Emerging Health Threats Journal 2011, **4**:s101. **doi:** 10.3134/ehtj.10.101

www.eht-journal.org

ABSTRACT

Building an automated Bayesian case detection system

F-C Tsui, J Espino, T Sriburadej, H Su, and J Dowling

RODS Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
E-mail: tsui2@pitt.edu

Objective

This paper describes the architecture and evaluation of our recently developed automated Bayesian case detection (BCD) system.

Introduction

Current practices of automated case detection fall into the extremes of diagnostic accuracy and timeliness. In regards to diagnostic accuracy, electronic laboratory reporting (ELR) is at one extreme and syndromic surveillance is at the other. In regards to timeliness, syndromic surveillance can be immediate, and ELR is delayed 7 days from initial patient visit.[1]

A plausible solution, a middle way, to the extremes of diagnostic precision and timeliness in current case detection practices is an automated Bayesian diagnostic system that uses all available data types, for example, freetext ED reports, radiology reports, and laboratory reports. We have built such a solution—BCD. As a probabilistic system, BCD operates across the spectrum of diagnostic accuracy, that is, it outputs the degree of certainty for every diagnosis. In addition, BCD incorporates multiple data types as they appear during the course of a patient encounter or lifetime, with no degradation in the ability to perform diagnosis.

Methods

The BCD system that we built has five components: real-time HL7 parsers, natural language processing (NLP) tools, Bayesian inference engine, a Bayesian network, and a database. The HL7 message parsers extract different data types from HL7 messages. Then, NLP tools, MedLEE,[2] and Topaz (a homegrown tool) find medical terms contained in each freetext report, including significant negative findings. We store the NLP results in a database. For non-freetext reports such as laboratory reports, we store the coded data directly to the database.

We built a Bayesian network with 57 nodes for detection of three diseases: flu, shigellosis, and measles. We formed the network structure and conditional probabilities by consulting a physician board certified in infectious diseases. We implemented the Bayesian inference engine in Java using the junction-tree algorithm. To make the BN portable, each node in the network is represented in an Unified Medical Language System (UMLS) Concept Unique Identifier.

We performed a preliminary evaluation of the BCD system using only freetext ED reports as input to detect influenza cases. The gold standard was laboratory-confirmed positive and negative reports.

Results

Our preliminary evaluation used 363 randomly selected reports (181 positive) from 12 January 2005 to 31 August 2007. We found an area under ROC curve of about 0.8 (95%CI: 0.76–0.85). When the posterior probability threshold was set to 0.8 (given $P(\text{flu}) = 0.1$), we found a sensitivity of 63.5% (95%CI: 56.5–70.5%), a specificity of 82.4% (95%CI: 76.9–87.9%), and a positive predictive value of 78.2% (95%CI: 71.5–84.9%).

We used BCD to estimate daily expected counts of flu cases presented in seven EDs of UPMC health system during H1N1 outbreak by summing the posterior probability for flu for each visit. Figure 1 shows chart of percent daily expected ED flu visits from July to December of 2009. Average daily ED visits was 569.

A demo web page of BCD is available at https://betaweb.rods.pitt.edu/casedetection-rest/demoPage.jsp. It demonstrates multiple data types' input for computing posterior probabilities of three diseases.

Conclusions

Our BCD system has good performance characteristics and is a solution to low diagnostic accuracy and timeliness in existing automated surveillance systems.

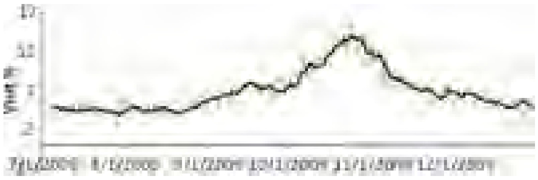

Figure 1 A flu chart showing daily percentage of ED visits with flu (green) and its 5-day moving average (black) between 7/1/2009 and 12/1/2009.

Acknowledgements

The project is funded by CDC grants P01 HK000086 and 1U38 HK000063-01. 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.

References

1 Que J, Tsui FC, Wagner MM. Timeliness study of radiology and microbiology reports in a healthcare system for biosurveillance. *AMIA Annu Symp Proc* 2006; 1068.

2 Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004;**11**:392–402.