HHS Public Access

Author manuscript

Prev Sci. Author manuscript; available in PMC 2022 November 01.

Published in final edited form as:

Prev Sci. 2021 November ; 22(8): 1173–1184. doi:10.1007/s11121-021-01255-2.

Using Machine Learning to Predict Young People's Internet Health and Social Service Information Seeking

W. Scott Comulada[1], **Cameron Goldbeck**[1], **Ellen Almirol**[1], **Heather J. Gunn**[1], **Manuel A. Ocasio**[2], **M. Isabel Fernández**[3], **Elizabeth Mayfield Arnold**[4], **Adriana Romero-Espinoza**[1], **Stacey Urauchi**[1], **Wilson Ramos**[1], **Mary Jane Rotheram-Borus**[1], **Jeffrey D. Klausner**[1], **Dallas Swendeman**[1], **Adolescent Medicine Trials Network (ATN) CARES Team**

[1]University of California, UCLA Center for Community Health, 10920 Wilshire Blvd Suite 350, Los AngelesLos Angeles, CA 90024, USA

[2]Tulane University, New Orleans, LA, USA

[3]Nova Southeastern University, Fort Lauderdale, FL, USA

[4]University of Texas Southwestern Medical Center, Dallas, TX, USA

Abstract

Machine learning creates new opportunities to design digital health interventions for youth at risk for acquiring HIV (YARH), capitalizing on YARH's health information seeking on the internet. To date, researchers have focused on descriptive analyses that associate individual factors with health-seeking behaviors, without estimating of the strength of these predictive models. We developed predictive models by applying machine learning methods (i.e., elastic net and lasso regression models) to YARH's self-reports of internet use. The YARH were aged 14–24 years old ($N = 1287$) from Los Angeles and New Orleans. Models were fit to three binary indicators of YARH's lifetime internet searches for general health, sexual and reproductive health (SRH), and social service information. YARH responses regarding internet health information seeking were fed into machine learning models with potential predictor variables based on findings from previous research, including sociodemographic characteristics, sexual and gender minority identity, healthcare access and engagement, sexual behavior, substance use, and mental health. About half of the YARH reported seeking general health and SRH information and 26% sought social service information. Areas under the ROC curve (≥ .75) indicated strong predictive models and results were consistent with the existing literature. For example, higher education and sexual minority identification was associated with seeking general health, SRH, and social service information. New findings also emerged. Cisgender identity versus transgender and non-binary identities was associated with lower odds of general health, SRH, and social service information seeking. Experiencing intimate partner violence was associated with higher odds of seeking

W. Scott Comulada, wcomulada@mednet.ucla.edu.

Ethics Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Consent to Participate Informed consent/assent was obtained from all participants in this study.

Conflict of Interest The authors declare that they have no conflicts of interest.

general health, SRH, and social service information. Findings demonstrate the ability to develop predictive models to inform targeted health information dissemination strategies but underscore the need to better understand health disparities that can be operationalized as predictors in machine learning algorithms.

Keywords

HIV; Digital health intervention; Internet health information; Social service information; Machine learning

Introduction

Digital communication is reshaping how we seek out and respond to health information. Most Americans use the internet and seek out health information (Fox & Duggan, 2013; Park & Kwon, 2018). Online health information impacts patient-provider relationships (Tan & Goonawardene, 2017), healthcare utilization (Yigzaw et al., 2020), and patients' medical decision-making (Chen et al., 2018). The power of digital health information to change health-related behavior drives the dissemination of information through digital health interventions (Cao et al., 2017). For example, an HIV prevention intervention for youth at risk for acquiring HIV (YARH) in two urban HIV epicenters successfully delivered daily text messages to promote healthy behaviors (Swendeman et al., 2019). The scalability of such digital health interventions hinge on the ability to predict who will use the interventions and how the interventions will be used. In this vein, we evaluate the ability of machine learning methods to predict internet health information-seeking behaviors among YARH with more replicability of the model in future samples from the same population than is typically provided by traditional analytic strategies.

Digital health interventions and health communication may play key roles in HIV prevention (Tomori et al., 2014), providing structural HIV intervention alternatives to existing, individually delivered interventions (Aidala et al., 2016; Sipe et al., 2017). Uptake of traditional clinic-based interventions is impeded by a range of challenges, especially for sexual and gender minority youth (Simoni et al., 2015): stigma, access to healthcare, healthcare costs, histories of substance abuse or mental illness, unemployment, and unstable housing (Bowleg & Raj, 2012; Grieb et al., 2013; Reback et al., 2015). Digital health interventions and access to information on social services offers broad access at low cost to critical health information for YARH with low risk of public disclosures of identity.

Yet, optimization of digital health interventions requires tailoring for specific populations (Claborn et al., 2018; Murray et al., 2016) and an understanding of types of health information sought, commonly classified as general health or sexual and reproductive health (SRH) information (Barman-Adhikari & Rice, 2011; Comulada et al., 2020; Mitchell et al., 2014). An understanding of motivations (e.g., privacy among sexual minority youth; Mitchell et al., 2014) and barriers to internet information access across populations is also needed, such as the "digital divide" for individuals in a low socioeconomic status (SES; McCloud et al., 2016; Nguyen et al., 2017).

Most studies on internet health-seeking behaviors have reported rates for general health and SRH information seeking on the internet, as well as statistically significant characteristics associated with health information-seeking behaviors. The statistical significance of internet health information-seeking correlates reported in prior studies does not inform the degree to which the correlates predict internet health information-seeking behaviors (Lo et al., 2015). Therefore, a primary goal of the paper is to build machine learning models from correlates of internet health information seeking from the literature and evaluate model performance in their prediction of internet health information-seeking behaviors.

