COACTION

Tuning a chief complaint text parser for use in DoD ESSENCE

Kenneth Cox[1]*, Howard Burkom[2] and Yevgeniy Elbert[2]

[1]Armed Forces Health Surveillance Center, Silver Spring, MD, USA; [2]Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA

Objective

Rather than rely on diagnostic codes as the core data source for alert detection, this project sought to develop a chief complaint (CC) text parser to use in the U.S. Department of Defense (DoD) version of the Electronic Surveillance System for Early Notification of Community-Based Epidemics (ESSENCE), thereby providing an alternate evidence source. A secondary objective was to compare the diagnostic and CC data sources for complementarity.

Introduction

An expanded ambulatory health record, the Comprehensive Ambulatory Patient Encounter Record (CAPER) will provide multiple types of data for use in DoD ESSENCE. A new type of data not previously available is the reason for visit (ROV), a free-text field analogous to the CC. Intake personnel ask patients why they have come to the clinic and record their responses. Traditionally, the text should reflect the patient's actual statement. In reality, the staff often 'translates' the statement and adds jargon. Text parsing maps keywords or phrases to specific syndromes. Challenges exist given the vagaries of the English language and local idiomatic usage. Still, CC analysis by text parsing has been successful in civilian settings (1). However, it was necessary to modify the parsing to reflect the characteristics of CAPER data and of the covered population. For example, consider the shock/coma syndrome. Loss of consciousness is relatively common in military settings due to prolonged standing, exertion in hot weather with dehydration, etc., whereas the main concern is shock/coma due to infectious causes. To reduce false positive mappings, the parser now excludes terms such as syncope, fainting, electric shock, road march, parade formation, immunization, blood draw, diabetes, hypoglycemic, etc.

Methods

First, a set of syndromes in an existing JHU-APL CC parser used in civilian versions of ESSENCE were selected for evaluation. The CC parser was then used to categorize the ROV from 3 months of records from all DoD facilities (about 12 million records total). From the records matched to each syndrome, 2000 were selected; the 1000 most common strings and an additional 1000 strings at random. Two analysts evaluated the sample strings independently and identified key decisions they each used to decide whether the match between the text string and the syndrome was accurate. They then attempted to reconcile those cases where they disagreed. Unresolved differences were reviewed by a DoD consultant who offered revised rules based on clinical experience and known practice patterns among DoD providers. The modified rules were incorporated into the CC parser and tested on miniature data samples for accuracy, and then the CC parser was rerun on the full record set. The cycle of independent analysis and review was repeated with additional modifications to correct any remaining errors followed by a final full run. Before and after contingency tables were used to compare CC-based versus diagnosis-based classifications. The final product was a modified CC parser input file tailored for use with DoD ambulatory healthcare records.

Results

The following before/after contingency Table 1 illustrates the results for two syndromes: one common, influenza-like illness (ILI), and one rarer syndrome, neurological (Neuro). Additional results will show effects of the CC noise removal on alerting using datasets with documented outbreaks.

Conclusions

This iterative method produced a CC parser with substantially improved performance, eliminating many obvious incorrect classifications and resulting in a smaller number of more meaningful alerts for public health investigation. The process could be used to tune parsers to meet the unique medical terminology used in different communities. Moreover, daily CC and diagnostic counts should not be crudely pooled, but do provide complementary views of the population health status.

Keywords

Text parser; chief complaint; syndrome; surveillance

References

1. Sniegoski CA. Automated syndromic classification of chief complaint records. Johns Hopkins APL Tech Dig. 2004;25:68–75.

*Kenneth Cox
E-mail: Kenneth.Cox7@us.army.mil

Table 1

Syndrome		CC Yes	CC No	Ratio CC only to ICD only
ILI before	ICD Y	194,957	1,382,852	0.6
	ICD N	762,842	38,904,014	
ILI after	ICD Y	182,601	1,395,208	0.4
	ICD N	595,013	39,071,844	
Neuro before	ICD Y	74	2,548	63.7
	ICD N	162,195	41,079,848	
Neuro after	ICD Y	10	2,612	16.0
	ICD N	41,717	41,200,327	

Abbreviations: ILI, influenza-like illness; neuro, neurological; Y, yes; N, no.

(page number not for citation purpose)