DDF Seeks Same: Sexual Health-Related Language in Online Personal Ads For Men Who Have Sex With Men

Oliver L. Haimson
University of California, Irvine
Department of Informatics
Irvine, CA, USA
ohaimson@uci.edu

Jed R. Brubaker
University of California, Irvine
Department of Informatics
Irvine, CA, USA
jed.brubaker@uci.edu

Gillian R. Hayes
University of California, Irvine
Department of Informatics
Irvine, CA, USA
gillianrh@ics.uci.edu

ABSTRACT

The HIV/AIDS crisis of the 1980s fundamentally changed sexual practices of men who have sex with men (MSM) in the U.S., including increased usage of sexual health-related (SHR) language in personal advertisements. Analyzing online personal ads from Craigslist, we found a substantial increase in SHR language, from ~23% in 1988 to over 53% today, echoing continuing concern about rising HIV rates. We argue that SHR language in Craigslist ads can be used as a sensor to provide insight into HIV epidemiology as well as discourse among particular communities. We show a positive significant relationship between prevalence rate of HIV in an ad's location and use of SHR language in that location. Analysis highlights the opportunity for SHR information found in Craigslist personal ads to serve as a data source for HIV prevention research. More broadly, we argue for mining large-scale user-generated content to inform HCI design of health and other systems, and explore use of such data to examine temporal changes in language to facilitate improved user-interface design.

Author Keywords

Health informatics; HIV/AIDS; personal ads; LGBT; online dating; digital identity; Craigslist; computational linguistics.

ACM Classification Keywords

H.4.3 Communication Applications; J.3 Life and Medical Sciences: Health; K.4.1 [Computers and Society]: Public Policy Issues: Computer-related health issues.

INTRODUCTION

When designing large-scale health systems, data giving insight into user practices and language choices can help HCI designers to inform choices in data structure and system features. Traditional data collection methods can be slow, expensive, and inaccurate, particularly when focusing on sensitive communities and practices. By exploring relationships between user-generated content and established data collection methods, we seek to augment existing practices and make data collection faster, cheaper, and possibly more accurate.

"Public health surveillance is the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice" [40]. Surveillance is frequently employed to provide early warning for public health emergencies, monitor health progress at the population level, and inform public health policy. However, public health professionals, in attempts to curtail HIV infection rates, have noted "an urgent need to address gaps in our ability to monitor changes in HIV, STDs, and sexual practices among MSM" [39:884] (MSM is short for men who have sex with men, an inclusive term used in public health literature). Conducting frequent population-based surveys along with facility-based surveillance, while effective [2,39], requires considerable time and resources, and often only reaches more visible segments of the MSM community [2].

In this paper, we explore the potential of using publicly available personal ads as a proxy for HIV and STI (sexually transmitted infection) statistics to augment current collection methods and provide more comprehensive data. Since the advent of online personal ads in the 1990s, MSM have willingly shared SHR information on sites such as Craigslist to facilitate sexual contact. Such data is free, plentiful, and readily accessible to researchers. The accessibility, affordability, and anonymity of Internet personal ads [11] make the Internet an "ideal medium for sexual pursuits" [9:74], but also an ideal environment for mining user-generated content.

As compared to survey-based research, online personal ads give researchers quick access to millions of anonymous ads, which contain information analogous to data found through surveys and surveillance [16,17]. Ads can be continuously and systematically collected, with only minimal costs such as computational time spent gathering data. Thus, online personal ads have the potential to make a substantial difference in HIV research and prevention efforts.

We analyzed online personal ad content and explored its relationship to HIV prevalence to demonstrate an

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2473-1/14/04...$15.00.
http://dx.doi.org/10.1145/2556288.2557077

application of mining online, publicly available data for use in real-world contexts. First, through our analysis of 252,786 MSM Craigslist ads, we identified SHR language currently used online. Second, through a comparison with print personal ads from the 1980s, we demonstrate an increase in the use of SHR language, signifying that even 30 years after the beginning of the HIV/AIDS crisis, health concerns among MSM persist and can be measured empirically online. Finally, by comparing use of SHR language in 95 locations, we find that HIV prevalence rates and SHR language in Craigslist ads have a significant positive relationship. Taken together, these contributions demonstrate the potential for publically available online data to be used as a surveillance tool and provide description of one method to create such tools.

We present this work as an example of mining user-generated online media content and using it as a sensor for secondary purposes. Large, publicly available online datasets are important sources of information for HCI researchers. To properly design large-scale health data systems, HCI research must be conducted on the information architecture of such data and its potential use as a sensor. Similar methods and techniques could be used in other domains (*e.g.*, urban or civic informatics). Additionally, our work highlights issues around the timescale of codifying large-scale data from online media, particularly given evolving language.

Our research also addresses gaps in the study of sexuality within HCI, particularly the shortage of research dealing with sexual orientation and homosexuality [24]. Research on the intersections of technology and sexuality contributes to the development and growth of the field of HCI [1]. By studying the online dating practices of MSM, we address the dearth of sexuality research within HCI.

The remainder of this paper is structured as follows: We first provide some background on HIV/AIDS and the use of personal ads by MSM, followed by a discussion of related research. We then describe the methods and results of our empirical work as conducted in three phases: developing a sexual health dictionary; determining presence of SHR language in Craigslist ads and comparing current metrics with those from the 1980s; and building and analyzing statistical models to explore the relationship between SHR language and HIV prevalence. We close with a discussion of HCI design implications and a summary of our findings.

BACKGROUND

HIV/AIDS

In the early 1980s, many gay men contracted and died from a mysterious disease initially known as “gay cancer.” The disease was eventually identified as AIDS, caused by the virus HIV and commonly spread through sexual contact. The HIV/AIDS crisis in the United States has been a considerable public health problem that has historically and continues to disproportionately affect MSM [5]. As a result, HIV/AIDS has fundamentally changed sexual practices in the U.S., particularly among MSM [39].

Highly Active Antiretroviral Therapy (HAART), a treatment first distributed in the United States in 1996, has succeeded in controlling HIV infections and decreasing AIDS deaths [28]. However, research has shown that some MSM conflate HAART’s benefits with a reduction in the risk of unsafe sex with HIV-positive partners, which has been shown to lead to a higher tendency to engage in unprotected sex [13]. Thus, HIV infection rates, especially those of MSM, have continued to increase.