As a starting point to building predictive models, variables included in machine learning models are chosen based on findings in the extant literature. Given the paucity of studies on YARH, we built our machine learning models based on findings from studies in other populations. There are a number of sociodemographic characteristics associated with seeking health information: being white (in contrast to African American or Latino), female, more educated, employed, and from a higher socioeconomic status (SES) are associated with more internet searches (Barman-Adhikari & Rice, 2011; Nguyen et al., 2017; Gonzalez et al., 2016; Jacobs et al., 2017; Li et al., 2016; Nikoloudakis et al., 2018; Calvert et al., 2013; Murray et al., 2016). Youth who identify as a gay, bisexual, or other men-who-have-sex-with-men (GBMSM) also search the internet more often for health information (Barman-Adhikari & Rice, 2011; Mustanski et al., 2011). Histories of health risks are associated with greater internet health seeking, including those who are homeless, engaging in risky sexual behavior, substance abuse, binge drinking, or being in poor physical or mental health (Aref-Adib et al., 2016; Barman-Adhikari & Rice, 2011; Brown et al., 2016; Curry et al., 2016; Gonzalez et al., 2016; Metrik et al., 2016). In contrast, non-smokers, having a usual source of healthcare, and healthcare engagement are also associated with searching the internet (Gonzalez et al., 2016; Graffigna et al., 2017; Nikoloudakis et al., 2018). Such patterns are not easily predictable. In some cases, risk leads to more searching (e.g., homelessness and binge drinking) while those who have more healthcare resources are also searching more often. In addition to research that links samples characterized by a specific characteristic, there is also a literature on correlates of internet searching. For example, the combination of experiencing trauma and having poor health (De Bellis & Zisk, 2014) is associated with internet searches. Those using social networking sites are also more likely to search for health information on the internet (Holloway et al., 2014). These findings are summarized in the first two columns of Table 1. This review of the literature established a starting point for our exploration of variables associated with internet searches.

Present Study

The goal of the current study was to use machine learning to build a predictive model of internet health information-seeking behaviors. Analyses were conducted on data collected to evaluate an HIV prevention intervention for YARH. We are only aware of one other study using machine learning models to predict health information-seeking behaviors (Comulada et al., 2020). Given the exploratory nature of the work, we casted a wide net in selecting candidate predictors and included a number of measures that have been linked to correlates of internet health seeking in the literature (see Table 1). We also hypothesized that White

participants would indicate more online social service information seeking compared to non-White participants (Curry et al., 2016).

Method

Sample

Study participants were recruited in Los Angeles, CA, and New Orleans, LA, between May, 2017, and August, 2019, to participate in a randomized controlled trial to evaluate technology-based interventions to improve HIV prevention continuum outcomes. The trial is one of a several studies referred to as ATN CARES and funded by the Adolescent Medicine Trials Network for HIV/AIDS Interventions (Swendeman et al., 2019). Eligibility criteria required participants to be 12 to 24 years old, test seronegative on a rapid HIV test, and to demonstrate a history that would indicate high risk for acquiring HIV. High risk was based on statistically based factors linked to HIV, and the criteria were increasingly narrowed over time to recruit those at highest risk. Risk criteria included self-identification as sexual or gender minority, having a history of multi-substance abuse, hospitalization for mental illness, incarceration, being African American or Latino, having had sexually transmitted infections, homelessness, and being treated for drug abuse. Ethical approval for all study procedures was obtained from the Institutional Review Board at the University of California, Los Angeles (IRB #16-001674-AM-00006). All participants provided consent to participate. Further study details are found in Swendeman et al. (2019).

Table 2 shows the sociodemographic characteristics of the sample ($N = 1287$). Approximately half of the participants were recruited in Los Angeles (58%) and the other half were recruited in New Orleans (42%). The average age was 21.0 years old (range = 14 to 24). Most of the sample reported non-White race/ethnicity: 45% reported being Black non-Latinx and 29% reported being Latinx, and 19% White. Most participants were assigned male at birth (81%) and were cisgender men (73%). Thirteen percent of the participants reported being transgender or non-binary. Two-thirds of the participants reported being gay or lesbian (40%) or bisexual (24%). Half had at least some education beyond high school, 45% were employed, and 27% were students. Three-quarters of participants had health insurance and approximately half (52%) of the insured participants had Medicaid.

Approximately one in five participants reported having participated in other HIV prevention programs (21%) and having been in substance use treatment over their lifetime (20%; Table 2). A high proportion of participants had engaged in risky sexual behavior, such as condomless sex (80%), used marijuana (89%), and other substances (62%). Close to a third of the participants had experienced traumatic events over their lifetime, such as intimate partner violence (37%), and had attempted suicide over their lifetime (34%).

Measures

Trained research staff administered the baseline assessment and entered responses on secure laptops or tablets. The assessment measures entered into models as predictors and outcomes are described below.

Predictors

Sociodemographic Characteristics—Study location, categorized as Los Angeles or New Orleans, age, race/ethnicity, sex at birth, gender identity, sexual orientation, education level, employment status, and monthly income during the prior month were recorded. Participants indicated the types of places they lived in the past 4 months. A binary predictor was created to indicate being homeless (e.g., homeless shelter or on the streets) or not during the past 4 months. Participants also reported participation in a diversion program or time in a juvenile detention center, jail, or prison (i.e., incarcerated) in their lifetime.

Support Services—Participants reported received any of the following support services over the past 4 months: housing, food, clothing, toiletries and hygiene products (toothbrush, deodorant, soap, etc.), transportation (including taxi, token, mileage reimbursement, etc.), employment services, case management, mental health counseling or treatment, substance use counseling or drug treatment, healthcare insurance counseling, healthcare service navigation (peer navigator, HIV navigator, etc.), hormone therapy/hormone therapy counseling, post-incarceration or parole services, child care, or other services.

Healthcare Utilization—Participants listed their current health insurance status. A binary predictor was created to indicate being insured versus not being insured or not knowing insurance status. They indicated if they had ever participated in HIV prevention programs (not including the current study) or substance use treatment programs. Participants also reported if they had ever used pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP) to prevent HIV.

Sexual Behavior—Participants reported the number of men, women, and transgender men and women with whom they had sexual intercourse over their lifetime. They were also asked if they ever had sex in exchange for something (e.g., food, money, or protection) and if they ever had vaginal or anal sex without a condom. A “No” response indicated 100% condom use.

Substance Use—The 3-item AUDIT-C was used to identify participants engaged in heavy or hazardous drinking during the past 4 months (Bush et al., 1998). Participants indicated whether they had ever smoked cigarettes or e-cigarettes for longer than 4 months. Participants also indicated if they used marijuana and other substances over their lifetime, including synthetic marijuana, cocaine or crack, heroin, ecstasy, methamphetamine, prescription stimulants, inhalants, hallucinogens, and prescription painkillers that were not prescribed.

Traumatic Events—Participants reported if any of the following traumatic events happened to them in their lifetime: forced or frightened to participate in a sexual (oral, vaginal, or anal) act; had sex with someone who was 5 or more years older than them before they turned 16 years old; attacked or robbed (including threats); seen someone seriously injured or killed; or had a close friend or family member murdered. They also reported if they had ever been the victim of intimate partner violence (e.g., having a current or former partner slap or throw something at them).

