevalent associated conditions for homeless patients were lack of housing (62.8%) and tobacco use disorder (61.5%), while for ACE patients it was mental disorders (36.6%–47.6%).

Conclusion: We provide an efficient solution for mining homelessness and ACE information from EHRs, which can facilitate large clinical and genetic studies of these social determinants of health.

Key words: text mining, homelessness, adverse childhood experiences, social determinants of health, EHR

INTRODUCTION

Social and behavioral determinants of health affect health and disease trajectories across an individual's lifespan. They have been shown to be associated with early onset and progression of various diseases, such as cardiovascular disease and type 2 diabetes, and a higher risk of premature death.[1–4] In 2014, the National Academy of Medicine underscored the importance of capturing these determinants of health in electronic health records (EHRs) to help guide clinical care.[5,6] However, the extent to which they are encoded in EHRs is unknown, and many of these phenotypes may only be available in clinical notes; hence, natural language processing (NLP) methodologies need to be developed to automatically extract relevant phenotypic information to be used as key data elements in large observational studies.[7] While multiple NLP methods have been successfully designed to identify various clinical diseases[8–12] and some behavioral determinants of health, such as tobacco use, alcohol use, and drug use, have been commonly investigated in EHRs,[13–16] approaches for extracting social determinants of health (SDH) from clinical text are less well developed.

Recently, reports have linked severe SDH including homelessness and adverse childhood experiences (ACEs) to poor health outcomes[17–19]; however, these are less often extracted from EHRs. In part, this could be due to a low prevalence of these determinants in health data repositories,[20–22] which makes the task of mining them more challenging. Previous studies on homelessness identification have tried utilizing administrative data such as diagnostic codes for classifying diseases[23,24] and residential address information acquired during patient registration.[25,26] However, the validity and reliability of these approaches have not been determined and administrative codes are not sufficient to ascertain the occurrence of homelessness in EHRs[21]; furthermore, address information is not available in deidentified clinical data, which are often used in research studies.[27,28] More important, these approaches are unable to capture significant related SDH such as vulnerability to homelessness or past history of homelessness. ACE is also not typically captured in structured EHR fields, since it is a construct comprising multiple components, many of which describe adverse events among adults that occurred long in the past.[29]

Motivated by the need for large observational studies to support specific interventions for the health risks associated with SDH,[29–34] the main goal of this study was to develop and test a large-scale text-mining approach for extracting homelessness and ACE from a big dataset of clinical notes. Using the Synthetic Derivative database, a deidentified version of the EHR at Vanderbilt University Medical Center, we integrated unsupervised learning and information retrieval methodologies to build SDH profiles and to identify the patients who best matched these phenotypic profiles. The primary requirement in the design of our system architecture was dictated by the need to achieve scalability on a large and dynamic EHR repository that is continuously updated with new clinical data. Notably, choosing an approach that requires 200–10 000 manually reviewed patient charts – a strategy adopted by the majority of NLP-based phenotype algorithms[35] – is infeasible for studying low-prevalence phenotypes such as homelessness and ACE. Similar to the majority of the phenotype algorithms from the Phenotype KnowledgeBase catalog,[36] our approach is highly customizable and can be iteratively refined to improve its performance. Details and updates about our SDH algorithm are published on the Phenotype KnowledgeBase website (https://phekb.org).

METHODS

Figure 1 depicts the general architecture of our system for searching SDH over the entire collection of clinical notes from the Vanderbilt Synthetic Derivative. As of October 4, 2016, this repository includes 100 485 756 clinical notes for 2 634 057 patients. A first step in this process consists of building a phenotypic profile for each SDH, homelessness and ACE, to query the note collection (query expansion module). Next, a retrieval model computes the similarity between the SDH profile (query) and the clinical document representation of each patient from the Synthetic Derivative (SDH retrieval module). Finally, the top-ranked patients are manually assessed for SDH relevance (relevance assessment module). We present these modules in greater detail below.

SDH query formulation

We adopted 2 strategies to build a list of query terms (ie, single- or multi-word expressions) that define an SDH profile. For ACE, we relied on a strategy based on domain expertise. The main reason we decided on this strategy is because ACE is a complex phenotype that covers multiple categories of adverse childhood experiences ranging from exposure to natural disaster to psychological abuse. Since each ACE category has a broad range of ways to be described in clinical text, we relied on our clinical experts to identify the most representative expressions of each category.

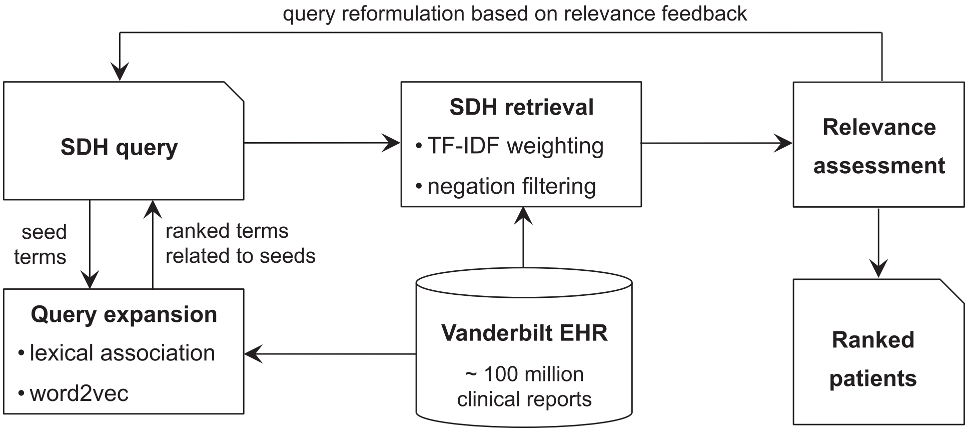

Figure 1. System architecture for identifying SDH in the Vanderbilt EHR

For homelessness, we implemented 2 data-driven methodologies, lexical association and word2vec, to expand an initial list of 2 relevant seed keywords, $S = \{homelessness,\ homeless\}$. Highly recommended for generating phenotype queries,[37] data-driven approaches have the ability to automatically extract a phenotypic profile from millions of clinical notes with minimal domain knowledge about the phenotype of interest. Both methods used for homelessness query expansion are designed on the premise that the best candidate words to describe the determinant are the ones that occur in similar contexts with the seed words from S. The top-ranked words generated by these methods were manually assessed and the highly relevant ones were included in the homelessness query.

