Mining Pattern Model of Influenza Surveillance

Tippa Wongstitwilairoong, M.Sc., Jariyanart Gaywee, Ph.D., Narongrid Sirisophana, M.D., Carl J. Mason, M.D., Julie A. Pavlin, M.D., Ph.D., MPH

Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand

OBJECTIVE

This paper presents an investigation using data mining techniques to model patterns of influenza from positive case demographics, symptoms and laboratory tests.

BACKGROUND

Traditionally, infectious disease research approaches have started with a hypothesis followed by data query and report generation. Today, data can be mined to generate new hypotheses, and then confirmed through data experimentation. This shift brings new opportunities to develop system analyses using advanced data mining techniques [1, 2]. The data mining paradigm is used to create synthetic data sets that simulate outbreaks of infectious diseases based on demographic and laboratory information [3]. Influenza is one of the most important diseases to model, as it causes epidemics every year worldwide and the potential exists for severe pandemics. Military populations are particularly susceptible to outbreaks due primarily to crowding, troop movement, physical stress, and immunological naïveté [4]. New methods are now available to analyze data using advanced data mining techniques. Data mining finds patterns and relationships in data by building models which may allow earlier recognition of an influenza outbreak compared to other circulating respiratory illnesses.

METHODS

In our study, we have used the Royal Thai Army (RTA) influenza surveillance dataset based on 578 respiratory illness cases during March 2007 – April 2008 from 6 RTA hospitals. The data set included the following information:

- Demographic: age, date, sex, site, travel history, vaccination history
- Symptom: oral or axillary temperature, chills, cough, fatigue, headache, malaise, snuffy nose, sore throat
- Laboratory test results: rapid diagnostic tests and reverse transcription polymerase chain reaction (RT-PCR) for influenza A or B

Out of the 578 cases, 106 were influenza A, 27 influenza B and 445 negative for influenza by RT-PCR and this was used as the gold standard for comparison to all other variables. We used the Waikato Environment for Knowledge Analysis (Weka) software tool from the University of Waikato [5]. It is open source software that has many well-known data mining algorithms. We analyzed the data using two data mining techniques: association rule and classification techniques. Several experiments were conducted using apriori and C 4.5 decision-tree generating algorithms to generate the model.

RESULTS

For the Association Rule: the apriori algorithm depicts attributes that occur together frequently in the RTA influenza dataset with a minimum confidence = 0.9. The 20 best pattern rules for a confirmed Influenza A case were the following attributes which occurred together: fatigue, headache, malaise, snuffy nose and positive for influenza A by rapid diagnostic test. For classification: using the C4.5 algorithm, we used a cross validation and ran the classifier 10 times with 10 folds. This model generated main predictive variables: demographics (age, site, date, vaccination status); symptoms (chills, cough, fatigue, headache, oral or axiliary temperature); and laboratory test results (rapid diagnostic tests). The model correctly classified variables in 83.4% of all cases. Sensitivity was 48.1% for a positive influenza case using the model variables, but had 94.6% specificity.

CONCLUSIONS

This study shows that the preliminary results are promising for application of the data mining methods in influenza databases. Data mining can be applied to influenza datasets to extract predictive information about influenza epidemics, including demographics, symptoms, and laboratory results. Decision tree models can be useful for influenza surveillance and strategic planning in control of influenza. The analysis can assist in promising an investigation of a case that has symptoms related to Influenza. To improve the model in the future we will include more cases, increase the number of variables and use data from multiple years.

REFERENCES

[1] Frank E, Hall M, Trigg L, Holmes G, Witten LH. Data mining in bioinformatics using Weka. *Bioinformatics* 2004 Apr 8; 20(15): 2479-81

[2] Gewehr JE, Szugat M, Zimmer R. BioWeka – extending the Weka framework for bioinformatics. *Bioinformatics*. 2007 Mar 1;23(5):651-3.

[3] Ngan PS, Wong ML, Lam W, Leung KS, Cheng JC. Medical data mining using evolutionary computation. *Artif Intell Med*. 1999 May;16(1):73-96.

[4] Canas LC, Lohman K, Pavlin JA, Endy T, Singh DL, et al. The Department of Defense laboratory-based global influenza surveillance system. *Mil Med*. 2000 Jul;165(7 Suppl 2):52-6.

[5] www.cs.waikato.ac.nz