Mental Health—The presence of anxiety and the presence of moderate to severe depression over the past 2 weeks was assessed using standard cutoff scores on the 7-item anxiety scale (GAD-7; Spitzer et al., 2006; Cronbach's alpha (α) = 0.88) and the 9-item Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001; α = 0.85), respectively. Participants also reported if they had ever attempted suicide.

Outcomes

Participants were asked the following question: "Have you ever used the Internet to do any of the following." They could check multiple responses from a list of 14 internet activities (e.g., "Look at general videos on YouTube or other video sites" and "Look for jobs"). Three yes-no binary outcome variables were created from the question to indicate participants who sought

- *General health* information, if they indicated looking for "Healthcare services (doctor, emergency room, hospital)," "Diet and nutrition information (Example: weight loss)," "Cold/flu symptoms," or "Medication information (not hormones)" on the internet (4 items; α = 0.79).
- *SRH* information, if they indicated looking for "Information about sex and sexuality," "HIV testing services," "STI symptoms (Example: fever)," "STI testing and treatment services," "PrEP/PEP," and "Where to get free condoms" on the internet (6 items; α = 0.85).
- *Social service* information, if they indicated looking for "social services (case management, mental health counseling, legal help, employment, food assistance, transportation services, legal information like how to change your name or gender marker, etc.)" through a single item.

We did not include information seeking on transgender health issues (e.g., hormone therapy or surgeries or procedures for transgender individuals) in the outcomes because the counts were too low to fit predictive models. Instead, we present the descriptive statistics.

Data Analysis

Machine learning was used to address the aims of the paper, mainly to build a predictive model. Traditional statistical approaches make it difficult to evaluate predictive performance because they use the same data to build and evaluate models. Alternatively, machine learning methods randomly split data into a training dataset to build models (75%; N = 965) and a testing dataset to evaluate performance (25%; N = 322). No more than 3% of the observations were missing for any predictors with one exception; 100% condom use (7.5% missing data) was dichotomized as 100% condom use versus less than 100% or unknown. We used data from participants without any missing observations (i.e., complete cases; N = 1287 of 1486 or 86.6% of the original sample size). Machine learning models were applied to training data to select variables that predicted the probability of having sought general health, SRH, and social service information on the internet in three separate models. The focus on variable selection was the overall minimization of error in predicting outcome responses, not on levels of statistical significance for individual predictors.

Elastic net and lasso logistic regressions were implemented to address a secondary aim to evaluate individual predictors. Both regression techniques are common machine learning approaches that produce interpretable coefficient estimates for predictors like traditional variable selection methods (Barrett & Lockhart, 2019; James et al., 2014). Lasso regression works by fitting a model to all candidate predictors and shrinking regression coefficients to zero for predictors that do not adequately contribute to error minimization. Elastic net combines the shrinkage penalties of lasso and ridge regression, which shrinks but never fixes any of the coefficients to zero. Thus, lasso typically shrinks more coefficients to zero compared to elastic net. Both lasso and elastic net procedures provide an additional benefit over traditional variable selection methods. They use a validation process (e.g., *k*-fold cross-validation) that further splits the training data into training and validation datasets to select optimal shrinkage parameters. Overfitting is minimized relative to traditional variable selection methods that uses a single dataset for variable selection. Models were fit through the *glmnet* R package (Friedman et al., 2010) using 10-fold cross-validation to select the optimal penalty parameters.

To aid interpretation of the relationship between predictors and outcomes, traditional logistic regressions that incorporated predictors selected by machine learning models were fit to the training data. Odds ratios are reported in Tables 3 and 4. Statistical significance levels are not presented due to difficulties in interpreting *p*-values for subsets of predictors selected through machine learning algorithms (Lo et al., 2015). We also note that reference categories for categorical predictors are not chosen a priori; machine learning algorithms select optimal reference groups, as well as the categories to collapse.

Model accuracy was gauged by using parameter estimates to predict internet health information and social service seeking in the test data. Predicted values were then compared to observed outcome values. Receiver operating curves were plotted to evaluate the sensitivity and specificity of predictions over a range of probability thresholds. Areas under the curve (AUC) were calculated to summarize the accuracy of predictions. An AUC of 0.5 as shown by a 45-degree line on the plot indicates a model that performs no better than chance.

Stratified subgroup analyses—Machine learning models were applied to subgroups based on gender identity, sexual orientation, and racial categories to explore differential rates of internet information seeking that have been evaluated through standard regressions and interaction effects on these characteristics in the literature. Stratified analyses were only conducted across a few subgroup categories due to small sample sizes in most categories that were not amenable to machine learning approaches.

Results

Internet health and social service information-seeking behaviors

A little less than half of the participants reported seeking general health and SRH information on the internet over their lifetime (44% and 46%, respectively). A quarter of the participants had sought social service information over the internet (26%). More than half of the transgender and non-binary participants reported having sought transgender health

information on the internet (54%; n = 91 of 170). Figure 1 is a histogram of general health, SRH, and social service health topics that participants searched for on the internet over their lifetime.

Predictors of internet health and social service information seeking

Figure 2 shows AUC for predictions of general health, SRH, and social service information seeking in the test data based on elastic net and lasso regression models. AUC above the 45-degree line (i.e., greater than .5) indicate predictability beyond chance. AUC were fairly high (.75–.77) across elastic net and lasso models for all three outcomes, indicating a reasonably high level of accuracy in predicting internet health seeking. AUC plots also help to set sensitivity thresholds for recruitment in a digital health study. For example, AUC for general health (Fig. 2) shows that a reasonably high true positive rate of .75 can be obtained for classifying individuals who seek general health information online if a false-positive rate of .375 is tolerable in screening participants. AUC were nearly identical between elastic net and lasso regression models indicating similar rates of accuracy. Lasso results are favored because more coefficients tend to be shrunk to zero compared to elastic net, which increases interpretability of the model.

Tables 3 and 4 present odds ratios from logistic models for all three information-seeking outcomes regressed on predictors selected by lasso models. We highlight results that are consistent across outcomes. Odds of seeking general health, SRH, and social service information on the internet were all lower in New Orleans compared to Los Angeles (odds ratios = 0.86–0.94). Participants reporting White race had higher odds of seeking general health (odds ratio = 1.21) and SRH information (odds ratio = 1.15) relative to other racial/ethnic groups. Black participants had lower odds of seeking social service information relative to Latinx and White participants (odds ratio = 0.77). Cisgender participants had lower odds of seeking general health, SRH, and social service information relative to transgender or nonbinary participants (odds ratios = 0.40–0.98). There were also differences in seeking information by sexual orientation. The most consistent pattern across different types of information-seeking behaviors was that heterosexual participants had the lowest odds of seeking general health, SRH, and social service information relative to participants reporting other sexual orientations. Participants who reported having health insurance had higher odds of seeking SRH (odds ratio = 1.17) and social service information (odds ratio = 1.21) relative to participants who did not have or were unsure if they had health insurance.

