Mining phenotypic keywords from a large collection of clinical narratives

Cosmin A. Bejan, PhD[1], Robertson Nash, PhD(c), ACNP, BC[2], Douglas Conway, BA[3], Erica A. Bowton, PhD[3], 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

Abstract: We describe a data-driven approach for extracting phenotype-related keywords, applied to the particular example of terms related to "homelessness." The algorithm runs over millions of clinical documents and requires minimal domain knowledge about the phenotype of interest. Our evaluation shows substantial improvements of this algorithm over a baseline using a lexical association measure.

Introduction and Background: Clinical phenotype algorithms often rely on natural language processing technologies to accurately identify cohorts of patients.[1,2] This is because relevant clinical information for a specific phenotype is often found only in narrative form or is not entirely captured by structured data. However, to integrate this information into phenotypic algorithms, domain experts need to be involved in creating phenotype-specific lexicons,[3] a process which typically demands substantial human effort. Our purpose was to automatically extract such a lexicon from the entire electronic medical record (EMR) using an algorithm which requires minimal domain knowledge about the phenotype of interest.

Method: The main steps of the extraction procedure are listed in Algorithm 1. Starting with a small set S of seed keywords representative for the phenotype of interest, the algorithm's scope is to identify those salient words which tend to occur more often than chance in the context of a seed keyword. In step 1, we extracted the set D of documents with at least one word from S. Next, using the documents in D, we constructed the set of candidate keywords N from the words at a distance of no more than $\delta/2$ from any seed in S. In steps 3–6, for each word w in N, we computed a contingency table by checking how many times w occurred in the context of a seed keyword. Finally, based on its contingency table, we computed the salience of w using the formula described in step 8. Here, AM denotes an association measure, $freq(w)$ is the frequency count of w, and λ is a parameter which controls the degree to which the inverse word frequency of w should weight in the salience score. An increased λ will correspond to an increased salience score of the less frequent words. For λ=0, the salience order of words is identical with the order of the association measure.

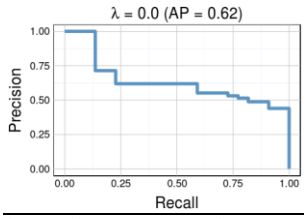
Input: S – set of seed keywords, δ – word context size, λ – inverse word frequency weight
Output: a set of words related to the seeds in S
Notation: $\Delta(w, d, \delta)$ – set of words from document d in the vicinity of no more than $\delta/2$ distance from the word w in d

1 extract D – documents with at least one seed
2 build $N = \bigcup_{d \in D, s \in S} \Delta(s, d, \delta)$
3 **foreach** $d \in D$ **do**
4 **foreach** $w \in d$ **do**
5 **if** $w \in N$ **then**
6 $\forall s \in S$ test if $s \in \Delta(w, d, \delta)$ and update the contingency table associated with w
7 **foreach** $w \in N$ **do**
8 $score(w) = \log(AM(w, S)/(1 + freq(w))^{\lambda})$
9 **return** the top words in N sorted by $score$

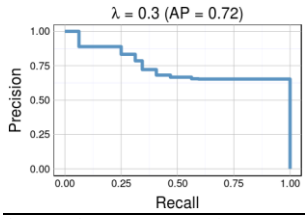
Algorithm 1 Phenotypic keywords extraction.

Evaluation: We ran the algorithm over >77 million documents from the Synthetic Derivative, a de-identified version of the Vanderbilt EMR, to extract keywords related to homelessness. First, we tokenized each document; next, we converted the tokens to lowercase, discarded punctuations and tokens with *freq*<5. In our experiments, we selected only two seed keywords: *homeless* and *homelessness*. After running step 1, we extracted 22,260 documents with at least one seed. For measuring the association between the candidate words and the two seeds, we used the χ^2 test, *t*-test, Fisher exact test, Dice coefficient, and pointwise mutual information. Our initial analysis of the results showed that the best results were achieved when employing the χ^2 test. To evaluate the keywords, first we ran the algorithm using all possible configurations of the χ^2 test, $\delta \in \{10,20,40,60\}$, and $\lambda \in \{0.0,0.1, \ldots,1.0\}$; then, a board certified clinician manually annotated the top 50 terms from each word set for relevance to homelessness.

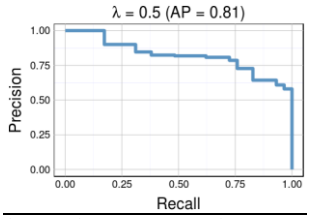
Results and Conclusion: Table 1 lists four such sets represented by the top 10 most salient words (parameter settings: χ^2 test, $\delta = 40$, $\lambda \in \{0.0,0.3,0.5,1.0\}$). In the table, the words emphasized in bold were marked as relevant for homelessness. For instance, *social* and *history* were highly ranked because most of the homelessness related information is described in the 'social history' section. Also, *mission* is a vernacular term used for "Nashville Rescue Mission." In Table 1, we also plot the precision-recall curves associated with the corresponding lists of annotated keywords. The best result (average precision of 0.81) for the experiments using a word context size of 40 was obtained when λ=0.5. In general, for all the experiment sets grouped by δ, the best results were achieved when the inverse word frequency, which will deemphasize frequent words, played a role in computing the keyword salience score (i.e., when λ>0). Thus, these results proved that, for the task of phenotypic keyword extraction, our algorithm outperforms a baseline system using only an association measure to compute the salience scores.



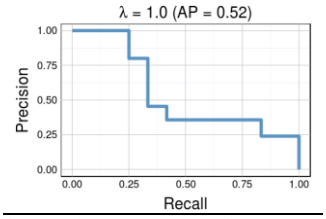
keyword	freq	score
social	24577	10.41
shelter	2504	9.92
history	88384	9.92
currently	13942	9.85
is	184677	9.60
lives	4686	9.39
living	8856	9.36
he	211750	9.35
review	15185	9.27
mission	3769	9.07



keyword	freq	score
shelter	2504	7.57
social	24577	7.38
currently	13942	6.99
lives	4686	6.86
unemployed	1605	6.75
living	8856	6.63
mission	3769	6.60
history	88384	6.50
shelters	445	6.41
review	15185	6.38



keyword	freq	score
shelter	2504	6.01
social	24577	5.36
unemployed	1605	5.27
shelters	445	5.19
lives	4686	5.17
currently	13942	5.08
mission	3769	4.95
living	8856	4.82
staying	1769	4.80
jobless	85	4.75



keyword	freq	score
trauma/burns	32	2.69
prisoners	25	2.68
estranged-father	24	2.68
use/htn	22	2.68
depression/etoh	22	2.68
legpain in right leg	21	2.67
***id-num**med4on*	20	2.67
spouse-alive	20	2.67
mutilple	18	2.67
cvah-o-cva-9on	17	2.66

Table 1 Results for extracting homelessness related keywords. Algorithm configuration: association measure = χ^2 test, $\delta = 40$, $\lambda \in \{0.0,0.3,0.5,1.0\}$. AP, average precision, freq, frequency count; score, salience score.

References

1 Bejan CA, Xia F, Vanderwende L, *et al.* Pneumonia identification using statistical feature selection. *J Am Med Inform Assoc* 2012;**19**:817–23.

2 Carroll R, Thompson W, Eyler A, *et al.* Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;**19**:e162–e169.

3 Gundlapalli A, 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–e364.