ISDS

Emerging Health Threats Journal 2011, **4**:s112. **doi:** 10.3134/ehtj.10.112

www.eht-journal.org

ABSTRACT

Identification of features for detection and prediction of homelessness from VA clinical documents

S Shen[1,2], BR South[1,2], M Palmer[1], S Duvall[1], M Samore[1,2], and AV Gundlapalli[1,2]

[1]*Salt Lake City VA Health Care System, Salt Lake City, UT, USA; and* [2]*University of Utah, Salt Lake City, UT, USA*
E-mail: shuying.shen@hsc.utah.edu

Objectives
We demonstrate a semi-automated approach to induce and curate lexical domain knowledge for identification of evidence and risk factors for homelessness found in VA clinical documents. This domain knowledge can be used to support training and evaluation of automated methods such as Natural Language Processing (NLP) systems for detection and prediction of homelessness among veterans. This could serve as a proxy for public health and other surveillance involving homeless individuals. Similar methods could be used to identify other conditions of interest.[1]

Introduction
Homelessness in general is a major issue in the US today. The risk factors of homelessness are myriad, including inadequate income, lack of affordable housing, mental health and substance abuse issues, lack of social support, and non-adherence to treatment/follow-up appointments. Early identification of these factors from clinical documents may help detect or even predict homelessness cases, allowing adequate intervention and prevention measures.

Methods
Using a think out loud approach, we developed an initial lexicon of features related to homelessness using expert inputs and available literature sources. This lexicon consists of social stressors (that is, recent divorce, unemployment), behavioral factors (that is, drug abuse), evidence (that is, lives in shelter, no housing), other risk factors (that is, exposure to war-related trauma) and direct mention of homelessness in the medical record (that is, homeless patient). This initial list was applied as pre-annotations to 600 VA clinical documents extracted from the VA Region one and four Data Warehouse for the time period 1/1/200-12/31/2009. Documents were pre-annotated using a prototype system that supports interactive annotation and semi-automated curation of user-defined information classes.

Refining the Lexicon
Domain experts reviewed pre-annotated documents to determine if information was correctly identified, make modifications to annotations, add missing annotations, or reject annotations found to be incorrect or irrelevant. We applied an iterative process of revising the lexicon until further refinements were exhausted.

Results
Our initial lexicon had 83 entries. After two rounds of semi-automated curation on 75 documents, 38 concepts were added. Pre-annotations were helpful for reviewers to focus attention around the surrounding context, revealing important cues and textual patterns that would inform guideline development and creation of a reference standard for NLP system implementation.

Conclusions
Our approach can effectively generate lexical domain knowledge combining information from literature and expert feedback via iterative refinement. This method could be easily adapted to other surveillance efforts for case identification and prediction.

Acknowledgements
This paper was presented as an oral presentation at the 2010 International Society for Disease Surveillance Conference, held in Park City, UT, USA, on 1–2 December 2010.

Reference
1 Rosenheck R, Fontana A. A model of homelessness among male veterans of the Vietnam War generation. *Am J Psychiatry* 1994;**151:** 421–7.