Odds of seeking general health, SRH, and social service information on the internet were higher for participants who had participated in HIV prevention programs, reported using PrEP and reported using PEP in the past (odds ratios = 1.12–1.43; Table 3). Odds of seeking SRH information were lower for participants who had been in a substance use treatment program (odds ratio = 0.83). Odds of seeking social service information were higher for participants who received support services in the past 4 months (odds ratio = 1.41). Odds of seeking general health, SRH, and social service information were higher for participants reporting riskier sexual behavior. For example, general health and SRH information-seeking odds were higher for participants reporting 10 or more partners (odds ratios = 1.29 and 1.40, respectively) and social service information-seeking odds were higher for participants

reporting 3 or more partners (odds ratio = 1/0.93 = 1.08). Problematic or binge drinking was associated with a slightly higher odds of seeking general health and SRH information on the internet (both odds ratios close to one). Marijuana use was associated with higher odds of seeking social service information (odds ratio = 1.28). Traumatic events, such as intimate partner violence (odds ratios = 1.06–1.29), and anxiety (odds ratios = 1.06–1.37) were associated with higher odds of seeking general health, SRH, and social service information on the internet. Moderate to severe depression was associated with higher odds of Internet social service information seeking (odds ratio = 1.74).

Subgroup analyses

AUC appeared to be similar in subgroups relative to AUC in the entire sample (i.e., .75–.77), except for among Black participants (N = 757) with lower AUC for internet general health information seeking (.63), SRH (.68), and social service information seeking (.82). Among participants identifying as a cisgender men (N = 1078), AUC was .76 for general health, .77 for SRH, and .79 for social service information seeking. Among participants identifying as gay or lesbian (N = 594), AUC was .71 for general health, .68 for SRH, and .72 for social service information seeking. Overall, AUC appeared to be similar except for the lower general health information-seeking AUC among Black participants. Analyses were not conducted for other racial, gender identity, or sexual orientation subgroups; sample sizes were smaller.

Discussion

This study demonstrated that models can be developed to predict internet health information-seeking behaviors among YARH with a fairly high degree of accuracy (AUCs = .75–.77 for internet general health, SRH, and social service information). The analytic strategy adopted, machine learning, emerges as a strategy which both validates existing HIV research and points to novel findings. This is promising for the development of future digital health interventions in HIV prevention research, especially since many of the measures selected by machine learning algorithms as predictors are commonly assessed in HIV prevention research. However, machine learning is likely to be useful far beyond HIV prevention and could be utilized as an analytic strategy for prevention researchers working in many areas. Machine learning allows us to build on the existing scientific literature, for example, by utilizing candidate predictors identified in previous studies to boost the accuracy of machine learning, such as internet experience and efficacy (Jacobs et al., 2017; Lagoe & Atkin, 2015). It concurrently has the ability to estimate the robustness of our findings.

Rates of seeking online information (44% for general health, 46% for SRH, and 26% for social service information) are in line with prior studies of youth. In Table 1, columns 3 and 4 summarize and contrast the results of this study with the existing literature and point to the novel findings that emerged from these analyses. Many of the predictors selected by machine learning algorithms were validated by prior studies and tended to be selected across models for general health, SRH, and social service information seeking, indicating common motivations and barriers in seeking online health and social service information.

Almost all of the sociodemographic predictors were found in this study that were also evident in earlier studies. However, there was additional information provided in several cases. For example, we had sufficient samples of young people who had not graduated from high school in contrast to high school graduates and those with some college. We were able to document that it is not simply that higher education is associated with more internet searches, but rather that having a high school diploma or equivalent and having higher education are associated with more internet searches. Similarly, our findings were quite particular about the relationships between employment and health searching on the internet: students, in contrast to those who were employed or not, were more likely to conduct internet searches. This study was also among the first to identify findings regarding gender diverse young people, specifically that YARH identifying as transgender or non-binary were more likely to search for health and social service information on the internet than cisgender YARH. Finally, the type of information (e.g., about diet vs. sexual health) and variations associated with different regions (e.g., Los Angeles vs. New Orleans) were unique. Each of the novel findings offers new opportunities for both shaping interventions for young people and for determining how to present information on websites aiming to influence health behaviors of young people. Unexpected findings underscore the importance of using machine learning tools to test our understanding of existing conceptual frameworks, generate new hypotheses, and develop predictive models as appropriate a priori groupings may be difficult to conceptualize.

Predictors selected by machine learning algorithms also led to new measures for consideration in the development of future algorithms to predict health-seeking behaviors. For example, predictors for PrEP/PEP utilization and participation in prior HIV prevention programs were selected in our models and have not previously been evaluated as correlates of internet health seeking. These measures are in line with access to care and health engagement measures shown to correlate with internet health seeking in prior studies (Gonzalez et al., 2016; Graffigna et al., 2017). Selection of access to care and engagement measures, along with racial/ethnic minority status and homelessness (as an SES marker), underscores the importance of considering structural technology barriers in digital health intervention design. For example, ATN CARES is delivering health promotion messages through text messages instead of smartphone apps to maximize digital access to intervention content.

We note several study limitations. First, study assessments were developed to evaluate efficacy of HIV prevention interventions. Internet information seeking was not queried in detail. Knowing the recency, frequency, and type of information seeking with greater granularity may have led to better predictive models. For example, there are likely to be key differences between YARH who seek out physical and mental health information on the internet (treated as general health information in our analyses) that would lead to better prediction if they were modeled. Furthermore, expertise in using digital communication tools, characteristics that have been shown to correlate with internet health seeking in prior research, was not evaluated in our study, and should be included in future studies. Second, analyses were exploratory. Selection of individual characteristics as predictors was not a confirmation of their utility in machine learning algorithms for future interventions. While minimized, overfitting may still be present in machine learning algorithms. This is especially

true for algorithms applied to smaller sizes such as the one used in this study. Splitting the data to build models and then evaluate performance may have resulted in small numbers of observations in some categories. The sample size also prevented the full exploration of interactions that have been shown to be associated with internet health seeking in the literature. Therefore, we used stratified subgroup analyses to evaluate whether AUC differed in response categories that would likely play important roles in the interactions. There were other response categories that are important (e.g., transgender identity) and should be explored in future research.

