Large-Scale Text Mining of Social Determinants from Electronic Health Records: Case Studies of Homelessness and Adverse Childhood Experiences

Cosmin A. Bejan, PhD[1], John Angiolillo, MD[2], Douglas Conway, BA[3], Robertson Nash, PhD, ACNP[2], Jana K. Shirey-Rice, PhD[3], Loren Lipworth, ScD[2], Robert M. Cronin, MD, MS[1,2,4], Jill Pulley, MBA[3], Sunil Kripalani, MD, MS[2], Shari Barkin, MD, MSHS[4], Kevin B. Johnson, MD, MS[1,4], Joshua C. Denny, MD, MS[1,2]

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

Introduction

Social determinants of health (SDH) have recently received a growing interest as they provided evidence for being associated with an early onset and progression of various diseases, and a risk of premature death.[1] This interest was also reflected in a report from the National Academy of Medicine emphasizing the importance of capturing SDH in Electronic Health Records (EHRs).[2] The objective of this study was to implement a big data approach for extracting two severe SHD, homelessness and adverse childhood experiences (ACE), from a large dataset of notes.

Methods

Figure 1 illustrates the architecture of our approach.[3] Using the Vanderbilt de-identified version of the EHR, which includes >100 million notes, we integrated unsupervised learning and information retrieval (IR) methods to capture the two SDH profiles and to identify the patients that best matched these profiles. First, we built a list of query terms for each profile. Since ACE is a complex phenotype with experiences ranging from sexual to psychological abuse, we relied on domain expertise to identify the initial query of this SDH. For homelessness, we employed two data-driven methods, lexical association[4] and word2vec,[5] to expand two seed keywords, *homelessness* and *homeless*. We trained word2vec on 10 million notes randomly sampled from the EHR and used a vector dimension of 100. Next, we computed the similarity between a query and the notes of each patient by using an IR vector space model architecture based on the standard TF-IDF weighted cosine metric. We implemented the IR model on top of our EHR database, a secure IBM Netezza data warehouse consisting of a parallel computing architecture, to ensure the scalability of our system. Each search over the entire EHR lasted <20 min. After retrieval, negation detection[6,7] was employed to exclude the patients with a high prevalence of negated terms. Finally, the top retrieved results were manually examined by seven assessors. Query reformulation based on relevance feedback was adopted by analyzing the top 20 results of each retrieval. For homelessness query expansion, the most relevant terms were selected from the top 50 keywords generated by the data-driven methods. After each query was settled, the final retrieval was executed and the top ranked patients were manually assessed. Specifically, for ACE, the top 1,000 patients were categorized as either ACE, not ACE, or undetermined. For homelessness, the assessment was performed at visit level using five categories: homeless, settled, at risk, undetermined, and pediatric. Since our next objective is to investigate associations between homelessness and various diseases using a case-control study design, for assessment, 600 patients were selected from 3 distinct sets: 1) the case set contained 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 don't have homelessness terms in their notes. The evaluation was performed using precision-recall curves, the precision of top k ranked results (P@k), and the area under the precision-recall curve (AUPRC). The 95% confidence intervals (CI) of AUPRC estimators were computed using a bootstrap procedure.

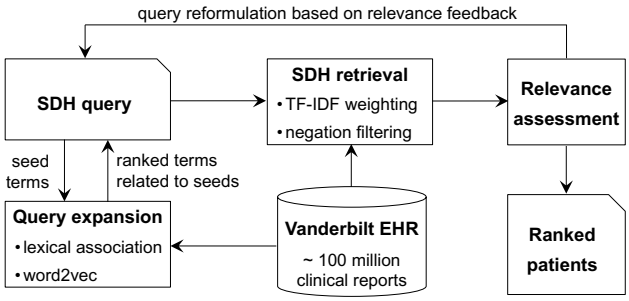

Figure 1 System architecture for SDH extraction.