It is worth mentioning that alternative techniques for homelessness query expansion include (1) ontology-based methods that explore the ontology relationships between various concepts and the seed keywords (eg, the synonymy and hypernymy-hyponymy relations from the ontologies available in the Unified Medical Language System)[38] and (2) methods that use homelessness-related information available in official documentation provided by different organizations such as the US Department of Health and Human Services.[39,40] However, these resources may not precisely reflect how homelessness is represented in a given EHR. For instance, some homelessness-related concepts may be specific to the region where the EHR is implemented. Furthermore, unlike data-driven approaches, these methods do not have the capability of discovering new homelessness-related concepts that are not available in the above-mentioned resources.

Query expansion based on lexical association

The first approach for homelessness query expansion extended our preliminary work on mining phenotypic keywords from large collections of clinical notes.[41] Starting with all the notes from the Synthetic Derivative, the method first selected all the documents with at least 1 seed word from S. Next, the content of these documents was tokenized, the tokens were converted to lowercase, and the low-frequency tokens and punctuation marks were discarded. After this preprocessing step, each word w occurring in the context of a seed was ranked according to the following formula:

$$\text{score}(w) = \log \frac{\text{LAM}(w,\ S)}{(1 + f(w))^{\lambda}}$$

Here, $f(w)$ represents the frequency counts of w, λ is a parameter that controls the degree to which the inverse word frequency of w should weight in the relevance score, and LAM denotes a metric of lexical association between w and the seed words. The empirical setup for this approach consisted of configurations using the word context size in $\{5,\ 10,\ 20,\ 30\}$, $\lambda \in \{0.0,\ 0.1, \ldots,\ 1.0\}$, and lexical association measures including Chi-square test, t-test, Fisher's exact test, Dice's coefficient, and pointwise mutual information.

Query expansion based on word2vec

Motivated by recent work on unsupervised distributional semantics for query expansion in open-domain question answering,[42–44] the second approach was based on word2vec.[45,46] word2vec, named after the popular software package developed at Google (https://code.google.com/p/word2vec/), is a collection of neural probabilistic language models that are able to learn distributional word representations or word embeddings from a large text collection. The resulting word embeddings, which are represented as multidimensional vectors of real numbers, were proven to predict semantically similar words with high accuracy in multiple NLP applications. In our query expansion module, we used the skip-gram model[46] of word2vec. For each word w in the 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. Specifically, 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 vector of w_k, c_k through a softmax function[46]:

$$p(w_k|w_j) = \frac{\exp(c_k^{\top} \cdot v_j)}{\sum_{i=1}^{|V|} \exp(c_i^{\top} \cdot v_j)}$$

The objective is to find the word embeddings that maximize 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 a collection of 10 million randomly sampled notes ($\sim$1.5 billion words) from the Synthetic Derivative. Before training, we preprocessed the notes by following the same steps as in the first query expansion approach. For model configuration, we used the hierarchical softmax to approximate the full softmax function, a vector dimension of 100, and context window sizes of 5 and 15. Once the word embeddings were learned, we used the cosine similarity metric to measure the similarity between the vectors corresponding to all words in V and the vectors associated with the seeds in S.

SDH retrieval

Despite the availability of many open-source search engines,[47,48] we decided to implement our own retrieval model that worked directly with the Synthetic Derivative database. Our decision was dictated mainly by the policy regulations on using EHR data outside the Synthetic Derivative; nevertheless, maintaining an up-to-date inverted index of >100 million notes from a daily updated EHR database is difficult in practice. The data management server that stores the database provided the scalability of our retrieval system. This is a secure IBM Netezza 1000 data warehouse appliance, which consists of a parallel computing architecture allowing for rapid analysis of massive data volumes. In our experiments, each search lasted <20 min over the entire Synthetic Derivative.

The SDH retrieval model was implemented based on the vector space model architecture in which both patients and SDH queries were represented as multidimensional vectors and each element of the vector was associated with a word or word expression from the notes stored in the Synthetic Derivative. In this model, the similarity of a patient to an SDH query was measured by the standard term frequency–inverse document frequency (TF-IDF) weighted cosine metric such that the most relevant patients to the query were ranked on top of the retrieved patient list. Notably, each patient in our framework was represented as a meta-document that included all the patient notes. To weight the query term i in the meta-document of patient j, we computed the TF-IDF weighting scheme as follows:

$$w_{i,j} = tf_{i,j} \cdot \log \frac{N}{df_i}$$

where $tf_{i,j}$ is the number of occurrences of the term i in the meta-document of patient j, df_i is the number of patients whose corresponding meta-documents contain the term i, and N is the total number of patients in the Synthetic Derivative. During the process of computing the term and patient counts, the punctuation marks were discarded.

Since SDH queries can contain terms corresponding to medical concepts, an investigator using our system has the option to select which of the query terms need to be checked for negation. To check whether terms were negated in patient notes (eg, "patient denies sexual abuse"), we used a simple and efficient rule-based system, NegEx.[49] Based on a simple majority voting approach for phenotype identification,[50] our retrieval model excluded patient data with a high prevalence of negated terms.

SDH assessment

We invested significant effort into better understanding how homelessness and ACE are encoded in clinical text and improving the process for retrieving them from the EHR. Seven assessors (4 for homelessness and 3 for ACE), who were either medical graduate students or faculty scientists involved in clinical research, performed the identification of SDH elements through manual examination of the retrieved patient notes. Of note, for each patient selected for assessment, the assessors analyzed all the corresponding patient notes, including those written by nurses, social workers, physical and occupational therapists, medical students, and others. Board-certified clinicians prepared and discussed with the assessors common annotation guidelines of the 2 determinants of health.

Query reformulation based on relevance feedback

Query reformulation based on relevance feedback is typically used when the query topic is unclearly defined in a document collection but can be easily evaluated through specific document examinations.[51] We adopted this approach to improve each search by iteratively refining the 2 SDH queries. For this, the assessors examined, on average, the top 20 results of each search to decide which query terms needed to be reformulated. An example of how the ACE query was reformulated is provided in Supplementary Note S1. To expand the homelessness query, a clinician who regularly treats homeless patients at Vanderbilt University Medical Center manually assessed the top 50 keywords generated by the 2 data-driven methods. The most prevalent and relevant keywords from each run were selected for homelessness query reformulation. During this process, the selection of terms to be checked for negation was also performed. Once the queries were finalized, all the patients analyzed at this step were excluded for the last SDH retrieval.