MSM accounted for 63% of all new HIV cases in 2010, and make up 52% of all HIV cases in the United States [5]. Taking into account the fact that MSM only make up 2% of the U.S. population [5], these statistics are especially alarming. Particular MSM subgroups, such as those under the age of 24 and young African-Americans, experience even higher rates of HIV infection [5]. Considering these statistics, it is unsurprising that disclosure of HIV status and use of SHR language is part of courtship for MSM.

Personal Advertisements and MSM

Personal ads have historically been useful in facilitating exchange between interested people when dating preferences lie outside of traditional markets, such as MSM [10]. Multiple studies have found that MSM meet sexual partners online significantly more often than others [19,30,31]. Given the relative frequency with which MSM use the Internet for sexual communication and their disproportionate risk of contracting HIV, how can the online content generated by MSM be used as a sensor for public health efforts to reduce the spread of HIV?

Contradictory research has argued that online personal ads can either help or hinder HIV prevention efforts. An increase in MSM sexual contact brought about by online personal ads may have had negative effects on disease control [7,18,27]. For example, those who met sex partners online were more at risk to contract HIV and other STIs than those who did not [27]. Likewise, the launching of Craigslist for a particular city was found to predict an increase in contraction of both AIDS and syphilis in that city, and the number of MSM personal ads linked to a particular city was found to be a significant predictor for AIDS cases [7]. Additionally, Craigslist’s search function may enable risky behavior by allowing users to search for behaviors that they desire, such as “bareback” (sex without a condom), a functionality that would not be possible offline or in print personal ads [18].

On the other hand, online dating could support HIV prevention by allowing sexual partners to discuss HIV status and protection preferences prior to meeting [31,38]. Just as Craigslist allows for searching for risky behaviors [18], it could also facilitate searching for safe behaviors. However, relying solely on information that sexual partners provide online can increase risk if it eliminates further discussion of safe sex practices, particularly if sexual partners are unsure or incorrect about their HIV status [37].

While this debate is ongoing, in this paper, we adopt a new approach. We focus on the information that can be gleaned through an analysis of personal ads rather than on the practices that surround them. We demonstrate that computational analysis of language in MSM Craigslist personal ads can provide one source of public health surveillance for MSM. The information found in these ads has potential to aid in HIV prevention strategies that, if successful, could mitigate the negative effects that Craigslist has arguably had on the spread of HIV and STIs.

RELATED WORK

This paper draws from and contributes to several bodies of literature that have explored health implications of personal ads by MSM. Although previous studies have examined the use of SHR language in MSM personal ads [14,18,22] and others have argued that Craigslist ads can be used for public health surveillance and HIV/STI epidemiology research [16,17], we posited that further insight could be gained by joining these two research methods. We thus build on previous research by combining linguistic analysis of personal ads with epidemiological analysis to understand how online MSM personal ads can be used as a sensor for public health surveillance.

Several studies have examined the content of personal ads on Craigslist and how it relates to sexual health and risk of HIV and other STIs in MSM [8,18,22,29,32]. Health-related language has been found to be more prevalent in ads posted by HIV-positive MSM [22], giving evidence of serosorting ("preferentially selecting sex partners with concordant HIV status and … using condoms with partners of discordant status" [4:2497]), a method shown to reduce risk of HIV transmission [4]. One risk indicator is the volume of ads posted by any individual MSM, which predicted more likeliness to engage in unsafe sexual practices [29], while the marital status of MSM can also correlate with perceived safety [8,32]. These studies show how content of personal ads correlates with the sexual risk behaviors of those posting and replying to these ads.

One notable focus relevant here can be seen in epidemiological work on HIV and Craigslist. Several studies have found relationships between the content or volume of Craigslist ads and real world prevalence of HIV and other STIs, showing that online personal ads and Craigslist in particular are effective tools for HIV epidemiology research [7,16,17]. For example, Fries *et al.* computationally extracted HIV status information from millions of Craigslist ads and found a positive predictive relationship with HIV rates by location, demonstrating that HIV status information disclosed in Craigslist ads can be used as a proxy for HIV rates among MSM [17]. These rates can in turn be used in "understanding or anticipating STI outbreaks" [17:13]. In addition to HIV rates, Craigslist posters include information about many risk behaviors that allow for public health surveillance [16]. Similarly, Chiasson *et al.* argue that the Internet is an ideal place to conduct research on the sexual health of MSM [9].

Personal ads have been used to study changes over time in the use of health-related language long before the advent of the Internet. Sociologist Alan G. Davidson analyzed the percentage of personal advertisements that included health-related language in each of four years: 1978, 1982, 1985, and 1988 [14]. He found a "significant increase in personal advertisements suggesting a concern with health" from 1982 – 1985 [14:125], the time during which many gay men first learned about AIDS [26], and again from 1985 – 1988, showing that the effects from the first time period persisted [14]. Davidson's work highlights how the gay community responded to the outbreak of HIV/AIDS by changing the language that they used to describe themselves and their sexual and dating preferences [14].

Personal ads can be "useful data sources for assessing the meanings people attach to their sexuality, as well as for assessing changes in these meanings over time" [14:136]. Although the format and medium for personal ads has shifted from newspapers to websites, the implications of their power to convey sexual representations and practices has persisted and grown along with their volume. Thus, Davidson's work led us to address the research question of how sexual health discourse among MSM has changed over time and across mediums, both in content and volume

The literature on Craigslist and HIV/STIs has shown that Craigslist ads can be used as a kind of sensor. We demonstrate that when used to collect and analyze health data, this sensor can provide information about disease rates, risk of spreading disease, and particular communities who may be at risk of contracting disease. When used in a public health context, this information could have powerful effects on HIV prevention and provides a real world example of the kind of outcomes promised by publicly available "big data". Our work leverages linguistic analysis of personal ads as a potential way to harness such data.

DATA

Our initial goal was to replicate Davidson's 1991 study, to determine how time and platform affected use of SHR language in MSM personal ads. Davidson compared the use of SHR language in gay male personal ads published during 1978, 1982, 1985, and 1988 in the *Village Voice*, a weekly New York City (NYC)-based newspaper [14]. Our goal framed choices in data analysis, which began with NYC for the sake of comparison with Davidson.

