

Project Name Innovative Surveillance and Assessment Techniques to Inform STD and HIV Prevention Action

Project Status Awarded

Point of Contact Matthew Hogben

Center National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention

Keywords HIV, STD/STI, Syphilis, Men who have sex with men, Qualitative data

Project Description: *Please describe your project and the area of exploration. Potential project ideas can include, but are not limited to: extensions to existing tools, processes, and systems; evaluations and assessments; experiments of tests; prototypes; pilots; and development of minimum viable products to explore multi-staged projects.*

Although the efficacy of PrEP shown in studies (Spinner et al. 2015; McCormack et al. 2016) suggests it could, with ART, significantly affect the US HIV/AIDS epidemic, only 2-5% of Americans who may benefit from PrEP are currently using it (Cha, 2016). It is suggested that factors including possible side effects, stigma, access (including access to knowledgeable health care providers), and inconsistent use (behavior) may be precluding broader uptake in the United States (Cha, 2016).

In particular, we are interested in observing the effects of HIV biomedical prevention upon STD prevention, especially syphilis, concentrated among gay, bisexual and other men who have sex with men (MSM). The studies cited above, as well as cohort data from a large US city (Volk et al. 2015), show high rates of STI incidence among MSM taking PrEP. A major putative reason is the “divergence” of prevention strategies for the STD epidemic among MSM (e.g., condoms, knowing partner status) from HIV prevention. Only a few small assessments have addressed MSM attitudes pertaining to STD prevention, directly or indirectly, in the context of PrEP/ARV. In particular, there are scant data on MSM perceptions of the complex relations among (especially rectal) STD incidence, which often predicts HIV incidence, using STI as an indicator for PrEP, and then the effect of PrEP on STD prevention. While conducting timely monitoring of such factors can inform policy or provide formative data for interventions, it is often expensive and time consuming to conduct traditional surveys, particularly for hard to reach populations.

In response to traditional surveillance and assessment challenges, we propose a pilot project to evaluate Reddit as a novel, untapped data source to monitor behaviors and attitudes relevant to PrEP, and to explore networks of attitudes and beliefs surrounding the issues of STD prevention in the context of PrEP. Reddit is a website where registered members can submit content in an online bulletin board system fashion. While Reddit has been used in other public health monitoring realms such as tobacco and alcohol use, it is yet to be utilized specifically in HIV/STI-related research (Cole et al. 2016; Tamersoy et al. 2015; Wang et al. 2016; Chen et al. 2015). Reddit data holds potential in shedding light into attitudes, beliefs, and behaviors related to PrEP in that data can be harvested in near real-time, providing a high volume of anonymized opinions of the public in this topic. While this pilot project will analyze retrospective data from 2015, the capacity for real-time surveillance exists if value is illustrated from this pilot.

Epidemico Inc. has harvested retrospective Reddit posts and comments from pushshift.io, a collection of public Reddit data that includes posts and comments dating back to October 2007. Pushshift.io uses

Reddit's Application Program Interface (API) to collect submissions; posts and comments are made available in newline JSON format. To demonstrate the volume of relevant conversations, in 2015 there were 2,340 comments mentioning PrEP and 1,671 mentioning Truvada. Further, there were 11,250 mentioning syphilis, 11,281 comments mentioning chlamydia, and 7,096 mentioning gonorrhea. We are also able to look specifically in subreddits (niche forums) that are of interest such as "askgaybros", "gaybros", "hivaid", or "PrEPared".

Following data collection, we will apply natural language processing (NLP) and machine learning (ML), two of Epidemico's core competencies, to the data. ML and NLP will support the review and analysis of Reddit content by extracting targeted information to distill meaningful insights from the large dataset. Epidemico has extensive experience in applying such technologies to infectious disease monitoring, adverse event detection, vaccine sentiment tracking, and various other public health topics (Powell et al. 2016; Freifeld et al. 2014; Brownstein et al. 2009). This experience, coupled with NCHHSTP's subject matter expertise, will enable the project team to devise optimized methods to analyze this novel, high volume datasource. Automated data processing will be followed with further curation and analysis by NCHHSTP to interpret this nuanced data and inform policy and/or interventions regarding PrEP and other STI prevention.

The goal of this proposed project is to assess whether insights can be gathered from Reddit to help inform relevant policies and interventions. Such insights pertaining to PrEP attitudes, experiences, and behaviors, collected from Reddit, could provide an expedited, innovative, cost-effective approach as compared to traditional surveys. Furthermore, social media listening overcomes other challenges tied to limitations in reaching remote populations and selection bias. With success of this pilot, we can envision a platform for real-time monitoring of Reddit data on these topics, to enable ongoing and timely formative data collection to inform NCHHSTP policy and intervention.

This project will be a collaboration between The CDC's National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) and Epidemico Inc. Originally spun out of Boston Children's Hospital, Epidemico is now a wholly-owned subsidiary of Booz Allen Hamilton known for pioneering global population health solutions that disrupt traditional detection, reporting, and engagement systems. Epidemico provides early insights, continuous monitoring, and consumer engagement for a wide range of population health domains, including disease outbreaks, drug safety, supply chain vulnerabilities, and more.

Research Question Topics

Topics fall into two categories: discussion around PrEP per se and discussion around STD prevention in the context of PrEP. As the data comprise very large quantities of text from thousands of persons posting on Reddit sites, the topics are themes we will explore.

- Looking at PrEP (Pre-exposure prophylaxis/Truvada/tenofovir/emtricitabine specific Reddit comments from 2015 to gain insights on: Side effects of PrEP, social stigma, high risk behavior associated with PrEP, diversion (black market distribution), adherence
- Inclusion of STIs: How often are STD (e.g., syphilis, gonorrhea, herpes) and STD prevention (e.g., condoms) mentioned in discussion of sexual behaviors? How do discussions reflect MSM attitudes or intentions related to preventing STD transmission or acquisition in the context of PrEP use? Are there discussions of STD prevention unrelated to HIV prevention?

Output/deliverable Ideas/suggestion

- Creation of a simple “data explorer”, which would include aggregate description data and a drill down to individual posts
 - Temporal visualization of mentions of PrEP and STIs
 - Visualization of the distribution of mentions across ‘digital spaces’ (e.g. isolating posts made in subreddits dedicated to conversations between MSM, women, etc.)
 - Visualization of relationships between topics (e.g. co-posting of PrEP and STI mentions)
- A summary report of major findings, and anticipated manuscript for a peer-reviewed journal

For more information about this project, please contact the CHIC at chiic@cdc.gov or Maria Michaels at maria.michaels@cdc.gov.