Homelessness assessment

Since homelessness is generally a temporary condition, the assessment for this determinant was accomplished at the patient visit level using 5 categories: homeless, settled, at risk, undetermined, and pediatric. Examples relevant for the homeless category include mentions of the patient being homeless, living in a shelter, motel, or car, or temporarily living with a friend or family member. Examples of at risk include multiple mentions of unemployment, criminal record, mental illness, substance abuse, and extreme poverty. Patients assessed on at least 1 visit as homeless or at risk were selected as relevant for homelessness, and their age at the first visit in either of these 2 categories was recorded in the EHR as the age when they first experienced homelessness. The notes from 2 consecutive visits were merged if the duration between visits was <15 days.

For assessment, all the notes of 600 patients were selected from 3 distinct sets: (1) the case set, corresponding to the top 200 ranked patients; (2) the fuzzy set, randomly selected patients with a rank >5000; and (3) the control set, randomly selected patients who did not have homelessness query terms in their notes. Since our objective was to use the assessed patients in case-control studies of homelessness, we estimated that most of the patients from the first and third set would be identified as valid cases and controls for this determinant, respectively. Of note, controls may be defined differently in homelessness association studies for a more precise definition of the undetermined category. The ability to evaluate the homelessness retrieval model for patients with a relatively higher rank was the main reason for assessing the patients from the fuzzy set. The corresponding patient visits were randomly shuffled to perform an unbiased assessment between consecutive visits of the same patient. Ten percent of the visits were double annotated, and conflicts were resolved by the clinician with expertise in homelessness. The interrater agreement was computed with Cohen's kappa.

ACE assessment

The ACE assessment was performed by manual examination of all the notes corresponding to the top 1000 retrieved patients using 3 categories: ACE, not ACE, and undetermined. The annotation guidelines for the ACE category comprised the 3 types of childhood abuse studied by Felitti and colleagues[29] (psychological, physical, and sexual abuse), poverty, food insecurity, and other forms of adverse childhood experiences. Examples of expressions found in clinical notes that are strong indicators of the ACE category included: "*Hx of sexual abuse and rape at **AGE[in teens] yo by father*," "*mother incarcerated for aggravated child neglect of this patient*," "*Pt has a history of sexual abuse and extreme domestic abuse of mother by father including an attempted shooting of mother by father and witnessed father shooting himself in the stomach*," and "*B/t ages of 2 and 3 (**DATE[Jun 96]-**DATE[Sep 96]) for a several-month period, mother reports that pt was sexually abused, possibly violently, at father's house by a male neighbor/friend of family.*" In cases where additional information was found for ACE patients, the assessors were also asked to indicate the abuse type, who was the abuser, and the patient's age when the abuse happened. When age was unclear (eg, "*sexual abuse during childhood*"), a default age range of 0–18 was assigned; also, appropriate age intervals were selected for different child development stages. Examples of comments provided by the annotators for ACE patients include "*physical abuse by mother until high school, sexual abuse by uncle as a child*" and "*history of sexual abuse by father and friend.*" Notably, since our goal was to conduct studies on the long-term effects of ACE in adult patients, a patient who was abused in adulthood (eg, by her husband) but not as a child was classified as not relevant for ACE. Most of the undetermined patients for ACE had expressions indicating possible abuse or history of abuse experiences at an unspecified age. The top 100 patients were assessed by 3 raters; their conflicts were solved by majority voting or discussion with a clinician, and their interrater agreement was measured with Fleiss' kappa.

Evaluation and data analysis

Evaluation of the query expansion and retrieval modules consisted of comparing the manual annotations with the automatically generated results. The resulting performance values were reported using precision (P), recall (R), F1 score (F1), precision-recall curves, precision of top *k* ranked results (P@*k*), and area under the precision-recall curve (AUPRC), which was estimated based on the average precision measure.[51] The 95% confidence intervals (CIs) of the AUPRC estimators were computed using a bootstrap procedure. Specifically, for the interval bounds of each point estimate, we used the empirical quantiles of the resampled data generated by 1000 bootstrap replicates.[52,53]

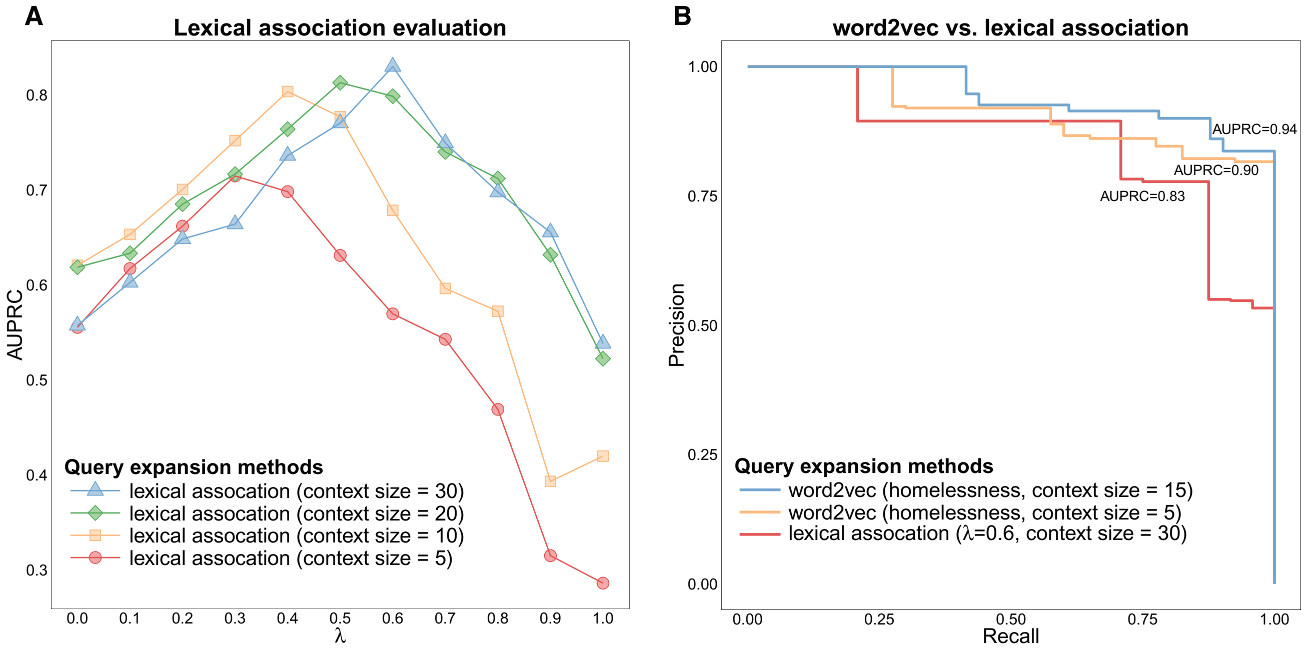

Figure 2. Evaluation of methods for homelessness query expansion

Table 1. Best method results for homelessness query expansion

Method	P@10	P@50	AUPRC (95% CI)
Lexical association ($\lambda = 0.6$, context size = 30)	0.80	0.48	0.83 (0.70–0.95)
word2vec (homelessness, context size = 5)	1.00	0.80	0.90 (0.84–0.96)
word2vec (homelessness, context size = 15)	1.00	0.82	0.94 (0.89–0.98)