Although methods of posting personal ads have changed in the last 25 years, we turned to Craigslist, a popular online classifieds website, as a modern equivalent of print personal ads. Like print personal ads, Craigslist posts are anonymous and stand-alone (as contrasted with profile-based online dating sites) and allow disclosure of sexual practices and health-related language. Differences between *Village Voice*

Population Range	Locations	Mean Population Density (SD) [33]	Mean HIV Estimated Diagnosis Rate (SD) [6]	Ads (% of Total)
> 5 Million	8	1897.6 (2231.7)	26.4 (7.2)	91,110 (36.04%)
2M – 5M	23	668.9 (444.3)	17.6 (10.0)	105,874 (41.88%)
1M – 2M	21	486.0 (317.4)	20.3 (11.6)	35,345 (13.98%)
< 1M	43	364.4 (252.8)	12.2 (8.7)	20,457 (8.09%)
TOTAL	95	594.1 (808.3)	16.5 (10.3)	252,786 (100%)
New York City		7231.6	36.5	10,737 (4.25%)

Table 1. Locations and Sample Sizes.

and Craigslist personal ads include message length, cost, and possibility of censorship. *Village Voice* personal ads had no word limit per se, but authors were charged on a per-line basis, while Craigslist ads are free with no word limit. Although the *Village Voice*'s censorship policies were not stated in the four 1978-1988 issues we accessed, the paper "reserves the right to reject or edit any advertisement" [35]. In comparison, Craigslist does not reject, remove, or edit ads unless other users flag ads for removal, and does not restrict adult content [12].

Our dataset comprises 252,786 personal ads posted to the "men seeking men" (m4m) subsection of Craigslist. Craigslist maintains a separate website for each of many cities and towns in the United States. Using a custom-built RSS scraper, we collected all m4m ads within a two-week period in August and September 2013 in 95 metropolitan statistical areas (MSAs) (see Table 1). Locations were selected to correspond with location-specific statistics on HIV prevalence rates as reported by the U.S. Centers for Disease Control and Prevention (CDC) in a 2011 report [6]. We excluded seven locations on the CDC's list of MSAs because a corresponding Craigslist site did not exist. Craigslist sites were selected to best approximate the geographic area of each MSA.

Our data collection methods captured each ad as it was first posted, meaning that our dataset includes ads that may have later been flagged by users and/or subsequently removed. In practice, ads are often removed when the poster wants no more responses; such ads are still relevant for analysis. Meanwhile, duplicate ads within a location were removed from our dataset prior to analysis. During manual coding of 500 ads, we identified a 0.2% rate of irrelevant ads.

There is a risk that people misrepresent their HIV status on Craigslist or may not be aware of it. However, at a population level, we are interested in capturing use of SHR language of any kind, not the specifics of any individual's personal status and claims. While there are almost certainly inconsistencies in individual ads, in aggregate, the data are relatively accurate [17].

Though we cannot claim that our sample is representative of all MSM in the U.S., research has shown that more than 85% of MSM find sexual partners online [3,20]. Additionally, the existence of a relationship between HIV prevalence rates in CDC data and the use of SHR language in our data, along with previous literature that has found similar positive correlation [17], signals the appropriateness of using Craigslist as a source to study MSM sexual health.

On average, large cities included more ads than small cities: the eight cities with populations over five million comprised 36.04% of total ads, and the 31 cities with populations over two million comprised 77.92% of total ads. A majority of ads (88.77%) included the poster's age. Of those ads with age included, excluding ads with reported age 99 or <10, the mean age in NYC was 34.59 (SD = 9.40; range 18-80), while the mean in the overall dataset was slightly older: 34.81 (SD = 10.24; range 10-98).

SEXUAL HEALTH DICTIONARY

Methods

To analyze the use of sexual-health related (SHR) content in personal ads, we first created a dictionary of SHR language. Davidson provided a dictionary of health-related language based on ads from 1978-1988 [14], but instead of adopting Davidson's dictionary or analyzing the data using a pre-supposed list of words and terms, we used open-coding techniques to find the SHR language that MSM used in our dataset [14]. This method allowed the discourse used by the MSM community in their personal ads to shape the dictionary [14]. Four coders, including two gay-identified men, were given a sample of Craigslist ads from the dataset, and were asked to independently determine SHR words or terms in each ad. 500 ads in total were coded, 125 by all four coders for inter-rate reliability. Use of multiple coders allowed us to detect SHR language that may not be familiar to or identified by one coder alone. Agreement levels were acceptable at 72% with a Fleiss' Kappa of .708. Words or terms were added to the dictionary if one or more of the coders identified the term as SHR. Two additional terms were added to our dictionary even though our coders did not identify them. These terms, "seeding" (and its shorter form, "seed") and "uninhibited", were mentioned in previous similar studies [18,32] and occurred in our full dataset, but not in the subset used for open-coding.

Next, to facilitate more detailed analysis, we divided our dictionary into 6 categories: disease, HIV (a sub-category of disease), protection, risk (a sub-category of protection), safety, and health. Table 2 shows our full dictionary as compared with Davidson's dictionary [14].

We intentionally excluded drug related terms to focus the scope of our paper on sexual health, replicate Davidson's methods, and compare current levels of SHR language to the 1980s. Terms that referenced drugs without explicitly

Sexual health-related language, Craigslist 2013	
Category	**Words and terms**
Disease	bug free, clean, ddf [drug and disease free], disease free, std free, tested
HIV (sub-category of Disease)	AIDS, HIV+, HIV-, neg, negative, pos, positive, poz, undetectable
Safety	safe, safely, safe sex
Protection	condom, protection
Risk (sub-category of Protection)	bareback, bb [bareback], bred, breed, breeding, raw, seed, seeding, uninhibited
Health	health, healthy
Davidson's health-related language, *Village Voice* 1978-1988 [14]	
Category	**Words and terms**
Health-Related	health, healthy, health conscious, clean, unaffected, unused, taking the current situation very seriously, interested in safe dating, safe for us both, HIV negative, safe-sex romance, safe, tested disease-free, AIDS-free relationship
Sexual Exclusivity Subcode	monogamous, monogamy, exclusive, 1-to-1, one to one, non-promiscuous, dislike promiscuity

Table 2. Sexual Health-Related Dictionaries, Craigslist 2013 and *Village Voice* 1978-1988.

mentioning sex were not included in Davidson's study and were subsequently omitted here. However, drug use and sexual health are often bound up with each other, and future research should examine drug use and its relationship to sexual health and HIV/AIDS among MSM.

