Using UMLS Semantic Network to Identify Search Terms for Biosurveillance

Debbie Travers, PhD, RN[1]; Matthew Scholer, MD, PhD[1];
John Crouch,BS[1]; Scott Wetterhall, MD, MPH[2]
[1]University of North Carolina at Chapel Hill, [2]RTI International

Objective

This paper describes and applies a new method for identifying biosurveillance search terms using the Semantic Network® of the Unified Medical Language System® (UMLS®)[1].

Background

The variability of free text emergency department (ED) data is problematic for biosurveillance [2], and current methods of identifying search terms for symptoms of interest are inefficient as well as time- and labor-intensive. Our ad hoc approach to term identification for the North Carolina Disease and Epidemiologic Collection Tool (NC DETECT) begins with development of clinical case definitions from which we build automated syndrome queries in standard query language (SQL). The queries are used to search free text clinical data from EDs, with the goal of identifying free text terms to match the case definitions. The free text search terms were initially collected from epidemiologists and clinical and technical staff at NC DETECT through informal review of ED data. Over time, we reviewed individual cases missed by our queries and identified additional search terms. We also manually reviewed records to find misspellings, abbreviations and acronyms for known search terms (e.g., *dypnea*, *diff. br*. and *SHOB* for **dyspnea**), and developed a pre-processor to clean text prior to syndromic classification.[3] The purpose of this project was to develop and test a more standardized approach to search term identification.

Methods

We used the UMLS Semantic Network to identify clusters of similar terms to add to our syndrome queries. The network organizes all UMLS concepts (and corresponding terms) into semantic categories (e.g., sign or symptom, body part, pathological function) with defined relationships (e.g., *antibiotic* is a pharmacologic substance)[1]. Related concepts are grouped as synonyms, narrower or broader.

We began by extracting all search terms from our current respiratory (RESP), fever rash (FR), and gastrointestinal (GI) syndrome queries. We then used the Semantic Network table MRREL to identify synonymous (e.g., *pyrexia, elevated temperature*), narrower (e.g., *febrile convulsions*) and broader (e.g., *body temperature*) terms for each of our search terms (e.g., **fever**). Two clinical experts then independently reviewed each synonymous, narrower and broader term to identify those to be added to the syndrome queries. For any records in which the first two reviewers differed, a third senior expert adjudicated.

Results

The new method generated 734 potential terms for RESP, 1,272 for FR and 756 for GI. Of those, the experts determined that 69 new RESP terms were appropriate to add to the respiratory query (49 synonymous, 10 narrower and 10 broader). In Table 1 are examples of respiratory terms reviewed.

Table 1- Example- UMLS Terms Reviewed

Original Search Term	Synonymous Concept	Narrower Concepts	Broader Concepts
Cyanosis	Selected	Selected	Selected
	Cyanosed	Blue lips	Skin color change
	Rejected	Rejected	Rejected
	(none)	Purple glove syndrome	Diagnostic findings

The review of terms for FRI and GI is ongoing. We are also conducting an analysis of the accuracy of the syndrome queries with the new terms included. The results of the revised queries will be compared to our gold standard which includes 10,692 records reviewed by a panel of clinical experts.[4]

Conclusions

The Semantic Network contains useful tools for identifying search terms for biosurveillance. Through systematic application of these tools, we supplemented our list of search terms that had been developed using ad hoc methods.

References

[1] National Library of Medicine (2006). Fact Sheet: UMLS Semantic Network. Retrieved July 30, 2008, from http://kswebp1.nlm.nih.gov/uPortal/frame.jsp?myvt-frame=http://www.nlm.nih.gov/research/umls/documentation.html

[2] Shapiro, A. (2004). "Taming variability in free text: Application to health surveillance." MMWR **53S**: 95-100

[3] Travers D, Wu S, Scholer M, Westlake M, Waller A, McCalla A. (2007). Evaluation of a chief complaint processor for biosurveillance. *Proceedings of the 2007AMIA Symposium*: 736-740.

[4] Scholer M, Ghneim G, Wu S, Westlake M, Travers D, Waller A, McCalla A, Wetterhall S. (2007). Defining and applying a method for improving sensitivity and specificity of an emergency department early event detection system. *Proceedings of the 2007AMIA Symposium*: 651-655.

Acknowledgments

We thank Shiying Wu, Matt Westlake, Anne-Lyne McCalla, and Anna Waller. Funded by CDC #R01-PH000038-01. Further Information: Debbie Travers, dtravers@email.unc.edu