Table 2. Final query terms used to search for homelessness and ACE phenotypes

Homelessness: homeless, homelessness, shelter, unemployed, jobless, incarceration

ACE: child abuse, sexual abuse, child neglect, childhood trauma, child protective service, physical abuse, psychological abuse, verbal abuse, poverty, food insecurity, cps supervisor, cps report, cps worker, cps investigation

Several analyses were performed to further examine the SDH elements identified by the 7 assessors. Analysis of long-time exposure of patients found as relevant for homelessness and ACE was performed by computing race and sex distributions over years. This type of demographic analysis provides the framework for measuring the dynamics of identified homeless and ACE patients and a better understanding of how the sex and racial composition of these patients changed over a long period of time. Another analysis extracted the most prevalent International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) billing codes assigned to these patients. For ACE patients, the assessor's comments were automatically processed to extract descriptive statistics of abuse types and the persons who committed abuse. Expressions that conveyed uncertainty (eg, "possibly/suspected sexual abuse") were discarded. Finally, the last analysis helped visualize the structural changes of the homelessness condition over time. The motivation for this type of analysis was to capture the major transitions between the homelessness categories and reveal various homelessness types, such as chronic and episodic.

RESULTS

Homelessness query expansion

The best results of our first query expansion approach (AUPRC = 0.83) were achieved when the Chi-square test was selected as the lexical association measure, the word context size was set to 30, and $\lambda = 0.6$. Our parameter search for this approach indicated that experiments using the inverse word frequency in the 0.4–0.6 range obtained better results than the baseline experiments relying only on the Chi-square test (Figure 2A). Supplementary Table S1 lists the top 50 results extracted by this method for various parameter configurations. However, 5 out of 6 word2vec experiments outperformed the best lexical association experiment. Figure 2B compares the precision-recall curves of 2 word2vec experiments (using the homelessness embedding and context sizes of 5 and 15) with the best-performing Chi-square test experiment. Table 1 lists the AUPRC, P@10, and P@50 values corresponding to these methods. As reflected in their performance, the word2vec experiments managed to maintain a high precision value even at the 50th ranked word (P@50 of 0.80 and 0.82). The top 100 words generated by all the word2vec experiments are listed in Supplementary Table S2. Additionally, Supplementary Table S3 shows the contingency tables of the optimum word ranks associated with the best F1 scores for each query expansion method.

SDH retrieval

Table 2 lists the final query terms that were used to rank the most relevant patients for the 2 SDH. Out of the 2 634 057 patients from the Synthetic Derivative, the searches corresponding to the homelessness and ACE queries from this table retrieved 35 220 patients (1.3%) and 27 861 patients (1.1%), respectively (Supplementary Table S4). Despite the difficulty of the assessment tasks and the number of categories to analyze, substantial interrater agreement was achieved for both homelessness (Cohen's κ = 0.775) and ACE (Fleiss' κ = 0.772). As a result of the homelessness assessment effort, 172 patients were identified as relevant for this SDH in the top 200 patients. Only 15 pediatric patients were found in this set. Thus, the overall precision for identifying relevant patients for homelessness in the top 185 retrieved adult patients was 93% (172/185). As expected, this is a significantly higher value when compared with the overall precision of 40% (70/175) for identifying the adult patients relevant for homelessness from the fuzzy set (Supplementary Table S5). Furthermore, as listed in Supplementary Table S5, only 5 at-risk and 0 homeless patients were identified in the control set. For ACE, our system achieved an overall precision of 70% (696/1000) for the top 1000 ranked patients. Supplementary Tables S5–S7 list additional information on the SDH assessment. Also, Supplementary Note S2 describes the error analysis for SDH retrieval.

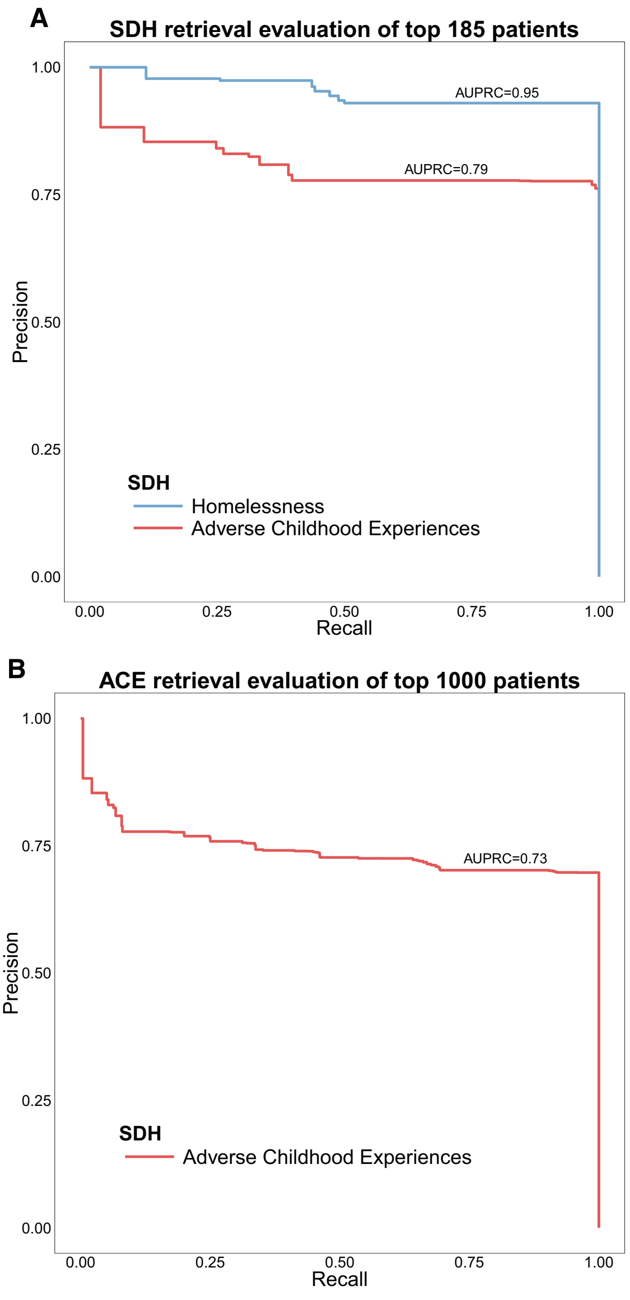

Figure 3. Precision-recall curves for homelessness and ACE evaluation

To gain a deeper insight into the SDH retrieval evaluation, Figure 3A compares the precision-recall curves of all the adult patients from the homelessness case set ($n = 185$) and the top 185 ranked patients for ACE. As observed, this comparative evaluation indicated a superior performance for homelessness identification (AUPRC = 0.95). Nevertheless, ACE identification maintained a high AUPRC value of 73% even when all top 1000 patients were evaluated (Figure 3B). Table 3 lists a more detailed evaluation for the identification of the 2 SDH.

