Optimizing A Syndromic Surveillance Text Classifier for Influenza-like Illness: Does Document Source Matter?

Brett R. South, MS1,2,6, Wendy W. Chapman, PhD3, Sylvain Delisle, MD,MBA4, Shuying Shen, MStat1,2,6, Ericka Kalp, MPH5, Trish Perl, MD,MSc5, Matthew H. Samore, MD1,2,6, Adi V. Gundlapalli, MD,PhD,MS1,2,6

1VA Salt Lake City Health Care System, 2Department of Internal Medicine, University of Utah School of Medicine, 3Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania 4VA Maryland Health Care System and University of Maryland School of Medicine, 5Johns Hopkins Medical Institutions and University, 6Department of Biomedical Informatics, University of Utah School of Medicine

Abstract
Syndromic surveillance systems that incorporate electronic free-text data have primarily focused on extracting concepts of interest from chief complaint text, emergency department visit notes, and nurse triage notes. Due to availability and access, there has been limited work in the area of surveilling the full text of all electronic note documents compared with more specific document sources. This study provides an evaluation of the performance of a text classifier for detection of influenza-like illness (ILI) by document sources that are commonly used for biosurveillance by comparing them to routine visit notes, and a full electronic note corpus approach. Evaluating the performance of an automated text classifier for syndromic surveillance by source document will inform decisions regarding electronic textual data sources for potential use by automated biosurveillance systems. Even when a full electronic medical record is available, commonly available surveillance source documents provide acceptable statistical performance for automated ILI surveillance.

Introduction
Automated text classifiers for syndromic surveillance have primarily focused on chief complaint (CC) fields, emergency department (ED) notes, nurse triage notes, and radiology reports as document sources for classification of concepts identified from electronic free-text data sources1-9. In 2006 AHIC outlined specific data elements for a minimum dataset used for biosurveillance10. These recommendations focus on establishing a minimum data set that incorporates both structured and unstructured data elements that could be used for biosurveillance. Listed in this recommendation were two specific textual data source documents: chief complaint text and emergency department notes. These document sources have been the focus of most investigations due to their wider availability from Electronic Medical Record (EMR) systems. In settings where other electronic notes are available, there has been limited work in the area of surveilling the full text of all available free-text encounter notes to build text classifiers for syndromic case detection.

One area that has also been previously unexplored is building text classifiers from free-text data sources extracted from the Veterans Affairs (VA) EMR for automated syndromic case detection based on specific document source and the full text of available documents. Specific data elements including ICD-9-CM codes extracted from the VA EMR have been incorporated in both the BioSense11 and ESSENCE12 biosurveillance systems. However, few applications utilizing electronic encounter note documents from the VA EMR have been implemented in these systems.

The VA currently maintains one of the few completely electronic and nearly paperless medical records that exist in the US health care sector. Free-text data are entered directly in the VA EMR via the GUI application called the Computerized Patient Records System (CPRS) This GUI application allows providers to enter documentation of patient encounters in real time. Since clinical notes are entered directly into the VA EMR by clinicians providing health
services, free-text electronic note documents are available immediately upon the completion of a patient encounter. Concurrent charting offers significant improvements in timeliness and availability of data compared with dictated or paper notes used by other health care systems. This last point is particularly important for the detection of potential bioterrorism threats or naturally occurring disease outbreaks, when the necessity of providing timely outbreak detection is paramount.

With improvements in data management and data extraction tools that have developed from VA national and local data warehousing efforts, electronic textual data are more readily available. However there are high workload requirements for implementing automated surveillance approaches using all available electronic documents.

This study evaluates and compares the performance of a simple text classifier developed for case detection of ILI applied to electronic document sources extracted from the VA EMR. Specific document sources were categorized into (1) document sources most often used for biosurveillance purposes (referred to as surveillance document sources) including (a) chief complaint string; (b) emergency department notes; (c) nursing or nurse triage notes; (2) routine clinic visit notes. We also evaluate an approach using a full text and full note corpus. Our hypothesis was that syndromic surveillance applied to a specific set of document sources alone would provide acceptable case detector performance over approaches utilizing the full text or full corpus of available note documents.

**Methods**

This study was carried out at the Baltimore VA Health Care System in Baltimore, Maryland and the Salt Lake VA Health Care System in Salt Lake City, Utah. Combined, these facilities provide care for nearly 90,000 patients in Maryland, Utah and surrounding states with an average of 40,000 emergency department, and over one million routine clinic visits per year.

**Setting**

Following approval by the Institutional Review Boards of both participating institutions a random sample of 15,377 patient encounters from two VA medical centers during the study period October 2003 through March 2004 was selected for manual chart review. Sampled patient locations included the emergency department and locations that provide outpatient care to Veterans.