Results

There are four important differences between our 2013 SHR language dictionary and Davidson's 1978-1988 dictionary (Table 2) [14]. First, Davidson's dictionary included a subset of terms related to monogamy and relationship exclusivity, while our dictionary does not [14]. In the 1980s, exclusivity was often motivated by HIV/STI prevention, which is no longer the case [23]. Davidson analyzed monogamy-related terms separately [14], allowing us to compare our findings to only his health-related terms.

Second, in the disease category and HIV sub-category, the 1978-1988 dictionary includes relatively few terms: "tested disease-free", "HIV negative", and "AIDS-free relationship." In 2013 the corpus of disease language has both increased (now including words such as "bug", "clean", and "std") and settled on shorter phrases and/or abbreviations to describe one's own STI status and to convey a preference for a partner's STI status: "ddf", "HIV+" or "poz", "HIV-" or "neg".

Third, in our 2013 Craigslist dictionary, we found the word "safe" primarily used in a list of words describing the poster or his desired partner (*e.g.* "very discreet, safe, sane, clean, d&d free, you also"). While Davidson's dictionary includes "safe", it also appears as part of several longer phrases that use "safe" in a dating context: "interested in safe dating", "safe-sex romance" [14]. Again, we see 2013 Craigslist ads using shorter phrases than the 1978-1988 dictionary.

Finally, words describing protection or lack of protection occur regularly in our 2013 data, but not at all in the 1978-1988 data [14]. This suggests that contemporary MSM are more likely to communicate about protection techniques and preferences in personal ads than 1980s MSM.

In the 25 years since Davidson's 1988 dataset was collected, MSM have standardized and abbreviated their SHR language, allowing for rapid and efficient communication. Surprisingly, even though physical and financial word count barriers have been eliminated in the Craigslist personal ad medium, our Craigslist SHR language dictionary includes shorter and more to-the-point terms than the print ad dictionary. This is likely a result of the fact that SHR information is now routinely communicated, demonstrated by the use of abbreviations. The differences between our dictionary and Davidson's show how SHR language has evolved beyond a topic of discussion or an ideal, and into the world of personal stats – akin to height, age, and weight.

Additionally, our analysis shows how personal ad text evolves over time, leaving open questions about adoption of language as indicated by the changes between Davidson's dictionary and ours. Differences in SHR language between 1980s personal ads and current Craigslist ads highlight the constant evolution of SHR language.

SHR LANGUAGE IN PERSONAL ADS OVER TIME

Methods

The next step in our comparison to Davidson's work was to detect the presence of SHR language across our larger dataset. We started our analysis with ads posted to NYC-area Craigslist. The text of each ad was compared to the terms included in our dictionary, and each ad was assigned a single binary indicator for the presence of SHR language.

In addition to the dictionary entries listed in Table 2, we also searched for variations in phrasing, punctuation, and spacing. For example, in addition to "disease free", we also searched for "disease-free", "no diseases", "diseased free", and "diseased-free", all of which appeared in the data. A term such as "ddf" (an abbreviation for "drug and disease free") likewise was expanded to include "dd free", "d&d free", "d&df", "d/df", "dd/f", etc. Other terms in our dictionary (*e.g.* "clean") had to be distinguished from longer terms (*e.g.* "clean cut") that were not part of our dictionary and would not result in an indication of SHR language. To reduce false positives in our risk category, negations of the SHR term were detected and ignored or reassigned as necessary. For example, an ad including the phrase "no bareback" would be put in the protection category, but not the risk category.

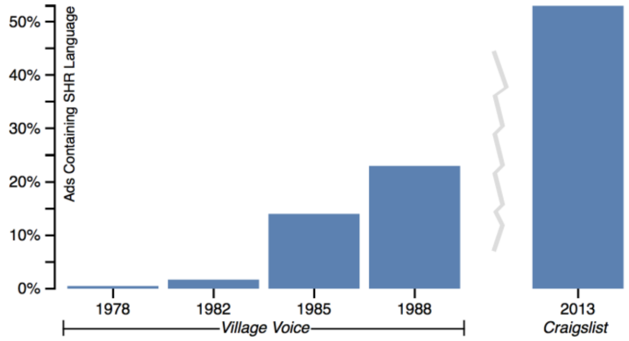


Figure 1. Sexual Health-Related Language in MSM Personal Ads Over Time, New York City Area.

Results

We began by comparing data from the NYC metro area with historical data. Use of SHR language increased substantially, from 22.99% to 53.50% from 1988 to 2013 (see Table 3 and Figure 1). In our full dataset, we see an even greater percentage of SHR language: 53.96%.

The increase in usage of health-related language that Davidson observed in the 1980s and attributed to concern with and response to HIV/AIDS [14] has not only persisted but has continued to grow throughout the last 25 years. Use of SHR language, including disclosing HIV status and safe sex preferences, has become a standard in MSM personal ads, occurring in over half of the ads in our sample.

Across our categories of SHR language, disease-related is the most common (48.34%). Within that category, 30.53% of ads with disease-related content include mention of HIV. This provides evidence to support our argument that the increase in health-concern in MSM personal ads over the past 25 years is greatly motivated by disease prevention, HIV prevention in particular.

Following disease, safety is the next most common category (12.85% of ads). Protection-related terms, including many connoting risky sexual practices, were less common (4.78%), however, ads that used language specifically associated with a preference for unsafe sexual activity occurred in 3.38% of ads. This is comparable to Grov's finding that 3% of Craigslist MSM ads in 2009 sought unprotected sex [18]. Although a relatively small percentage compared to our other categories, 3.38% corresponds to approximately 8,544 ads seeking risky sex encounters in only two weeks. In NYC alone, a smaller percentage of ads included risky SHR language (2.35%), but this still corresponds to 252 ads in two weeks.