SDH analysis

Table 4 shows the demographic information for all the patients found as relevant for homelessness ($n = 247$) and ACE ($n = 696$). The mean age of homeless and ACE patients was 49 and 28.1 years, respectively; also, most of the homeless patients were men (70.4%), while 67.8% of ACE patients were women.

To examine trends over a long period of time (ie, since they were first recorded in the Synthetic Derivative as homeless or experienced ACE), race and sex distributions of these patients are depicted in Figure 4. A higher number of homeless patients is recorded for the first time in 2005 and 2006 (Figures 4A and 4B), which could be explained in part by the fact that this was the period when comprehensive outpatient visit information was first recorded in the Synthetic Derivative. In contrast, a wider span of ACE events is noticed in Figures 4C and 4D, with the first event recorded in 1936. From these 2 figures, an increasing trend of ACE events is revealed, with the number of ACE men being very close to the number of ACE women over the last decade.

Our analysis of the assessors' comments for ACE indicated that most of the abuse types were sexual and physical (Figure 5A), and the top abuser keywords were "father" and "mother" (Figure 5B). Extraction of ICD-9 codes revealed that the top codes for ACE patients were for mental disorders (Figure 5C), while the ICD-9 code for homelessness, V60.0, was the most prevalent code assigned to 62.8% of patients assessed as relevant for homelessness. Additional information on the descriptive statistics from Figure 5 is provided in Supplementary Tables S8–S11.

A particularly interesting finding is the homelessness status trends across patient visits of the assessed patients for this determinant of health. Figure 6 depicts an alluvial diagram in which blocks represent proportions of patients assessed for a homelessness

Table 3. Evaluation results of homelessness and ACE retrieval

SDH	P@10	P@185	AUPRC (95% CI) Top 185	P@1000	AUPRC (95% CI) Top 1000
Homelessness	1.00	0.93	0.95 (0.92–0.97)	–	–
ACE	0.80	0.76	0.79 (0.73–0.84)	0.70	0.73 (0.70–0.76)

Table 4. Demographics of patients found as relevant for homelessness and ACE

Characteristic	Homelessness	ACE
Total	247	696
Age[a]	49.0 ± 11.9	28.1 ± 17.7
Age at event[a]	42.8 ± 12.1	6.7 ± 3.6
Sex, *n* (%)		
Male	174 (70.4)	224 (32.2)
Female	73 (29.6)	472 (67.8)
Ethnicity, *n* (%)		
White	177 (71.7)	496 (71.3)
African American	63 (25.5)	141 (20.3)
Hispanic	2 (0.8)	29 (4.2)
Other	3 (1.2)	12 (1.7)
Unknown	2 (0.8)	18 (2.6)

[a]Mean ± standard deviation in years

category at a particular patient visit and transitions between blocks represent changes in block composition over consecutive patient visits. For instance, 97 patients were assessed as homeless at their first visit (the H1 block), out of whom, at their second visit, 75 (77.3%) were reidentified as homeless (the H1→H2 transition) and only 3 (3.1%) were found as settled (the H1→S2 transition). Similarly, from the total number of 57 patients labeled as at risk at their first visit (the R1 block), 46 of them were equally split into homeless (|R1→H2| = 23) and at risk (|R1→R2| = 23) at their second visit. Notably, the patients included in this analysis have at least 2 visits in the Synthetic Derivative. As observed, homelessness was chronic in the majority of homeless patients, and cyclic transitions between the at-risk and homeless categories, which may capture the episodic homeless patients, were prominent in this diagram. Moreover, very few transitions from homeless or at risk to settled (and, conversely, from settled to homeless or at risk) were observed. Similar trends are observed for different patient subgroups and number of visits (Supplementary Figures S1 and S2).

DISCUSSION

The big-data approach presented in this study demonstrates the feasibility of systematically capturing and extracting low-prevalence