A trained non-physician reviewed all CPRS data elements including clinical parameters, for vital signs, orders, imaging results, medications, and electronic textual data for documentation indicating ILI. A patient was considered to have a clinical standard diagnosis of ILI if (1) the patient had a positive influenza culture or influenza antigen; or (2) any two of the following symptoms were present for more than seven days duration: cough, fever or chills or night sweats, pleuritic chest pain, myalgia, sore throat, headache; and (3) illness not attributable to non-infectious etiology. The preliminary reviewer performed a screening review, flagging those cases having any possibility of being positive for ILI based on case definition criteria. All cases flagged as ILI from the initial screen and a 10% random sub-sample of negative patient encounters were subjected to secondary review by a physician. All discordant cases between the screening and secondary review were then arbitrated by three physicians specializing in pulmonary medicine and infectious disease. Final physician arbitrated cases and those cases where there was agreement between the first and second reviewer were used as the reference standard for evaluation of the text classifier for ILI case detection.

**Document Corpus and Text Classifier Development**

The document corpus for this study included the full text of 76,500 electronic note documents extracted from the VA EMR corresponding with the 15,377 sampled patient encounters. Note documents traditionally used for automated surveillance purposes were separated from routine visit encounter documents and
categorized into chief complaint, emergency department notes, and nurse triage notes. Since the VA EMR does not contain a separate chief complaint field, chief complaint strings were extracted from sampled encounter documents where chief complaint or reason for visit was specified in the text of the encounter note.

The text classifier used in this study was based on string matching for concepts from the ILI case definition mapped to the Unified Medical Language System (UMLS)\textsuperscript{13} coupled with the unmodified version of a negation algorithm called NegEx\textsuperscript{14}. Our final list of search strings included 186 synonyms term variants, and additional terms and misspellings for relevant identification of ILI based on the case definition and clinician review mapped to concept unique identifier (CUI) codes obtained using the NLM UMLS Metathesaurus\textsupersaurus search tool. Our goal was to demonstrate a simple text-based biosurveillance approach that requires limited resources to implement and maintain. We also sought to evaluate the performance of the unmodified version of NegEx applied to specific categories of VA electronic note documents.

We compared statistical performance of the text classifier applied to different logical combinations of document sources from the VA clinical record. These were as follows: (1) surveillance document sources including (a) chief complaint string; (b) nursing or nurse triage note; (c) emergency department notes; (d) combined surveillance document sources (a), (b) and (c); (2) routine clinical documents; and (3) the full corpus of all combined document sources. Patient encounters having document sources with two or more unique non-negated concepts were flagged as potential ILI cases. We measured sensitivity, specificity, positive and negative predictive values (PPV and NPV), and area under the receiver operating curve (ROC) with 95% confidence intervals (95% CI) for the text classifier applied to each document source using Stata statistical software version 10\textsuperscript{15}.

**Results**

Initial screening review identified 355 patient encounters potentially meeting ILI case definition criteria. Patient encounters identified by initial screen plus a 10\% random sample of negative patient encounters were subjected to a secondary physician review (n=505). Between screening and secondary review, there was disagreement for 33 patient encounters arbitrated by a panel of three physicians. The final reference standard identified 280 patients (sample prevalence 1.8\%) meeting the case definition for ILI.

Chief complaint text string was available for 50\% of patients (13,883 notes), emergency department encounter notes were available for 16\% of patients (5,178 notes), and nursing or nurse triage notes were available for 35\% of patients (10,111 notes). However, when combined two thirds (68\%) of patients (25,403 notes) had any one of the surveillance note types (Table 1). There was some overlap between routine visit notes, nurse triage and note documents from which chief complaint string was extracted.

**Statistical Performance of the ILI Text Classifier on Specific Document Sources**

When the text classifier was applied to each document source separately or in various logical combinations, specificities were always greater than 95\%. Sensitivity was highest when the text classifier was based on the full note corpus at 89\% and lowest at 4\% when nurse triage notes were used. Positive predictive values were highest when the text classifier was applied to CC, ED and combined surveillance document sources. Additionally, PPV based on all combined surveillance document sources (27\%) was higher than for the full note corpus and had a sensitivity of 75\% and high specificity 96\%. Text classifier performance based on routine note documents produced only marginal sensitivity 30\% and PPV 20\%. Finally area under the receiver operator curve was highest at 92\% when the full corpus was used as the text classifier document source (Table 2).

**Error Analyses of the ILI Text Classifier on Specific Document Sources**

An important consideration for biosurveillance systems incorporating various document sources for case detection is the number of false positive
and false negative cases that will be generated by the text classifier. It is important to note that the highest number of false positives (5%) (1-specificity) and lowest number of false negatives (11%) (1-sensitivity) were produced when the classifier was applied to the full text and full note corpus. Whereas the highest numbers of false negatives were generated when the classifier was applied to nurse triage notes (96%) and CC string (73%). Although almost one third of patients had some type of nurse triage note, and half of all patients had a readily extractable CC string, the number of ILI concepts extracted by the text classifier in these document sources did not meet our inclusion rule of two or more unique non-negated concepts. When the text classifier was applied to combined surveillance document sources, or the full note corpus there was a dramatic reduction in false negative cases (25% versus 11%), suggesting that ILI concepts are also found in other note sections or VA note document types besides surveillance document sources.