SHR LANGUAGE AND HIV PREVALENCE

Methods

To explore relationships between characteristics of MSM and their locations and use of SHR language in personal ads, we constructed seven multivariate logistic regression models (see Table 5). Dependent variables were binary indicators of the presence of SHR language overall in each of our six SHR language categories: disease, HIV, safety, protection, risk, and health. We used the author's age, the ad's word count, and the ad location's HIV prevalence (rate per 10,000 population) [6], population [34], and population density [33] as regressors. Ads in which authors did not indicate an age, as well as those with a reported age of 99 or <10 were excluded from analysis (11.33% in total).

Category	*Village Voice* (1988)	Craigslist, NYC only (2013)	Craigslist, 95 locations (2013)
Overall SHR	22.99%	53.50%	53.96%
Disease	NA	46.30%	48.34%
HIV (sub-category of Disease)	NA	17.03%	14.76%
Safety	NA	17.87%	12.85%
Protection	NA	3.56%	4.78%
Risk (sub-category of Protection)	NA	2.35%	3.38%
Health	NA	2.12%	1.23%
Davidson's health-related dictionary [14]	22.99%	34.35%	32.73%
Davidson's sexual exclusivity subcode [14]	13.41%	0.07%	0.15%

Table 3. Presence and Types of Health-Related Language in Craigslist 2013 and *Village Voice* 1988.

Ads were placed into one of six age groups based on the poster's age, as reported in Table 4. Because use of SHR language varies significantly among age groups, age is used as a categorical rather than a continuous variable. Great care was taken to construct age groups based on HIV/AIDS history. Those in the youngest two age groups (born after 1986) represent those "too young to remember life before HAART" [21:176]. Further divisions were made based on stage of life during the 1980s, when most people learned of HIV/AIDS. Those 45 and over compose one group, as these men were likely sexually active during the 1980s. Another group includes men 21 and younger, who may exhibit higher risk behaviors than older men [25]. Each age group has a similar sample size, with the exception of the youngest age group, which includes substantially fewer people, but is still important to keep separate due to the potential differences in risk preference between this group and the others [36]. Age groups were used as dummy variables, with the oldest age group as reference value.

Age in 2013	Sample size	During AIDS outbreak (~1983-1988)	Remember life before HAART?
10 - 21	15,832	not born	no
22 - 26	41,557	not born	no
27 - 31	40,800	born	yes
32 - 37	39,315	child	yes
38 - 44	41,029	teenager	yes
45 - 98	45,619	adult – likely sexually active	yes

Table 4. Age Groups.

	Sexual Health-Related (1)	**Disease (2)**	**HIV (3)**	**Safety (4)**	**Protection (5)**	**Risk (6)**	**Health (7)**
Predictor	Odds ratio (95% confidence interval)						
Word count	1.0089**** (1.0086, 1.0091)	1.0074**** (1.0072, 1.0076)	1.0044**** (1.0041, 1.0046)	1.0052**** (1.0049, 1.0054)	1.0053**** (1.0050, 1.0056)	1.0037**** (1.0033, 1.0040)	1.0062**** (1.0058, 1.0066)
HIV prevalence rate	**1.0004**** (1.0003, 1.0004)**	**1.0003**** (0.0002, 0.0003)**	**1.0010**** (1.0009, 1.0011)**	**1.0005**** (1.0005, 1.0006)**	**1.0003**** (1.0001, 1.0004)**	**1.0003*** (1.0001, 1.0004)**	**1.0005**** (1.0002, 1.0007)**
Population	1.0000**** (1.0000, 1.0000)	1.0000**** (1.0000, 1.0000)	1.0000**** (1.0000, 1.0000)	1.0000**** (1.0000, 1.0000)	1.0000**** (1.0000, 1.0000)	1.0000**** (1.0000, 1.0000)	1.0000** (1.0000, 1.0000)
Population density	0.9999**** (0.9999, 0.9999)	0.9999**** (0.9999, 0.9999)	0.9997**** (0.9997, 0.9997)	0.9999**** (0.9999, 1.0000)	0.9998**** (0.9998, 0.9999)	0.9998**** (0.9998, 0.9999)	1.0000 (1.0000, 1.0001)
(Ages 45 - 98)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Ages 38 - 44	1.0625**** (1.0337, 1.0922)	1.0513**** (1.0232, 1.0802)	1.1363**** (1.0943, 1.1801)	1.0356* (0.9968, 1.0762)	1.2033**** (1.1319, 1.2792)	1.2693**** (1.1825, 1.3625)	0.5887**** (0.5317, 0.6513)
Ages 32 - 37	0.9996 (0.9722, 1.0278)	0.9797 (0.9544, 1.0070)	1.1834**** (1.1398, 1.2294)	1.0495** (1.0096, 1.0908)	1.2361**** (1.1622, 1.3146)	1.2570**** (1.1698, 1.3506)	0.3554**** (0.3142, 0.4010)
Ages 27 - 31	0.8251**** (0.8029, 0.8481)	0.8347**** (0.8123, 0.8576)	1.1035* (0.9955, 1.0751)	0.9394*** (0.9033, 0.9770)	1.0009 (0.9385, 1.0673)	0.9746 (0.9033, 1.0514)	0.2995**** (0.2628, 0.3403)
Ages 22 - 26	0.7319**** (0.7122, 0.7520)	0.7748**** (0.7541, 0.7960)	1.0165 (0.9781, 1.0565)	0.7121**** (0.6831, 0.7422)	0.8249**** (0.7712, 0.8821)	0.7742**** (0.7145, 0.8386)	0.2108**** (0.1813, 0.2439)
Ages 10 - 21	0.5947**** (0.5731, 0.6172)	0.6544**** (0.6306, 0.6791)	0.8586**** (0.8125, 0.9069)	0.5357**** (0.5024, 0.5707)	0.8662*** (0.7901, 0.9483)	0.6596**** (0.5851, 0.7415)	0.0861**** (0.0596, 0.1198)

* p < .10; ** p < .05; *** p < .01; **** p < .001

Table 5. Results of Logistic Regression Models.