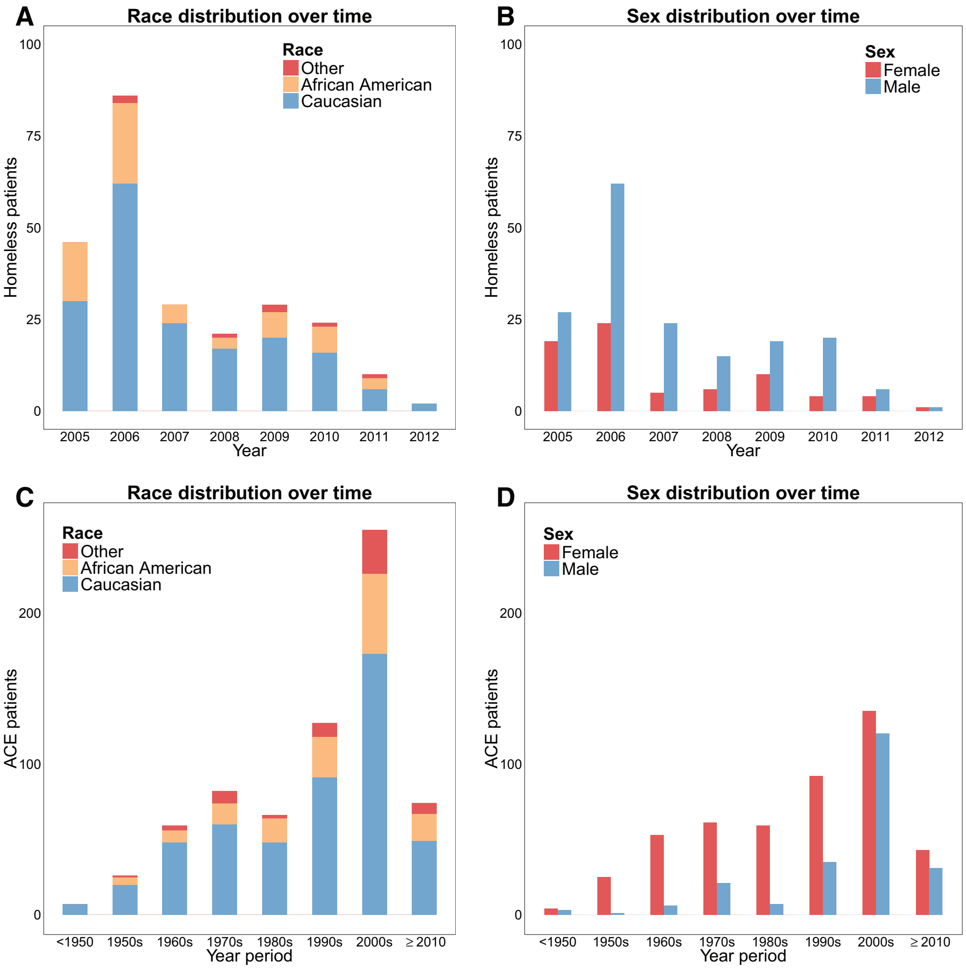

Figure 4. Race and sex distributions over time for the identified homeless and ACE patients

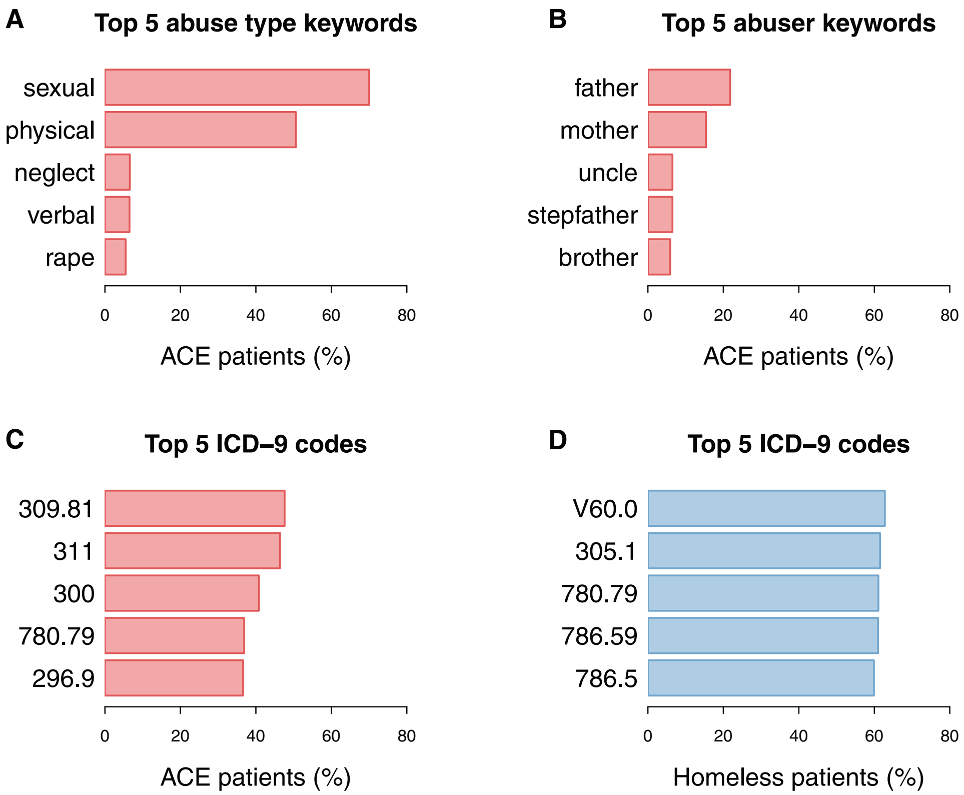

Figure 5. Descriptive statistics of the patients found as relevant for ACE and homelessness

ICD-9 code descriptions: 309.81, Posttraumatic stress disorder; 311, Depressive disorder, not elsewhere classified; 300, Anxiety, dissociative and somatoform disorders; 780.79, Other malaise and fatigue; 296.9, Other and unspecified episodic mood disorder; V60.0, Lack of housing; 305.1, Tobacco use disorder; 780.79, Other malaise and fatigue; 786.59, Other chest pain; 786.5, Chest pain.

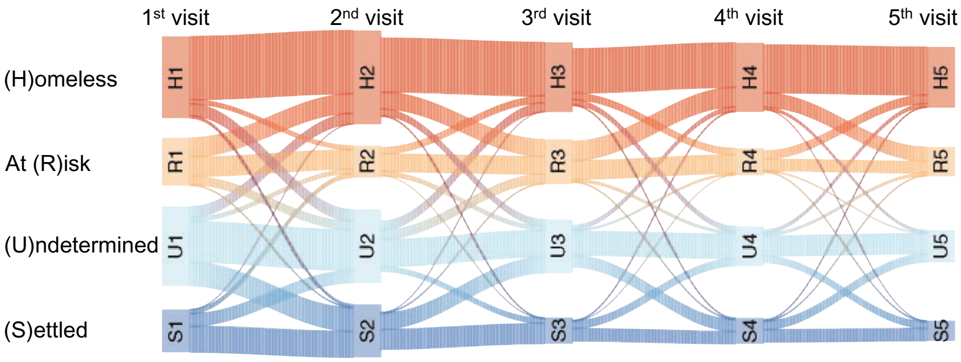

Figure 6. Trends in homelessness status across patient visits

determinants of health from millions of clinical notes. In addition to its main characteristic of scanning large volumes of clinical data in real time to accurately identify a specific phenotype, our system is able to rank all patients by their relevance to the corresponding phenotypic profile. This characteristic is not available for the phenotype algorithms relying on Boolean retrieval approaches through SQL keyword search. Another significant contribution of our approach is the data-driven generation of a phenotypic profile using millions of notes. Large-scale lexicons of relevant concepts could be automatically extracted with high precision and without a significant amount of human effort. Notably, the data-driven approaches we proposed for mining homelessness-relevant keywords are also able to capture keywords specific to the region where the EHR was implemented. For instance, “mission,” which is a vernacular term used for Nashville Rescue Mission, is highly ranked in all the extracted lexicons (Supplementary Tables S1 and S2). This is one relevant example that shows the advantage of employing data-driven approaches over ontology-based methods for phenotype query expansion.