Results

Lexical association and word2vec proved to be viable methods for homelessness query expansion. Table 1 lists the top 8 ranked keywords relevant for this determinant. However, a comparative study on the top 50 keywords extracted by the two methods revealed superior performance values achieved by word2vec (Figure 2 and Table 2). As a result of query expansion and reformulation, the final query terms used for retrieval assessment were 1) for homelessness: *homeless*, *homelessness*, *shelter*, *unemployed*, *jobless*, and *incarceration*; and 2) for 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*, and *cps investigation*. Based on these terms, the system retrieved 35,220 (1.3%) and 27,861 (1.1%) patients for homelessness and ACE, respectively. For both SDH, a substantial inter-rater agreement (kappa=0.8) was achieved. Out of the top 200 patients of the homelessness retrieval, 15 were pediatric, and 172 were found as relevant for this determinant (i.e., they have at least one visit assessed as relevant for homelessness). As listed in Table 3, the homelessness precision of the top 185 adult patients was 93% (172/185). This is substantially higher when compared with the precision of 40% achieved on the fuzzy set. For ACE, the system obtained a precision of 70% for the top 1,000 ranked patients. A comparative evaluation of the top 185 patients retrieved for the two SHD (Figure 3) indicated a higher AUPRC of 94% for homelessness identification (see details in Table 3). An analysis of the ACE patients revealed that most adverse experiences included sexual and physical abuse and top prevalent abusers were father and mother (Table 4). An alluvial diagram of the patients assessed for homelessness across consecutive visits (Figure 4) indicated that this condition was chronic for most of the homeless patients. The prominent cyclic transitions between the at risk and homeless categories captured an episodic trend of homeless patients.

Lexical association	word2vec
shelter	homeless
unemployed	polysubstance
shelters	methamphetamine
social	sober
lives	schizo-affective
ambulation/mobility	schizophrenia
jobless	prostitute

Table 1 Homelessness related words.

Method	P@10	P@50	AUPRC (95% CI)
Lexical assoc.	0.80	0.48	0.83 (0.70–0.95)
word2vec	1.00	0.82	0.94 (0.89–0.98)

Table 2 Query expansion evaluation.

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

Table 3 SDH retrieval evaluation (top 185).

Abuse type		Abuser	
sexual	70.0%	father	21.8%
physical	50.6%	mother	15.4%
neglect	6.6%	stepfather	6.5%
verbal	6.5%	uncle	6.5%
rape	5.5%	brother	5.9%
emotional	5.3%	parents	4.6%
child	2.2%	cousin	4.3%

Table 4 ACE assessment analysis.

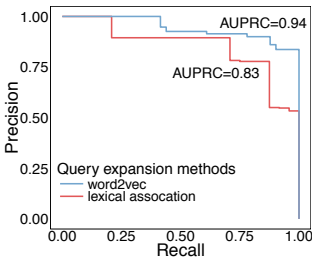

Figure 2 Homelessness query expansion evaluation.

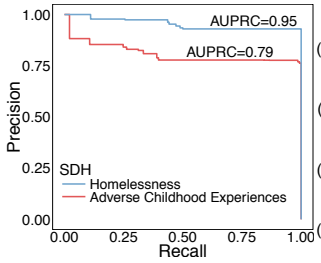

Figure 3 Homelessness and ACE retrieval evaluation.

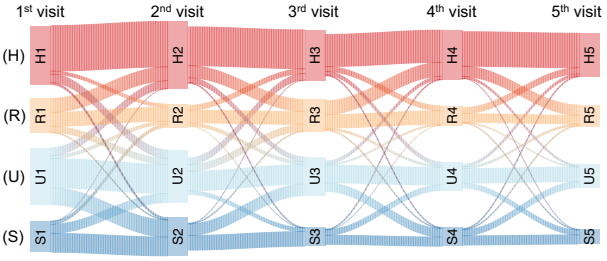

Figure 4 Trends over homelessness categories: (H)omelessnes, At (R)isk, (U)ndetermined, and (S)ettled.

Discussion

We described a feasible solution for mining SDH from a large-scale EHR, which opens the way for conducting large phenome wide association studies[8] on the impact of these determinants on the overall quality of life. To the best of our knowledge, this is the first study on SDH text mining from all the notes of a big EHR repository. Our system performance could be improved by integrating structured data in the retrieval model. However, the only structured data we could use from our EHR is the V60.0 administrative code for homelessness, which is not sufficient for extracting this determinant.[9] In this study, we found that only 62.8% of patients assessed as relevant for homelessness have the V60.0 code. Future work could expand the applicability of our system to identify additional determinants of health.

References

1 Braveman P, Gottlieb L. The Social Determinants of Health: It's Time to Consider the Causes of the Causes. *Public Health Rep* 2014.

2 National Academy of Medicine. *Capturing Social and Behavioral Domains in Electronic Health Records* 2014.

3 Bejan CA, *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* 2017.

4 Bejan CA, *et al.* Mining phenotypic keywords from a large collection of clinical narratives. *AMIA Jt Summits Transl Sci Proc* 2015.

5 Mikolov T, *et al.* Distributed Representations of Words and Phrases and their Compositionality. *NIPS* 2013.

6 Chapman WW, *et al.* A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001.

7 Bejan CA, *et al.* Assertion modeling and its role in clinical phenotype identification. *J Biomed Inform* 2013.

8 Denny JC, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013.

9 Gundlapalli A, *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.