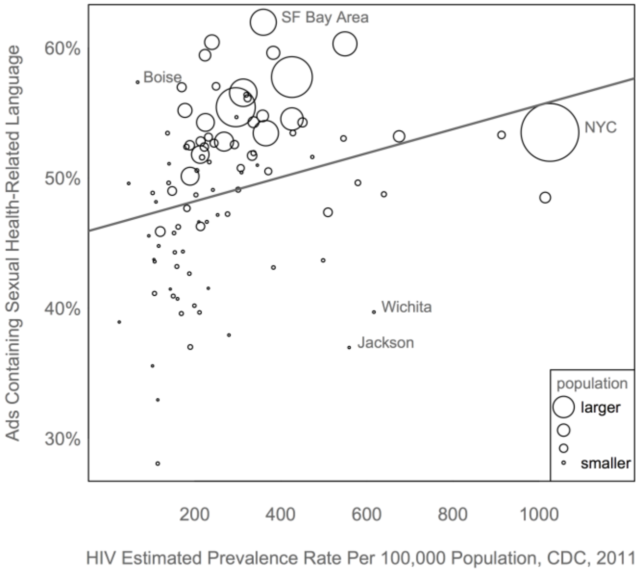
Results

Across all regression models (1-7), the prevalence rate of HIV in the ad's location was a significant positive predictor of use of SHR language, even after controlling for population, population density, word count, and age. Although the effect size is relatively small and the high significance level may be due to the very large sample size, our results show a meaningful effect. In model 1, for instance, an increase in HIV prevalence rate of 100 people per 100,000 population (*i.e.* an increase of 0.1%) predicts an increase of 4% in SHR language.

Though our results are correlational, the positive relationship between HIV prevalence and use of SHR language supports an extension of Davidson's findings that the HIV/AIDS crisis may have triggered increased health-concern in MSM personal ads [14]. In the previous section we showed that overall use of SHR language has increased substantially in the past 25 years. Here we see that the magnitude of this trend varies geographically, and may be motivated in part by HIV prevalence.

Cities with larger populations are more likely to have higher incidence of SHR language in MSM Craigslist ads (see Table 5 and Figure 2). In the regression models (1-7), population is a significant positive predictor of SHR language in all six categories and overall. The scatterplot (Figure 2) demonstrates this relationship and also allows us to see several outliers, such as the San Francisco Bay Area, which has the highest use of SHR language but only an average HIV prevalence rate. Boise, Idaho, similarly, has a very low prevalence of HIV but a high use of SHR language. Outliers in the opposite direction include Jackson, Mississippi and Wichita, Kansas, both of which have a relatively high prevalence of HIV and low use of SHR language. Outliers may signify the effectiveness of HIV prevention strategies in these locations. Identifying HIV/STI outreach and education approaches in locations such as Boise and the San Francisco Bay Area may be of use to outliers in the opposite direction (Jackson, Wichita), where similar approaches may help with HIV prevention.

Use of SHR language differs significantly across age groups, with men 21 and under using significantly less SHR



SHR language = 46.360 + 0.009 * prevalence rate, p=0.009

Figure 2. Relationship Between HIV Estimated Prevalence Rate and Use of SHR Language in 95 Locations.

language in all categories (models 1-7) than men in the reference group. This may be a result of the fact that young MSM exhibit higher risk preferences, or do not remember life before HAART [25]. However, MSM under 21 also use significantly less SHR language demonstrating a preference for unsafe sexual behavior than other age groups (model 6).

Overall (model 1), men aged 38 – 44 are the most likely to use SHR language, even more so than men who were adults during the beginning of the HIV/AIDS crisis. Perhaps these men, who were teenagers during the 1980s HIV/AIDS outbreak, came of age during a time when the importance of safe sex was especially emphasized.

Models 3 and 5 show that men aged 32 – 37 are relatively highly likely to use terms in our HIV and Protection categories. This may be a result of the multitude and prevalence of HIV/AIDS outreach and awareness campaigns in the 1990s [15], when these men were teenagers, which may have instituted a higher awareness of HIV/AIDS and sexual safety.

Interestingly, in our health category (model 7), age and use of SHR language have a positive relationship, meaning that as age increases, use of terms in this category ("health" or "healthy") increases. "Health" is a term that was featured prominently in Davidson's dictionary and in his paper, implying that use of the words "health" and "healthy" were important ways for MSM of the time to convey their HIV/STI status and preferences for safe sex [14]. Use of these words in this context has since tapered off in favor of standardized and abbreviated terms such as "ddf". However, the results of the regression model (7) indicate that those MSM who were using the words "health" and "healthy" to convey HIV/STI status in the 1980s may still be using these terms, while younger MSM are not.

Our analysis showing that an ad location's HIV prevalence rate is a significant predictor for use of SHR language in that ad (models 1-7) led us to further question whether the relationship holds in the other direction, *i.e.* whether the use of SHR language in a location predicts HIV prevalence rate in that location. Table 6 shows the results of a linear regression model testing this hypothesis (model 8). Though this relationship is not as strongly significant as the former (models 1-7), it is significant at the 10% level, even after adding control variables. The sample size is limited to the 95 locations from the CDC's dataset [6]. A similar model might show the same effect at a higher significance level if data were available from more locations, showing potential for the use of Craigslist ads for public health surveillance.

	HIV Prevalence Rate (8)	
Variable	Coefficient	Standard error
Intercept	537.30*	(315.18)
Mean SHR language	613.02*	(311.24)
Population	-0.00	(0.00)
Population density	0.16****	(0.03)
Count of ads	0.01	(0.01)
Mean word count	-3.62	(3.03)
State's Internet penetration	-605.46*	(348.14)
Adjusted R^2	*0.32*	

* p < .10; ** p < .05; *** p < .01; **** p < .001

Table 6. Results of Linear Regression Model, Dependent Variable = HIV Prevalence Rate Per 10,000 Population.

LIMITATIONS

There are several limitations to our analysis. First, the full dataset may include SHR terms that were not identified by our coders. Next, the circulation area of the *Village Voice* and the readership area of NYC Craigslist ads do not match exactly. Another limitation of comparing our dataset with Davidson's is that while his includes only single gay men, excluding bisexual or heterosexual MSM and couples [14], ours includes all MSM. Further research could examine the differences in SHR language between gay-identified and non-gay-identified MSM.

