Proceedings of the Society for Academic Emergency Medicine Annual Meeting (SAEM) 2002.
Academic Emergency Medicine 2002;9(5):511.

Identifying Medical Concepts in Free Text Chief Complaint Data

Debbie A Travers and Olivier Bodenreider

University of North Carolina: Chapel Hill, NC, National Library of Medicine: Bethesda, MD

ABSTRACT

Objectives: A controlled vocabulary for chief complaint (CC) will facilitate clinical guidelines, emergency department (ED) surveillance, and third-party claims review. In an earlier pilot study, automated efforts to develop a CC vocabulary were conducted using natural language processing (NLP) of free text ED CC terms. The resulting standardized terms were then compared to the controlled vocabularies in the Unified Medical Language System (UMLS); only 14% of the ED CC free text terms matched a UMLS concept. The objective of this study was to apply and evaluate a combination of automated and manual methods for processing ED CC terms, in order to improve the match rate with UMLS concepts. **Methods:** In this retrospective secondary data analysis, we used corpus linguistics methods to collocate free text CC data from 3 university tertiary referral center EDs (rural, urban, and suburban) for all ED visits during 1/01. First, we evaluated how many CC terms matched a UMLS concept prior to any processing. The terms that did not match a UMLS concept were then manually examined by the investigator, who used domain knowledge to identify patterns in the data. NLP techniques were then applied, moving from simple to more aggressive techniques. The resulting CC terms were again compared to the UMLS to determine the match rate with standard concepts. **Results:** There were 6,900 visits, and 6,054 unique CC terms recorded during the 1/01. Prior to any processing, 22% of the terms matched a UMLS concept. 3 patterns in the data were identified: punctuation (N/V/D), acronyms (BRBPR), and modifiers (mild/moderate/severe). After addressing the patterns with NLP techniques, 35% of the remaining CC terms then matched a UMLS concept. **Conclusions:** We increased the match rate of CC terms with UMLS concepts with a combination of manual and automated techniques. Additional review by domain experts and further NLP are needed to identify concepts for the nonmatched ED CC terms. This study provides the foundation for a controlled ED CC vocabulary.