Conclusions

We demonstrated that characteristics assessed in an HIV prevention trial can be used to build a predictive model for internet health-seeking and social service-seeking behaviors. In doing so, we added to an emerging discussion on the utility of machine learning tools to target appropriate populations for public health interventions (e.g., Barrett & Lockhart, 2019). Data that feed machine learning algorithms are central to the discussion. Digital phenotyping harnesses passive mobile phone-based monitoring (Onnela & Rauch, 2016) and provides a tantalizing framework in which to combine passive and self-reported data collection for better prediction. Even so, a basic need remains to understand health disparity markers, some of which emerged in our findings (e.g., social determinants of health, sexual and gender minority identity), to collect better data to optimize prediction and develop quality digital health content for vulnerable populations.

Acknowledgements

The members of Adolescent Medicine Trials Network CARES are Sue Ellen Abdalian, Elizabeth Mayfield Arnold, Robert Bolan, Yvonne Bryson, W. Scott Comulada, Ruth Cortado, M. Isabel Fernandez, Risa Flynn, Tara Kerin, Jeffrey Klausner, Marguerita Lightfoot, Norweeta Milburn, Karin Nielsen, Manuel Ocasio, Wilson Ramos, Cathy Reback, Mary Jane Rotheram-Borus, Dallas Swendeman, Wenze Tang, Panteha Hayati Rezvan, and Robert E. Weiss.

Funding

CARES is a program project grant funded by the ATN for HIV/AIDS Interventions Research Program Grant at the National Institutes of Health (U19HD089886). The Eunice Kennedy Shriver National Institute of Child Health and Human Development is the primary funder of this network, with the support of the National Institute of Mental Health, National Institute of Drug Abuse, and National Institute on Minority Health and Health Disparities. Additional support was provided by the National Institute of Mental Health through the Center for HIV Identification, Prevention, and Treatment Services (CHIPTS; P30MH058107) and an HIV training grant (T32MH109205).

References

Aidala AA, Wilson MG, Shubert V, Gogolishvili D, Globerman J, Rueda S, et al. (2016). Housing status, medical care, and health outcomes among people living with HIV/AIDS: A systematic review. American Journal of Public Health, 106, e1–e23.

Aref-Adib G, O'Hanlon P, Fullarton K, Morant N, Sommerlad A, Johnson S, & Osborn D (2016). A qualitative study of online mental health information seeking behavior by those with psychosis. BMC Psychiatry, 16. 10.1186/s12888-016-0952-0

Barman-Adhikari A, & Rice E (2011). Sexual health information seeking online among runaway and homeless youth. Journal of the Society for Social Work and Research, 2, 89–103.

Barrett TS, & Lockhart G (2019). Efficient exploration of many variables and interactions using regularized regression. Prevention Science, 20, 575–584. [PubMed: 30506295]

Bowleg L, & Raj A (2012). Shared communities, structural contexts, and HIV risk: Prioritizing the HIV risk and prevention needs of black heterosexual men. AJPH, 102, S173–S177.

Brown JL, Gause NK, & Northern N (2016). The association between alcohol and sexual risk behaviors among college students: A review. Current Addiction Reports, 3, 349–355. [PubMed: 27896039]

Bush K, Kivlahan DR, McDonell MB, Fihn SD, & Bradley KA (1998). The AUDIT Alcohol Consumption Questions (AUDIT-C): An effective brief screening test for problem drinking. Archives of Internal Medicine, 3, 1789–1795.

Calvert JK, Aidala AA, & West JH (2013). An ecological view of internet health information seeking behavior predictors: findings from the CHAIN Study. Open AIDS Journal, 7, 42–46.

Cao B, Gupta S, Wang J, Hightow-Weidman LB, Muessig KE, Tang W, et al. (2017). Social media interventions to promote HIV testing, linkage, adherence, and retention: Systematic review and meta-analysis. JMIR, 19, e394. [PubMed: 29175811]

Chen YY, Li CM, Liang JC, & Tsai CC (2018). Health information obtained from the Internet and changes in medical decision making: Questionnaire development and cross-sectional survey. JMIR, 20, e47. [PubMed: 29434017]

Claborn KR, Meier E, Miller MB, Leavens EL, Brett EI, & Leffingwell T (2018). Improving adoption and acceptability of digital health interventions for HIV disease management. Translational Behavioral Medicine, 8, 268–279. [PubMed: 29385547]

Comulada WS, Step MM, Fletcher J, Tanner AE, Dowshen N, Arayasirikul S, et al. (2020). Predictors of Internet health information seeking behaviors among young adults living with HIV across the United States. JMIR, 22, e18309. [PubMed: 33136057]

Curry SR, Rhoades H, & Rice E (2016). Correlates of homeless youths' stability-seeking behaviors online and in person. Journal of the Society for Social Work and Research, 7, 2334–2315.

De Bellis MD, & Zisk A (2014). The biological effects of childhood trauma. Child and Adolescent Psychiatric Clinics of North America, 23, 185–222. [PubMed: 24656576]

Fox S, & Duggan M (2013) Pew Research Center. https://www.pewresearch.org/internet/2013/01/15/health-online-2013/. Accessed 31 May 2020.

Friedman J, Hastie T, & Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33, 1–22. [PubMed: 20808728]

Gonzalez M, Sanders-Jackson A, & Emory J (2016). Online health information-seeking behavior and confidence in filling out online forms among Latinos: a cross-sectional analysis of the California Health Interview Survey, 2011–2012. JMIR, 18, e184. [PubMed: 27377466]

Graffigna G, Barello S, Bonanomi A, & Riva G (2017). Factors affecting patients' online health information-seeking behaviours: The role of the Patient Health Engagement (PHE) Model. Patient Education and Counseling, 100, 1918–1927. [PubMed: 28583722]

Grieb SMD, Davey-Rothwell M, & Latkin CA (2013). Housing stability, residential transience, and HIV testing among low-income urban African Americans. AIDS Education and Prevention, 25, 430–444. [PubMed: 24059880]

Jacobs W, Amuta AO, & Jeon KC (2017). Health information seeking in the digital age: An analysis of health information seeking behavior among US adults. Cogent Social Sciences, 3, 1302785.

James G, Witten D, Hastie T, & Tibshirani R (2014). An introduction to statistical learning with applications in R. Springer, New York., Chapter 6.