Limitations
Further training and adaptations of the NegEx algorithm for negation patterns and templated note structures peculiar to VA note documents may improve text classifier accuracy. We did not evaluate the effect of the overlap between routine visit notes and note documents containing chief complaint string and nurse triage document sources. However, these effects should be minimal, since our evaluation of text classifier performance is at the level of case detection based on identification of unique non-negated ILI concepts. ROC values are based on single point estimates using previously published statistical methods. Finally, these results should also be evaluated against other syndromes and electronic note documents obtained from additional VA facilities.

Conclusions
Clinician notes represent a significant proportion of patient information in the VA EMR. Our work builds on previous efforts and expands the development of syndromic text classification to the full text of electronic clinical notes for ILI. Though gains in sensitivity were achieved using the full corpus of notes documents, these results suggest that applying a relatively simple text classifier to all available and combined surveillance document sources may provide acceptable statistical performance for routine ILI biosurveillance in situations where all encounter documents are not available. As a more diverse set of document sources becomes available in commercially available EMR products it may not be necessary to perform automated surveillance on all of them. This would conserve computing and public health resources and decrease investigations of false positives. A more thorough surveillance of the full corpus of notes could be reserved for researching targeted conditions.

The VA Health Care System provides care to approximately 5.5 million patients having 55.7 million outpatient visits per year. Efforts are underway to include data from the VA EMR in two of the largest biosurveillance systems available for public health agencies. Therefore, further work is necessary to evaluate potential gains offered by adapting NegEx to VA note documents to improve text classifier accuracy for ILI and other syndromes. Additional work is necessary to evaluate other variables that could be used for case detection as textual data sources are only one of many data elements for automated biosurveillance specified in the AHIC MDS.

Acknowledgements
This study was supported using resources and facilities at the VA Salt Lake City Health Care System and VA Maryland Health Care System. The authors also gratefully acknowledge funding from the CDC for the BioSense Evaluation Grant (Johns Hopkins University) and Center of Excellence in Public Health Informatics (University of Utah). Additional support was received from NLM Grant 1 R01LM009427-01.

References
2. Chapman WW, Dowling JN, Wagner MM. Classification of emergency department
15. Stata SE version 10 statistical software College Station, Tx: StataCorp LP.

Table 1. Distribution of Documents by Category and Source

<table>
<thead>
<tr>
<th>Document category</th>
<th>Document source</th>
<th>Patients</th>
<th>Note documents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surveillance</strong></td>
<td>*CC string</td>
<td>7,608</td>
<td>13,883</td>
</tr>
<tr>
<td></td>
<td>**ED notes</td>
<td>2,485</td>
<td>5,178</td>
</tr>
<tr>
<td></td>
<td>Nurse triage notes</td>
<td>5,402</td>
<td>10,111</td>
</tr>
<tr>
<td></td>
<td>Combined surveillance</td>
<td>10,439</td>
<td>25,403</td>
</tr>
<tr>
<td><strong>Routine</strong></td>
<td>Routine clinic notes</td>
<td>12,802</td>
<td>51,097</td>
</tr>
<tr>
<td><strong>Full note corpus</strong></td>
<td>All electronic notes</td>
<td>15,377</td>
<td>76,500</td>
</tr>
</tbody>
</table>

*CC=Chief complaint  **ED=Emergency Department

Table 2. Accuracy of ILI Text Classifier by Specific Document Source

<table>
<thead>
<tr>
<th>Document category</th>
<th>Document source</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>PPV (95% CI)</th>
<th>NPV (95% CI)</th>
<th>ROC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surveillance</strong></td>
<td>CC string</td>
<td>27 (22,33)</td>
<td>99 (99,99)</td>
<td>33 (27,40)</td>
<td>99 (99,99)</td>
<td>63 (61,66)</td>
</tr>
<tr>
<td></td>
<td>ED notes</td>
<td>51 (45,57)</td>
<td>98 (98,99)</td>
<td>37 (32,42)</td>
<td>99 (99,99)</td>
<td>75 (72,78)</td>
</tr>
<tr>
<td></td>
<td>Nurse triage notes</td>
<td>4 (2.7)</td>
<td>100 (100,100)</td>
<td>17 (9,28)</td>
<td>98 (98,99)</td>
<td>52 (51,53)</td>
</tr>
<tr>
<td></td>
<td>Combined surveillance</td>
<td>75 (70,80)</td>
<td>96 (96,97)</td>
<td>27 (24,30)</td>
<td>100 (99,100)</td>
<td>85 (83,88)</td>
</tr>
<tr>
<td><strong>Routine</strong></td>
<td>Routine clinic notes</td>
<td>30 (24,35)</td>
<td>98 (98,98)</td>
<td>20 (16,24)</td>
<td>99 (99,99)</td>
<td>64 (61,66)</td>
</tr>
<tr>
<td><strong>Full text full corpus</strong></td>
<td>All electronic notes</td>
<td>89 (85,93)</td>
<td>95 (94,95)</td>
<td>23 (21,26)</td>
<td>100 (100,100)</td>
<td>92 (90,94)</td>
</tr>
</tbody>
</table>