An automated surveillance and analysis tool for online STI/HIV behaviors

Jonathan P. DeShazo\textsuperscript{1,2}, Bryant T Karras MD\textsuperscript{1,2},
\textsuperscript{1} Department of Medical Education & Biomedical Informatics, School of Medicine,\n\textsuperscript{2} Department of Health Services, School of Public Health and Community Medicine,\nCenter for Public Health Informatics, University of Washington, Seattle, WA;

Abstract
As part of a preliminary investigation, proof-of-concept software was developed to automatically retrieve, process, and classify online behaviors related to STI/HIV transmission. 300 online sexual solicitations were automatically collected, processed and characterized by the software using a shallow natural language processor. The results of the software analysis were compared with expert manual coding for agreement. The study demonstrated the feasibility of an automated online behavioral surveillance system.

Introduction
Sexual behavior is the primary determinant for sexually transmitted infections (STIs), and behavioral surveillance has played an integral role in population STI/HIV management\cite{1, 2}. Online venues are often considered to have potential roles in STI/HIV intervention, as well as access to underrepresented subgroups. Use of the internet for sexual solicitation is growing, and online behaviors have been linked with higher risks of STI/HIV\cite{3, 4}. In contrast to conventional behavior surveillance, relatively little is known about online behaviors regarding STI/HIV. The overall goal of this research is to comprehensively characterize publicly available online sexual solicitations and evaluate quantitative methods of analyzing and reporting online risk behaviors of interest. This paper reports the preliminary findings of the proof of concept application.

Methods
A preliminary investigation was performed and proof of concept (POC) software tool was developed to: (1) determine the feasibility and estimate the difficulty of developing and online behavior analysis tool, (2) inform subsequent designs, (3) provide initial estimates of the landscape of online behavior found The POC was designed to automatically retrieve sexual solicitations from websites previously identified as having a large number of men who have sex with men (MSM) solicitation activity. The solicitation text is processed using a naïve natural language classifier, enhanced by the use of phrasing and structured web data elements. For the POC, the NLM processor targeted only one website and evaluated the messages using only six predetermined characteristics: Age and geographic location of solicitor, and whether the solicitor is (1) a MSM. (2) is self described as HIV+. (3) Willingness to engage in drug use(i.e. drug friendly). (4) is seeking unprotected anal intercourse (UAI).

The POC Tool was trained on 200 messages and tested using another 300 messages. Evaluation was performed by Kappa statistical analysis between the POC Tool and expert review of the presence of the identified risk factors.

Results

<table>
<thead>
<tr>
<th>Tool</th>
<th>Expert</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV+</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>MSM</td>
<td>237</td>
<td>232</td>
</tr>
<tr>
<td>Drugs</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>UAI</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

N=300
254 (85%) self reported geo locations found
233 (78%) self reported age found

Table 1: Inter rater agreement between software tool and expert review

Discussion / Conclusion
The proof of concept has demonstrated the feasibility of an automated online behavioral surveillance system. Based on the expert agreement as well as the high percentage of age and locations found, it is likely a production system can collect risky and safe online behaviors along with relevant demographic information. This study is limited by the sample size and rare occurrence of some reported risk factors, single expert review, shallow NLP techniques. Further work should to be done to advance this online behavior surveillance technology, and evaluate the value, and potential uses to Public Health.

References