Holloway IW, Dunlap S, del Pino H, Hermanstyne K, Pulsipher C, & Landovitz RJ (2014). Online social networking, sexual risk and protective behaviors: Considerations for clinicians and researchers. Current Addiction Reports, 1, 220–228. [PubMed: 25642408]

Kroenke K, Spitzer RL, & Williams JBW (2001). The PHQ9. Validity of a brief depression severity measure. Journal of General Internal Medicine, 16, 606–613. [PubMed: 11556941]

Lagoe C, & Atkin D (2015). Healthy anxiety in the digital age: An exploration of psychological determinants of online health information seeking. Computers in Human Behavior, 52, 484–491.

Li J, Theng Y-L, & Foo S (2016). Predictors of online health information seeking behavior: Changes between 2002 and 2012. Health Informatics Journal, 22, 804–814. [PubMed: 26261218]

Lo A, Chernoff H, Zheng T, & Lo S-H (2015). Why significant variables aren't automatically good predictors. Proceeding of the National Academy of Sciences of the United States of America, 112, 13892–13897.

McCloud RF, Okechukwu CA, Sorensen G, & Viswanath K (2016). Beyond access: Barriers to internet health information seeking among the urban poor. Journal of the American Medical Informatics Association, 23, 1053–1059. [PubMed: 27206459]

Metrik J, Caswell AJ, Magill M, Monti PM, & Kahler CW (2016). Sexual risk behavior and heavy drinking among weekly marijuana users. Journal of Studies on Alcohol and Drugs, 77, 104–112. [PubMed: 26751360]

Mitchell KJ, Ybarra ML, Korchmaros JD, & Kosciw JG (2014). Accessing sexual health information online: Use, motivations and consequences for youth with different sexual orientations. Health Education Research, 29, 147–157. [PubMed: 23861481]

Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. (2016). Evaluating digital health interventions: Key questions and approaches. American Journal of Preventive Medicine, 51, 843–851. [PubMed: 27745684]

Mustanski B, Lyons T, & Garcia SC (2011). Internet use and sexual health of young men who have sex with men: A mixed-methods study. Archives of Sexual Behavior, 40, 289–300. [PubMed: 20182787]

Nikoloudakis IA, Vandelanotte C, Rebar AL, Schoeppe S, Alley S, Duncan MJ, & Short CE (2018). Examining the correlates of online health information–seeking behavior among men compared with women. American Journal of Men's Health, 12, 1358–1367.

Nguyen A, Mosadeghi S, & Almario CV (2017). Persistent digital divide in access to and use of the Internet as a resource for health information: Results from a California population-based study. International Journal of Medical Informatics, 103, 49–54. [PubMed: 28551001]

Onnela JP, & Rauch SL (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology, 41, 1691–1696. [PubMed: 26818126]

Park E, & Kwon M (2018). Health-related Internet use by children and adolescents: Systematic review. JMIR, 20, e120. [PubMed: 29615385]

Reback CJ, Ferlito D, Kisler KA, & Fletcher JB (2015). Recruiting, linking, and retaining high-risk transgender women into HIV prevention and care services: An overview of barriers, strategies, and lessons learned. International Journal of Transgenderism, 16, 209–221.

Simoni JM, Kutner BA, & Horvath KJ (2015). Opportunities and challenges of digital technology for HIV treatment and prevention. Current HIV/AIDS Reports, 12, 437–440. [PubMed: 26412082]

Sipe TA, Barham TL, Johnson W, Joseph H, Tungol-Ashmon ML, & O'Leary A (2017). Structural interventions in HIV prevention: A taxonomy and descriptive systematic review. AIDS and Behavior, 21, 3366–3430. [PubMed: 29159594]

Spitzer RL, Kroenke K, Williams JB, & Lowe B (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166, 1092–7. [PubMed: 16717171]

Swendeman D, Arnold EM, Harris D, Fournier J, Comulada WS, Reback C, et al. (2019). Text-messaging, online peer support group, and coaching strategies to optimize the HIV prevention continuum for youth: Protocol for a randomized controlled trial. JMIR Research Protocols, 8, e11165. [PubMed: 31400109]

Tan SS-L, & Goonawardene N (2017). Internet health information seeking and the patient-physician relationship: A systematic review. JMIR, 19, e9. [PubMed: 28104579]

Tomori C, Risher K, Limaye RJ, Van Lith L, Gibbs S, Smelyanskaya M, & Celentano D (2014). A role for health communication in the continuum of HIV care, treatment, and prevention. JAIDS, 66, S306–S310. [PubMed: 25007201]

Yigzaw KY, Wynn R, Marco-Ruiz L, Budrionis A, Oyeyemi SO, Fagerlund AJ, & Bellika JG (2020). The association between health information seeking on the Internet and physician visits (the seventh Tromsø study - part 4): Population-based questionnaire study. JMIR, 22, e13120. [PubMed: 32134387]