Although we removed duplicate identical ads, our dataset does include some similar or nearly-duplicate ads, presumably posted by the same user during our data collection period. Were a single user to generate a substantial portion of the ads in a particular location, this may compromise our analysis. Other limitations include the possibility of foreign language or spam ads, which were minimal but were not excluded from our dataset.

With large-scale computational linguistics techniques, some margin of error is to be expected. Comparing our computational coding to a human-coded random sample of 100, we identified two non-SHR ads that the computer coded as false positives, and zero false negatives. Thus, we can assume a margin of error of approximately 2%.

One important limitation is the inability of our search algorithm to evaluate the context of a term. For example, the word "clean" is often used in a SHR context. However, it can also be used to discuss hygiene, drug use, etc. While a human can usually identify whether "clean" is being used in a SHR context or not, this is not an easy task for a computer. In fact, the two false positives that were found when calculating our computational margin of error both were instances of "clean" being used in a non-SHR context. The margin of error helps to mitigate this limitation.

DESIGN IMPLICATIONS

This study argues for computational analysis of online media content as a research method to discover insights about user populations, which is useful for HCI designers in a variety of contexts. First, when designing large-scale health data systems, designers must be cognizant of the ways in which users represent their health conditions, preferences, and activities. Computational linguistic analysis of data sources such as Craigslist ads can inform the information architecture of such systems. Next, many HCI domains would benefit from mining user-generated content to answer questions about how users communicate and represent themselves. In this paper we present an example of this in a health context, but similar methods could be used in the areas of urban informatics (to inform the design of city-wide technologies), user experience (to inform the design of user interfaces), and many more.

This work also highlights temporal issues surrounding the use of language in user interfaces. We have explored how SHR language used by MSM has evolved in the last quarter century. While dropdown menus for HIV status and safe sex preferences on online dating sites would make computational analysis easier, as well as supporting serosorting and negotiation of sexual safety before meeting [37,38], such solutions are perhaps too simple if they do not take into account the evolution of language. Analysis of the constantly changing nature of language over time opens up opportunities for natural language processing and machine learning techniques to identify new SHR terms as they emerge, an area for future research. We encourage HCI research to develop tools to detect evolving language for use in user-interface design.

CONCLUSION

In this work, we examined the potential for online personal ad content to serve as a sensor for real world health behavior. This kind of low-cost, efficient sensing can augment existing data collection methods. Additionally, data collected through these mechanisms can inform systems designers about the necessary information architecture for large-scale health systems as well as the potential for online health interventions. Concretely, we analyzed sexual health-related language in a large sample of MSM Craigslist personal ads. We first built a dictionary of SHR terms, then calculated the proportion of ads with SHR content and demonstrated a rising use of SHR compared to print personal ads in the 1980s. We then built regression models to determine demographic and location-based characteristics that predict use of SHR language.

Our SHR language dictionary allowed us to examine the differences between ways that MSM currently communicate health-concerns to practices directly following the HIV/AIDS epidemic. Despite the longer length of online personal ads, SHR language is more standardized and abbreviated when compared to language documented in the 1980s. At the same time, we found a substantial increase in the percentage of ads using SHR language, from 22.99% in 1988 to 53.50% in 2013. This suggests that health-concerned language has become a core attribute of MSM personal ad content. The largest category of SHR language was disease-related, including much HIV-related language, and 3.38% of ads indicated preferences for unsafe sexual practices.

Our statistical models (1-7) identified that the prevalence rate of HIV in an ad's location was a significant predictor for use of SHR language. This suggests that the substantial increase in SHR language that occurred over the past 25 years may have been motivated by concern with HIV. These results highlight the potential for Craigslist ads to be used as a way to quickly gather information about MSM sexual health and practices for public health surveillance purposes to assist in HIV prevention. Additionally, we found that young MSM, a group with higher preference for risky behavior [25], were less likely to use SHR language in their ads, and we identified several outlying locations with high HIV prevalence but low use of SHR language. A SHR language sensor could potentially help to identify communities such as these who may benefit from HIV education and outreach.

By employing data science techniques to analyze SHR language in a large, publicly available dataset, this research demonstrates that SHR language on Craigslist can serve as a sort of sensor. This sensor provides insight not only into epidemiological public health data and sexual practices and discourse among particular communities, but into the potential for similar sensors to transform publicly available, online data into design implications. Future research could consider ways to improve such a sensor, such as by computationally identifying more information about an ad's author's demographics and sexual practices, or could ascertain new applications for the sensor. Another future research direction could contrast use of SHR language among different communities on Craigslist, to identify differences in sexual practices, discourse, and representation of online identity. Finally, research could examine the use of SHR discourse, or lack thereof, in profile-based online dating sites and mobile apps such as Match.com and Grindr, and the implications of censorship of SHR language on disease control.

ACKNOWLEDGMENTS

Thank you to Karen Cheng, Kate Ringland, Steven Frost, and members of STAR and LUCI for their generous help and support with this paper. This work was conducted as part of the UC Irvine Competitive Edge program and funded in part by a Eugene Cota-Robles Fellowship.

REFERENCES

1. Brewer, J., Kaye, J., Williams, A., and Wyche, S. Sexual interactions: why we should talk about sex in HCI. *Proc. CHI 2006*, ACM (2006), 1695–1698.
2. Brown, T. Behavioral surveillance: Current perspectives, and its role in catalyzing action. *J. of Acquired Immune Deficiency Syndromes International Perspectives on HIV 32*, (2003), S12–S17.
3. Bull, S.S., McFarlane, M., Lloyd, L., and Rietmeijer, C. The process of seeking sex partners online and implications for STD/HIV prevention. *AIDS Care 16*, 8 (2004), 1012–1020.
4. Cassels, S., Menza, T.W., Goodreau, S.M., and Golden, M.R. HIV serosorting as a harm reduction strategy: Evidence from Seattle, Washington. *AIDS 23*, 18 (2009), 2497–2506.
5. Centers for Disease Control and Prevention. *HIV in the United States: At A Glance*. 2013.

