Project Summary

Mining Pre-exposure Prophylaxis trends in social media

One specific application will focus on the use of antiretroviral therapy to preventatively treat individuals who are at-risk for HIV infection. This use of antiretroviral therapy for preventative treatment is referred to as Pre-Exposure Prophylaxis (PrEP) and uses the small molecule nucleotide reverse transcriptase inhibitor trade named Truvada. PrEP is highly effective at preventing HIV, but because it is a new treatment there remain some social and medical obstacles preventing it from achieving its full potential. In this section we use various supervised and unsupervised text mining techniques to identify relevant keywords, hashtages, positive tweets and negative tweets in Twitter data that can be used to gain insight into the social concerns surrounding PrEP adoption. By addressing these concerns, we can make PrEP more effective and improve HIV prevention efforts.

Combining Social Media and Phylogenetic information to infer HIV outbreak dynamics

One of our goals to better understand HIV outbreak dynamics relies on the ability to combine hard medical data and social media data to monitor and predict social media analyses. In this section we propose work similar to previous work conducted in the San Diego regional health system, in which we infer an HIV pylogenetic transmission network from anonymity genetic medical data in the northern Georgia health system, and associate this with geolocation data taken from social media over the same time period. If successful, this project would measure the network of HIV transmission in the northern Georgia health system, link these dynamics to keywords and metadata present in social media data geotagged to the same locations, and determine the regional risk of drug resistant serotypes.

Determining associations between personality and drug perception on social media

Our previous work has uncovered complex social sentiments, favorable and unfavorable, to the PrEP drug Truvada. Our work and other work suggests that these social sentiments that may be driving adoption of and adherence to Truvada. Personality, commonly measured with the use of the Big Five profile, has been found to be predictive of thoughts and behaviors. Other researchers have used personality to predict disease outbreaks using social media, and adherence to drugs using traditional survey data. However, we have found that the direct association of personality and drug perceptions using social media data is lacking. Here, we suggest a project that would investigate the association between personality and perceptions of drugs from three major categories, psychiatric drugs, pain killers, and HIV related drugs including Truvada. By studying these associations, we can determine social concerns and barriers, that while not directly related to the medical functionality of these drugs, are critical to ensure that preventable and treatable diseases are optimally combated.

Project Description

1 Mining Pre-Exposure Prophylaxis Trends in Social Media

Pre-Exposure Prophylaxis (PrEP) is a recently developed method for the prevention of Human Immunodeficiency Virus (HIV) via the administration of an oral pharmaceutical trade named Truvada. Truvada contains active ingredients tenofovir and emtricitabine, both Nucleotide Reverse Transcriptase Inhibitors (NRTIs). In the last four years, since Truvada was approved for PrEP in 2012, PrEP has shown demonstrated efficacy at preventing HIV for HIV negative individuals in serodiscordant relationships[17]. Though existing methods of effective HIV prevention exist, data shows that they may not be used in case of unexpected sexual contact or personal preference[21]. PrEP is also well suited to individuals in socially or economically underprivileged groups, and for the HIV negative partner in serodiscordant couples[24]. In experimental studies, PrEP has been studied as a method to safely conceive a child[15].

Initial studies of PrEP have shown that it is highly effective[9], however because it is still a new treatment, it is facing a number of medical and social obstacles before it reaches full adoption. Incomplete clinical and patient knowledge, social stigma, and uncertain insurance status have been identified as challenges preventing continued adoption[5]. Also, since Truvada is an oral NRTI, it must be taken daily. Cases where patients do not adhere to their full prescription have led to loss of viral protection, and some patients and clinicians worry that lack of adherence could lead to increased risk of infection with drug resistant strains[4].

The goal presented in this proposal is to improve PrEP adoption and efficacy by reviewing PrEP patients and other HIV community members perceptions of the challenges and successes that have taken place during PrEP's initial adoption. While local clinical monitoring has uncovered some challenges facing PrEP efficacy and adoption[22], large scale data mining has the potential to capture a broader perceptive than local clinical reports can. In addition, by mining the massive data sentiments present on social media platforms, researchers are able to capture unfiltered opinions, and can update disease monitoring analyses in real time. Previous work has used Twitter to predict county-level HIV prevalence[26], and general HIV discussion monitoring[25]. However to our knowledge, prior to the previous work described in the next subsection, data mining social media specifically to determine thoughts and sentiments surrounding PrEP has not been performed.

1.1 Previous Work

Our goal in this previous research was to determine perceptions and sentiments of PrEP using social media, in order to address challenges related to PrEP adoption and efficacy. Though we have investigated the use of multiple social media platforms including Facebook, Grindr, Reddit and Twitter, we have focused primarily on the use of Twitter for past projects and intend to use Twitter for future analyses. Twitter is especially useful because in addition to textual data, various useful metadata is also available including datetime, geolocation, username, hastags, and external hyperlinks. Twitter also features a convenient well documented and free API and has over 300 million monthly active users.