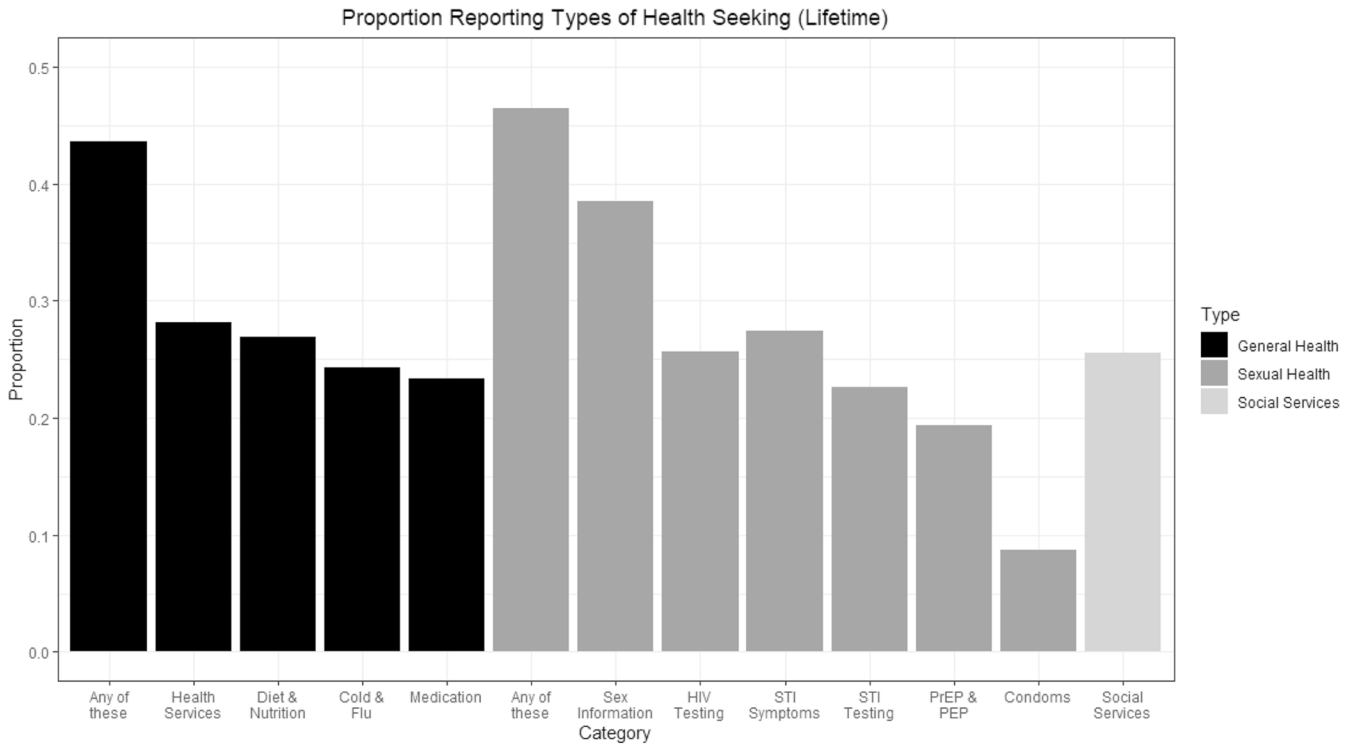

Fig. 1.

Proportion seeking different types of health and social service information on the Internet over their lifetime

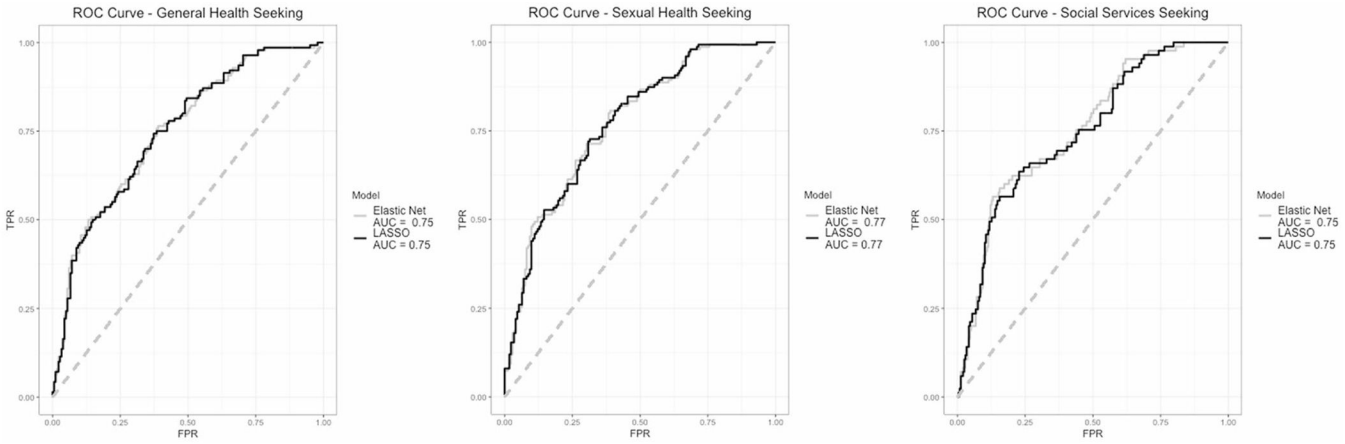

Fig. 2.

ROC for seeking general health, sexual and reproductive health, and social service information on the Internet

Table 1

Summary of existing literature examining the predictors of health seeking on the Internet, the findings from the current study that are similar, and novel findings from this study. When applicable, novel findings are described on same rows as related variables from existing literature

Variables positively associated with internet health seeking in the extant literature		**Findings replicated in this study**	**Novel observations in this study**
White	Barman-Adhikari and Rice (2011); Comulada et al. (2020); Nguyen et al. (2017)	Similar	
Higher education	Comulada et al. (2020); Gonzalez et al. (2016); Jacobs et al. (2017); Li et al. (2016); Nikoloudakis et al. (2018); Nguyen et al. (2017)	Similar	
Employment/higher SES	Calvert et al. (2013); Comulada et al. (2020); Gonzalez et al. (2016); Jacobs et al. (2017); Li et al. (2016); Nguyen et al. (2017)	Similar	
Female gender	Comulada et al. (2020[a]); Gonzalez et al. (2016); Li et al. (2016)	Similar	
Cisgender (vs. transgender)	Comulada et al. (2020)	Inverse relationship	
Gay/bisexual orientation	Comulada et al. (2020); Mitchell et al. (2014); Mustanski et al. (2011)	Similar	
Homelessness	Barman-Adhikari and Rice (2011); Curry et al. (2016)	Similar	
Having usual source of care	Gonzalez et al. (2016)	Similar	
Healthcare engagement	Graffigna et al. (2017)	Similar	Prior HIV prevention program
Self-reported ART adherence[b]	Comulada et al. (2020)	Not applicable	PrEP/PEP use
Poor physical health	Gonzalez et al. (2016); Li et al. (2016); Nikoloudakis et al. (2018)	Similar	
Poor mental health	Aref-Adib et al. (2016); Curry et al. (2016)	Similar	
Binge drinking	Gonzalez et al. (2016)	Similar	Substance abuse treatment
Not smoking tobacco	Comulada et al. (2020[a]); Gonzalez et al. (2016); Nikoloudakis et al. (2018)	Similar	
Trauma + poor health	De Bellis and Zisk (2014)	Similar	
Risky sexual behavior	Brown et al. (2016); Metrik et al. (2016)	Similar	Sex exchange
Use of social networking sites	Holloway et al. (2014)	Not evaluated	
			Geographic area (LA > NO)

The variables/subgroups listed seek more information on the Internet compared to the subgroups not listed (e.g., White participants seek health information on the Internet more often than non-White participants)

[a]Inverse relationship (e.g., male participants seek health information on the Internet more often than female participants)

[b]Self-reported antiretroviral therapy (ART) evaluated in a sample of young people living with HIV

Table 2

Candidate predictors of digital health seeking ($N = 1287$)