6. Centers for Disease Control and Prevention. *HIV Surveillance Report, 2011*. 2013.
7. Chan, J. and Ghose, A. Internet's dirty secret: Assessing the impact of online intermediaries on HIV transmission. Soc. Sci. Research Network, 2013.
8. Cheeseman, K., Goodlin-Fahncke, W., and Tewksbury, R. "Looking for a married hookup": An examination of personal ads posted by men seeking sex with married men. *J. of Men's Studies 20*, 2 (2012), 144–157.
9. Chiasson, M.A., Parsons, J.T., Tesoriero, J.M., *et al.* HIV behavioral research online. *Journal of Urban Health 83*, 1 (2006), 73–85.
10. Cocks, H.G. *Classified: The Secret History of the Personal Column*. Random House, 2009.
11. Cooper, A. Sexuality and the Internet: Surfing into the new millennium. *CyberPsy&Beh 1*, 2 (1998), 187–193.
12. Craigslist Terms of Use. *Craigslist.org*, 2012.
13. Crepaz N, Hart TA, and Marks G. Highly active antiretroviral therapy and sexual risk behavior: A meta-analytic review. *JAMA 292*, 2 (2004), 224–236.
14. Davidson, A.G. Looking for love in the age of AIDS: The language of gay personals, 1978-1988. *The Journal of Sex Research 28*, 1 (1991), 125–137.
15. Davis, J. Evolution of an Epidemic: 25 Years of HIV/AIDS Media Campaigns in the U.S. Henry J. Kaiser Family Foundation, 2006.
16. Fries, J.A., Ho, T.Y., Polgreen, P.M., and Segre, A.M. Using Craigslist messages for syphilis surveillance. *Int. Meeting on Emerg. Diseases and Surveil.* (2011).
17. Fries, J.A., Segre, A., Polgreen, L., and Polgreen, P. The use of Craigslist posts for risk behavior and STI surveillance. *Int. Society for Disease Surv.* (2011), 13.
18. Grov, C. Risky sex- and drug-seeking in a probability sample of men-for-men online bulletin board Postings. *AIDS and Behavior 14*, 6 (2010), 1387–1392.
19. Gudelunas, D. Online personal ads. *Journal of Homosexuality*, 49, 1 (2005), 1–33.
20. Gudelunas, D. There's an app for that: The uses and gratifications of online social networks for gay men. *Sexuality & Culture*, 16, 4 (2012), 347–365.
21. Handel, M.J. and Shklovski, I. Disclosure, ambiguity and risk reduction in real-time dating sites. *Proc. GROUP 2012*, 175–178.
22. Hatala, M.N., Baack, D.W., and Parmenter, R. Dating with HIV: A content analysis of gay male HIV-positive and HIV-negative personal advertisements. *J. of Social and Personal Relationships 15*, 2 (1998), 268–276.
23. Hoff, C.C. and Beougher, S.C. Sexual agreements among gay male couples. *Archives of Sexual Behavior 39*, 3 (2010), 774–787.
24. Kannabiran, G., Bardzell, J., and Bardzell, S. How HCI talks about sexuality: discursive strategies, blind spots, and opportunities for future research. *Proc. CHI 2011*.
25. Mansergh, G. and Marks, G. Age and risk of HIV infection in men who have sex with men. *AIDS 12*, 10 (1998), 1119–1128.
26. Martin, J.L. The impact of AIDS on gay male sexual behavior patterns in New York City. *American Journal of Public Health 77*, 5 (1987), 578–581.
27. McFarlane, M., Bull, S.S., and Rietmeijer, C.A. The Internet as a newly emerging risk environment for sexually transmitted diseases. *Journal of the American Medical Association 284*, 4 (2000), 443–446.
28. Mocroft, A., Brettle, R., Kirk, O., *et al*. Changes in the cause of death among HIV positive subjects across Europe: results from the EuroSIDA study. *AIDS 16*, 12 (2002), 1663–1671.
29. Moskowitz, D.A. and Seal, D.W. "GWM looking for sex—SERIOUS ONLY": The interplay of sexual ad placement frequency and success on the sexual health of "men seeking men" on Craigslist. *Journal of Gay & Lesbian Social Services 22*, 4 (2010), 399–412.
30. Mowlabocus, S. *Gaydar Culture: Gay Men, Technology and Embodiment in the Digital Age.* Ashgate Publishing, 2010.
31. Mustanski, B., Lyons, T., and Garcia, S.C. Internet use and sexual health of young men who have sex with men: A mixed-methods study. *Archives of Sexual Behavior*, 40, 2 (2011), 289–300.
32. Rosenbaum, M.S., Daunt, K.L., and Jiang, A. Craigslist exposed: The Internet-mediated hookup. *J of Homosexuality 60*, 4 (2013), 505–531.
33. U. S. Census Bureau. Population, housing units, area, and density: 2010 - United States -- metropolitan and micropolitan statistical areas. 2010. http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml.
34. U. S. Census Bureau. Annual estimates of the population: April 1, 2010 to July 1, 2012. 2012. http://www.census.gov/popest/data/metro/totals/2012/.
35. Village Voice. Classifieds. *Village Voice*, 1978-1988.
36. Wickman, M.E., Anderson, N.L.R., and Smith Greenberg, C. The adolescent perception of invincibility and its influence on teen acceptance of health promotion strategies. *J of Pediatric Nursing 23*, 6 (2008), 460–468.
37. Winchester III, W.W., Abel, T.D., and Bauermeister, J. The use of partner-seeking computer-mediated communication applications by young men that have sex with men (YMSM): Uncovering human-computer interaction (HCI) design opportunities in HIV prevention. *Health Systems 1*, 1 (2012), 26–35.
38. Wohlfeiler, D., Hecht, J., Volk, J, *et al*. How can we improve online HIV and STD prevention for men who have sex with men? Perspectives of hook-up website owners, website users, and HIV/STD directors. *AIDS and Behavior*, 17, 9 (2013), 3024–3033.
39. Wolitski, R.J., Valdiserri, R.O., Denning, P.H., and Levine, W.C. Are we headed for a resurgence of the HIV epidemic among men who have sex with men? *Amer. J. of Public Health 91*, 6 (2001), 883–888.
40. World Health Organization. WHO | Public health surveillance. *World Health Organization*, 2013. http://www.who.int/topics/public_health_surveillance/.