To the best of our knowledge, this is the first study on mining SDH from all the notes of a large-scale EHR repository. The closest study to our work is a text-mining approach using an off-the-shelf search engine to identify ACE patients in a subset of 44.7 million notes from the US Department of Veterans Affairs (VA) data repository.[22] However, information retrieval was used only as an intermediary step in this approach; thus, no final list of ACE query terms or evaluation was provided for the retrieval part. In another VA study,[21] 1000 notes were used to train and test a machine learning system for homelessness identification. While a precision of 94% was achieved on the test set, a system evaluation on 10 000 different notes sampled from the same repository as the training and test sets showed a significant drop in precision to 70%.

As could be observed from the system design, our focus was on extracting the 2 SDH with high precision at the cost of a potential increase in the number of false negatives. Indeed, we suspect that our identified counts of ACE and homeless patients likely underrepresents the true burden of these 2 conditions in the EHR. The reason for this approach was to support subsequent analyses on other clinical and potential genetic associations with ACE and homelessness, where the selection of highly accurate case and control patients is critical for achieving scientifically valid association results. For the initial experiments in this direction, we will select the disease conditions reported by Felitti and colleagues,[29] including ischemic heart disease, any cancer, stroke, chronic obstructive pulmonary disease, and diabetes.

Our work could be reproduced by either implementing the same information retrieval architecture on top of a database server or employing a search engine to index the clinical notes and retrieve the SDH relevant patients. Nevertheless, we acknowledge the limitation on exactly replicating our results (eg, achieving a P@1000 of 70% when using the ACE keywords listed in Table 2) due to various factors that contribute to differences in data distribution across multiple EHR repositories, such as different prevalence values and ways of representing the determinants of health in clinical notes. One domain adaptation methodology that could improve the performance of our system when implemented on a different EHR is to use unsupervised data-driven methods for phenotype query refinement as described in this manuscript.

Unbiased prevalence estimation of the 2 determinants of health requires manual assessment of a large random sample of patients. Since one important objective during the development and evaluation of our system was to minimize the manual assessment effort, a significant time-intensive process that lasted ∼6 months for evaluation of each determinant, the SDH prevalence estimation has been kept out of the scope of our study. For instance, if the prevalence of these social determinants of health is taken into account in the comparative evaluation illustrated in Figure 3A, the performance of the 2 corresponding algorithms could be better discriminated. Nevertheless, an estimate based on the V60.0 code indicated a 0.1% prevalence for homelessness, which is highly underestimated. Figure 5D and Supplementary Table S10 show that out of 247 patients found as relevant for homelessness, only 62.8% had at least one V60.0 code. Furthermore, Gundlapalli and colleagues[21] estimated a V60.0 code–based homelessness prevalence of 0.6% among VA patients. After running a supervised learning classifier on 10 000 randomly selected documents from the VA and manually reviewing the patients relevant for homelessness, they adjusted their prevalence estimate to 3.3%. Assuming a 3.3% homelessness prevalence in the Synthetic Derivative, a baseline system using random sampling will need to retrieve, on average, 5212 patients in order to identify 172 patients relevant for homelessness. As listed in Supplementary Table S5, for the same number of homeless patients, our system needed to retrieve the top 185 patients.

Another limitation of our system is the capability of automatically extracting more specific phenotypic information, such as locations where homeless patients sleep or information about the persons who commit child abuse. For this, human annotations with phenotype-specific information need to be provided in clinical text such that information extraction systems can learn their representation. Nevertheless, the majority of high-performing information extractors for such tasks rely heavily on machine learning technologies and advanced NLP tools that are not scalable to run on large text collections. To make this process feasible, our system could be applied in a prefiltering phase to significantly reduce the number of patients to analyze. For instance, out of the approximately 2.6 million patients in the Synthetic Derivative, our system narrowed down the search space to 35 220 patients (1.3%) and 27 861 patients (1.1%) with high relevance to homelessness and ACE, respectively (Supplementary Table S4).

Motivated by our success in mining 2 different SDH from big EHR data, in future work we plan to investigate the feasibility of our method for identifying additional determinants of health. Further plans to improve our system performance include integrating structured data in the retrieval model (eg, ICD-9 codes) and implementing additional term-weighting schemes (eg, BM25 and InL2c1) and query-expansion methods.

CONCLUSION

Mining SDH from large-scale EHRs is feasible and could be carried out in the context of high-throughput clinical and genetic studies to determine their impact on overall quality of life. This study shows that an unsupervised learning method based on Google's word2vec model is a successful approach for automatically extracting rare phenotypic profiles from millions of clinical notes. Our system was tested in real time to extract SDH relevant patients from >100 million clinical notes. Experimental results indicate that both determinants could be extracted with high precision from clinical notes, although our results suggest that homelessness may be easier to extract than ACE. Future work could expand the applicability of our system to identify additional determinants of health.

CONTRIBUTORS

CAB designed and implemented the system, performed the evaluation and data analysis, and wrote the initial draft of the manuscript. CAB, DC, RN, JP, KBJ, and JCD contributed to the assessment and analysis of homeless patients. CAB, JA, JKS-R, LL, JP, and SB were involved in the assessment and analysis of patients relevant for adverse childhood experiences. RN, RMC, SK, SB, KBJ, and JCD provided critical suggestions and clinical insights into the analysis of social determinants of health in electronic health records. All authors contributed to the final manuscript.

FUNDING

This study was supported in part by grants from the National Library of Medicine (R01 LM010685), the National Institute of General Medical Sciences (R01 GM103859), the National Center for Advancing Translational Sciences (UL1 TR000445), and the Patient-Centered Outcomes Research Institute (CDRN-1306-04869). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and the Patient-Centered Outcomes Research Institute.

COMPETING INTERESTS

None.