Characteristic	*n*	%	Characteristic	*n*	%	Characteristic	*n*	%
Sociodemographics			Sociodemographics (cont)			Lifetime substance use		
Study location			Employment status			Problematic/binge drinking, past 4 months	512	39.8
Los Angeles	752	58.4	Employed	575	44.7	Smoking, 4 months or longer	517	40.2
New Orleans	535	41.6	Unemployed	368	28.6	Marijuana use	1146	89.0
Age group			Student	344	26.7	Use of any drug, excluding marijuana	798	62.0
14–18 years	187	14.5	Below federal poverty level	914	71.0	Polydrug use, excluding marijuana	594	46.2
19–21 years	538	41.8	Support services, past 4 months	660	51.3			
22–24 years	562	43.7	Homeless, past four months	626	48.6	Traumatic events over lifetime		
Race/ethnicity			Incarceration, lifetime	324	25.2	Forced do something sexually as a child	399	31.0
Black or African American	577	44.8				Force or threats to attack/rob you	393	30.5
Latinx	377	29.3	Current health insurance status			Seen someone seriously injured or kill	632	49.1
White	249	19.3	Medicaid	505	39.3	Close friend or family member murdered	544	42.3
Asian/HPI/NA/AN	68	5.3	Medicare	109	8.5	Forced sexual act someone 5+ years older	387	30.1
Other race/ethnicity	13	1.0	Private	327	25.4	Intimate partner violence	476	37.0
Refusal/non-response	3	0.2	Other type/uncertain of type	26	2.0			
Sex at birth			Uninsured	219	17.0	Mental health, past 2 weeks		
Male	1041	80.9	Unsure of insurance status	100	7.8	Probable anxiety (GAD-7 > 7)	499	38.8
Female	246	19.1				Moderate/severe depression (PHQ-9 > 9)	409	31.8
Gender identity			Lifetime healthcare utilization			Suicide attempt(s), lifetime	437	34.0
Cisgender man	934	72.6	HIV prevention programs	272	21.1			
Cisgender woman	183	14.2	Hospitalization, mental health	381	29.6			
Transgender woman	107	8.3	Outpatient care, mental health	548	62.6			
Transgender man	63	4.9	Substance use treatment	261	20.3			
Sexual orientation			PrEP use	175	13.6			
Gay or lesbian	508	39.5	PEP use	59	4.6			
Bisexual	307	23.9						
Heterosexual	350	27.2	Lifetime sexual behavior					
Other sexual orientation	122	9.5	Number of partners					

Characteristic	*n*	%	Characteristic	*n*	%	Characteristic	*n*	%
Education level			0–2	298	23.2			
Below high school (HS)	288	22.4	3–10	468	36.4			
HS diploma/equivalent	328	25.5	More than 10	521	40.5			
Some higher education	547	42.5	100% condom use	234	19.6			
Completed higher education	124	9.6	Sex exchange	323	25.1			

GNC gender nonconforming

Table 3

Odds ratios from logistic regressions of three outcomes for having sought information through the Internet. Predictors were selected by lasso regression models. Results continue in Table 4

General health information		**Sexual reproductive health (SRH) information**		**Social service information**	
Predictor	**OR**	**Predictor**	**OR**	**Predictor**	**OR**
Sociodemographics					
Study location: NO vs. LA	0.84	Study location: NO vs. LA	0.89	Study location: NO vs. LA	0.94
Race/ethnicity		Race/ethnicity		Race/ethnicity	
African American	Ref	African American	Ref	African American	0.77
Latinx	Ref	Latinx	Ref	Latinx	Ref
White	1.21	White	1.15	White	Ref
Other	Ref	Other	Ref	Other	1.01
Male vs. female sex at birth	0.99			Male vs. Female sex at birth	0.64
Cisgender vs. transgender/non-binary	0.40	Cisgender vs. transgender/non-binary	0.50	Cisgender vs. transgender/non-binary	0.98
Sexual orientation		Sexual orientation		Sexual orientation	
Gay or lesbian	1.19	Gay or lesbian	1.36	Gay or lesbian	Ref
Bisexual	Ref	Bisexual	Ref	Bisexual	Ref
Heterosexual	0.48	Heterosexual	0.45	Heterosexual	0.82
Other sexual orientation	1.24	Other sexual orientation	1.45	Other sexual orientation	Ref
Education level		Education level		Education level	
Below high school (HS)	0.86	Below high school (HS)	Ref	Below high school (HS)	0.94
HS diploma/equivalent	Ref	HS diploma/equivalent	0.93	HS diploma/equivalent	Ref
At least some higher education	1.49	At least some higher education	1.41	At least some higher education	1.06
		Employ: Student vs. employ/not employment	1.04		
				Support services, past 4 months	1.42
Homeless, past 4 months	0.74	Homeless, past 4 months	0.60		
Current health insurance status					
Yes vs. no/unsure	1.002	Yes vs. no/unsure	1.17	Yes vs. no/unsure	1.21

OR odds ratio, *NO* New Orleans, *LA* Los Angeles, *HPI* Hawaiian or Pacific Islander, *NA* Native American, *AN* Alaska Native, *GNC* gender nonconforming, *HS* high school

Table 4

Odds ratios (OR) from logistic regressions of three outcomes for information seeking through the Internet. Predictors were selected by lasso regression models

General health information		**Sexual reproductive health (SRH) information**		**Social service information**	
Predictor	**OR**	**Predictor**	**OR**	**Predictor**	**OR**
Lifetime healthcare utilization					
HIV prevention programs	1.38	HIV prevention programs	1.40	HIV prevention programs	1.21
Substance use treat program	0.99	Substance use treat program	0.83		
Historical PrEP use	1.12	Historical PrEP use	1.31		
Historical PEP use	1.12	Historical PEP use	1.43		
Lifetime sexual behavior					
Number of sexual partners		Number of sexual partners		Number of sexual partners	
0–2	Ref	0–2	Ref	0–2	.93
3–10	Ref	3–10	Ref	3–10	Ref
More than 10	1.29	More than 10	1.40	More than 10	Ref
100% condom use	0.68	100% condom use	0.63		
Sex exchange	1.07	Sex exchange	1.16		
Substance use					
Problematic or binge drinking	1.03				
Smoking, 4 months or longer	0.86	Smoking, 4 months or longer	0.91		
				Marijuana use	1.28
Lifetime traumatic events					
Intimate partner violence	1.29	Intimate partner violence	1.21	Intimate partner violence	1.06
				Force or threats to attack/rob you	1.08
Mental health, past 2 weeks					
Probable anxiety (GAD-7 > 7)	1.37	Probable anxiety (GAD-7 > 7)	1.24	Probable anxiety (GAD-7 > 7)	1.06
				Moderate/severe depression (PHQ-9 > 9)	1.74