

## ABSTRACT

# The use of Craigslist posts for risk behavior and STI surveillance

JA Fries, A Segre, L Polgreen, and P Polgreen

Computational Epidemiology Research Group, University of Iowa, Iowa, IA, USA  
 E-mail: jason-fries@uiowa.edu

## Objective

This paper describes a novel method of obtaining large scale, geographically diverse behavioral data about men who use the Internet to seek Sex with Men (MSM) by examining anonymous Craigslist message posts to predict HIV/AIDS.

## Introduction

The rise and associated risks of using the internet to find sexual partners among men who have sex with men (MSM) has been noted by many researchers.<sup>1,2</sup> The anonymity and relative ease of finding partners on the internet has facilitated casual sexual encounters that can encompass a variety of unsafe sexual practices, from anonymous partners to 'Party and Play' activities (PNP), slang for illegal drug use, unprotected sex, group sex and so on. These anonymous sexual encounters make it more difficult for public health officials to notify exposed partners.

In addition, detailed data regarding risk behaviors are generally obtained via conventional survey techniques, which are expensive to conduct. Thus, a general method of empirically deriving large scale, location-specific behavioral data could be immensely useful in understanding or anticipating STI outbreaks.

Craigslist is a website specializing in online classified advertisements around the world. Our hypothesis is that Craigslist contains rich behavioral data regarding MSM communities and that such information can function as proxy for external prevalence rates for diseases (that is, HIV/AIDS).

## Methods

Beginning 1 July, 2009, daily Craigslist RSS feeds were collected for eight personals categories in 416 local Craigslist sites around the United States. As of 1 September, 2010, 54,450,547 individual posts have been collected. Of the 6,951,603 posts in California, ~57% can be classified as MSM specific. Approximately 60% of these MSM posts can be naively geocoded to specific counties using the Google Maps API. All geocoded messages are then searched to identify two message categories: (1) self-disclosed, positive HIV status, (2) PNP.

California's department of public health provides quarterly statistics on HIV/AIDS cumulative case counts for all

counties in the state. March 2010 summary data were used to calculate the prevalence rate used in this analysis.<sup>3</sup>

To find proxies for HIV prevalence in each county in California, we used ordinary least squares (OLSs). The dependent variable in all models was the actual HIV rate in each county (number of HIV cases divided by the county population). Independent variables included broadband penetration by housing unit as well as variables generated from Craigslist. We considered the number of MSM posts, and within the MSM categories, we also considered the number of posts with self-disclosed HIV status and PNP.

## Results

MSM posts with HIV status disclosed as a fraction of MSM personal posts is a positive predictor of HIV rates (coefficient = 5.4;  $P < 0.001$ ;  $R^2 = 0.85$ ). MSM posts as a fraction of all personal posts is a positive predictor of HIV rates but the  $R^2$  was relatively low (coefficient = 0.0068;  $P = 0.008$ ;  $R^2 = 0.24$ ). In contrast, PNP posts as a fraction of all personal posts is a positive predictor of HIV rates with a higher  $R^2$  (coefficient = 0.56;  $P < 0.001$ ;  $R^2 = 0.75$ ).

## Conclusions

Our Craigslist HIV-positive self-disclosure rate is a proxy for HIV among Californian MSM communities at the county level. Future models will explore this relationship in more depth. The second two models show that there are meaningful behavioral data embedded within messages. When considering just counts of MSM posts, the amount of variation that can be explained is rather modest but including posts denoting high-risk behavior, the  $R^2$  increases considerably—much more of the variability in HIV rates is explained. These results suggest that more sophisticated data mining techniques could yield an important source of behavioral data to help understand and perhaps anticipate STI activity.

## Acknowledgements

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 McFarlane M, Bull SS, Rietmeijer CA. The internet as a newly emerging risk environment for sexually transmitted diseases. *J Am Med Assoc* 2000;**284**:443–6.
- 2 Rosser BR, Oakes JM, Horvath KJ, Konstan JA, Danilenko GP, Peterson JL. HIV sexual risk behavior by men Who use the internet to seek Sex with men: results of the MEN'S INternet Sex Study II (MINTS-II). *AIDS Behav* 13:488–98.
- 3 California Department of Public Health, Office of AIDS, HIV/AIDS Surveillance Section. <http://www.cdph.ca.gov/data/statistics/pages/oahivaidstatistics.aspx>.