ACKNOWLEDGMENTS

We would like to thank Dr William Stead for his comments and suggestions on earlier versions of this paper. We also thank Gabriella Papa for her contribution to the assessment of homeless patients.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Centers for Disease Control and Prevention. *Tobacco-Related Mortality*. 2016. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/tobacco_related_mortality/. Accessed October 2, 2016.
2. Centers for Disease Control and Prevention. *Alcohol Use and Your Health*. 2016. http://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm. Accessed October 2, 2016.
3. Mokdad AH, Marks JS, Stroup DF, *et al*. Actual causes of death in the United States, 2000. *JAMA*. 2004;291:1238–45.
4. Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. *PLoS Med*. 2010;7:e1000316.
5. National Academy of Medicine. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. Washington, DC: National Academies Press; 2014.
6. National Academy of Medicine. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington, DC: National Academies Press; 2014.
7. Hripcsak G, Forrest CB, Brennan PF, *et al*. Informatics to support the IOM social and behavioral domains and measures. *J Am Med Inform Assoc*. 2015;22:921–24.
8. Yu S, Liao KP, Shaw SY, *et al*. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc*. 2015;22:993–1000.
9. Mo H, Thompson WK, Rasmussen LV, *et al*. Desiderata for computable representations of electronic health records–driven phenotype algorithms. *J Am Med Inform Assoc*. 2015;22:1220–30.
10. Lin C, Karlson EW, Canhao H, *et al*. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. *PLoS One*. 2013;8:1–10.
11. Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc*. 2012;19:e162–69.
12. Bejan CA, Xia F, Vanderwende L, *et al*. Pneumonia identification using statistical feature selection. *J Am Med Inform Assoc*. 2012;19:817–23.
13. Wang Y, Chen ES, Pakhomov S, *et al*. Automated extraction of substance use information from clinical texts. *AMIA Annu Symp Proc*. 2015;2015:2121–30.
14. Chen ES, Carter EW, Sarkar IN, *et al*. Examining the use, contents, and quality of free-text tobacco use documentation in the electronic health record. *AMIA Annu Symp Proc*. 2014;2014:366–74.
15. Melton GB, Manaktala S, Sarkar IN, *et al*. Social and behavioral history information in public health datasets. *AMIA Annu Symp Proc*. 2012;2012:625–34.
16. Uzuner O, Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*. 2008;15:14–24.
17. Schanzer B, Dominguez B, Shrout PE, *et al*. Homelessness, health status, and health care use. *Am J Public Health*. 2007;97:464–69.
18. Hwang SW, Dunn JR. Homeless people. In: Galea S, Vlahov D, eds. *Handbook of Urban Health: Populations, Methods, and Practice*. New York: Springer Science & Business Media; 2006.
19. Austin A, Herrick H, Proescholdbell S. Adverse childhood experiences related to poor adult health among lesbian, gay, and bisexual individuals. *Am J Public Health*. 2016;106:314–20.
20. Gundlapalli AV, Redd A, Carter M, *et al*. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. *J Am Med Inform Assoc*. 2013;20:e355–64.
21. Gundlapalli AV, Carter ME, Palmer M, *et al*. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. *AMIA Annu Symp Proc*. 2013;2013:537–46.
22. Hammond KW, Ben-Ari AY, Laundry RJ, *et al*. The feasibility of using large-scale text mining to detect adverse childhood experiences in a VA-treated population. *J Trauma Stress*. 2015;28:505–14.
23. Austin J, McKellar JD, Moos R. The influence of co-occurring axis I disorders on treatment utilization and outcome in homeless patients with substance use disorders. *Addict Behav*. 2011;36:941–44.
24. Birgenheir DG, Lai Z, Kilbourne AM. Datapoints: trends in mortality among homeless VA patients with severe mental illness. *Psychiatr Serv*. 2013;64:608.
25. Zech J, Husk G, Moore T, *et al*. Identifying homelessness using health information exchange data. *J Am Med Inform Assoc*. 2015;22:682–87.
26. Salit SA, Kuhn EM, Hartz AJ, *et al*. Hospitalization costs associated with homelessness in New York City. *N Engl J Med*. 1998;338:1734–40.
27. Roden DM, Pulley JM, Basford MA, *et al*. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther*. 2008;84:362–69.
28. Uzuner Ö, South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011;18:552–56.
29. Felitti VJ, Anda RF, Nordenberg D, *et al*. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The Adverse Childhood Experiences (ACE) Study. *Am J Prev Med*. 1998;14:245–58.
30. Kushel MB, Vittinghoff E, Haas JS. Factors associated with the health care utilization of homeless persons. *JAMA*. 2001;285:200–06.
31. Levy BD, O'Connell JJ. Health Care for Homeless Persons. *N Engl J Med*. 2004;350:2329–32.
32. O'Connell J. *Premature Mortality in Homeless Populations: A Review of the Literature*. Nashville, TN: National Health Care for the Homeless Council; 2005. http://www.nhchc.org/wp-content/uploads/2011/10/Premature-Mortality.pdf. Accessed October 3, 2016.
33. Hibbs JR, Benner L, Klugman L, *et al*. Mortality in a cohort of homeless adults in Philadelphia. *N Engl J Med*. 1994;331:304–09.
34. Heim C, Nemeroff CB. The role of childhood trauma in the neurobiology of mood and anxiety disorders: preclinical and clinical studies. *Biol Psychiatry*. 2001;49:1023–39.
35. Chaitanya Shivade, Preethi Raghavan, *et al*. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2013;21:221–30.
36. Kirby JC, Speltz P, Rasmussen LV, *et al*. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc*. 2016;23:1046–52.
37. Hripcsak, Albers. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2012;20:117–21.
38. Agarwal V, Podchiyska T, Banda JM, *et al*. Learning statistical models of phenotypes using noisy labeled training data. *J Am Med Inform Assoc*. 2016;23:1166–73.
39. National Health Care for the Homeless Council. *What Is the Official Definition of Homelessness?* 2016. https://www.nhchc.org/faq/official-definition-homelessness/. Accessed October 2, 2016.
40. National Alliance to End Homelessness. *Changes in the HUD Definition of "Homeless."* 2012. http://www.endhomelessness.org/library/entry/changes-in-the-hud-definition-of-homeless. Accessed October 2, 2016.
41. Bejan CA, Nash R, Conway D, *et al*. Mining phenotypic keywords from a large collection of clinical narratives. In: *AMIA Jt Summits Transl Sci Proc*. 2015;242–43.
42. Zamani H, Croft WB. Embedding-based query language models. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. 2016;147–56.
43. Diaz F, Mitra B, Craswell N. Query expansion with locally-trained word embeddings. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016;367–77.

44. Sordoni A, Bengio Y, Nie J-Y. Learning concept embeddings for query expansion by quantum entropy minimization. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press 2014;1586–92.
45. Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. *ICLR*. 2013;1–12.
46. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. *NIPS*. 2013;3111–19.
47. Middleton C, Baeza-Yates R. *A Comparison of Open Source Search Engines*. [Online]. 2016. http://wrg.upf.edu/WRG/dctos/Middleton-Baeza.pdf. Accessed October 2, 2016.
48. Rappoport A. *Open Source Search Engines*. [Online]. 2012. http://www.searchtools.com/tools/tools-opensource.html. Accessed October 2, 2016.
49. Chapman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34:301–10.
50. Bejan CA, Vanderwende L, Xia F, *et al*. Assertion modeling and its role in clinical phenotype identification. *J Biomed Inform*. 2013;46:68–74.
51. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. New York: Cambridge University Press; 2008.
52. Efron B. Bootstrap Methods: Another Look at the Jackknife. *Ann Stat*. 1979;7:1–26.
53. Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. In: *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer; 2013:451–66.