In this analysis we collected over 1 million tweets from the Twitter streaming API filtered on PrEP and HIV related keywords. By using embedding techniques such as Word2Vec and Doc2Vec, we were able to identify new keywords and hashtags that are relevant to the PrEP conversation on

Table 1: Cosine similarity to document-vector “#PrEP”

Related hashtag/tweet	Cosine similarity to #PrEP
#lgbtmedia16	0.739128
#hiv	0.727602
#whereisprep	0.707165
#truvada	0.696113
#hivprevention	0.636068
tweet-702179860983189504	0.630055
user-711275699529764864	0.629254
tweet-708519265540907010	0.628778
tweet-712032637024653313	0.628646
#harrogatehour	0.628547

Twitter including unexpected political connections ”NancyReagan” and popular hashtags associated with PrEP that might not have been known to researchers in advance such as ”#whereisprep” (see table 1 for examples of hashtags and structurally-related tweets). In addition Doc2Vec can be used to query the top N tweets related to a given hashtag or the top N users related to a given hashtag. This allowed us to identify a small subset of the N tweets most structurally related to ”PrEP” that can be human-read without requiring humans to read the full dataset. Reading from this set of structurally important tweets, we uncovered important blog articles exploring some of the fears of drug resistant forms of HIV and concerns as to whether these strains could be caused in part by over use or misuse of PrEP medications. The linked blog articles that we found highlight the usefulness of the hyperlink metadata embedded in tweets. Through these hyperlinks, Twitter acts as an index, providing indirect access to a much larger external social media ecosystem.

Our Doc2Vec results also allowed us to identify the top N users most relevant to PrEP. By querying the Twitter REST API for these users’ timelines, and using topic modeling methods like Latent Dirichlet Analysis (LDA), we identified the word-distribution-topics present in the Twitter conversation (see figure 1). This analysis went beyond the simple keyword identification analysis from Word2Vec since it clustered keywords into topics and it operated on all tweets that PrEP-related users tweeted, not just PrEP related tweets. The LDA results showed a variety of related concerns such as other sexually transmitted diseases, LGBT related topics, health insurance and political topics. Neither PrEP or Truvada were present in the top 30 keywords related to HIV/AIDS demonstrating that PrEP is still a nascent rare topic in the online discussion. An extension to LDA, Dynamic Topic Modeling (DTM), was able to capture topic and word frequency over time. The DTM results showed that the keyword ”PrEP” is increasing in relative frequency over time, even relative to related words such as ”pill”, ”prevention” and ”drug”. This demonstrates increased interest in PrEP which may correlate with an increase in PrEP adoption over the data acquisition period.

One of the key terms identified through LDA topic modeling was ”stigma”, a concern that is a known issue impeding PrEP adoption in previous studies[17]. Stigma is an issue with sexually transmitted diseases in general, including HIV, resulting in many HIV positive indivisuals not knowing their status, and thus may not be aware that they are transmitting the disease when they come in sexual contact with HIV negative individuals. Previous studies have also suggested that

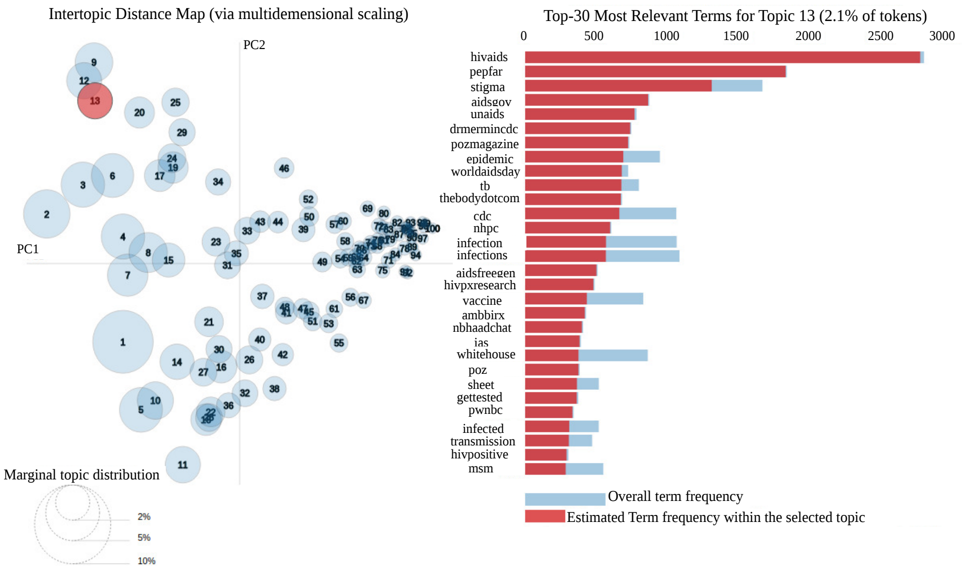

Figure 1: LDA topic modeling for the top 500 users related to PrEP.

stigma may contribute to issues surrounding PrEP adherence[5].

Using an open dataset of tweets which were labeled with binary sentiment labels, we performed a semi-supervised classification sentiment analysis. In the sentiment analysis, 5513 tweets were provided by Sanders Analytics. These tweets were human labeled either positive or negative. The sentiment analysis was performed by converting the tweets to numerical vectors using the unsupervised Doc2Vec embedding method. These numerical vectors were then classified according to the sentiment labels, and the trained model was used to infer sentiment labels for each of our PrEP-related tweets.

The sentiment analysis allowed us to identify N PrEP related tweets with the highest sentiment, and N PrEP tweets with the lowest sentiment. After performing the automated sentiment analysis, a human was inserted into the analysis loop to quickly read the top positive and negative tweets to get a sense of the most important successes and concerns present in public PrEP perception. In the positive tweets we found hyperlinks to blogs with positive firsthand accounts from individuals successfully using PrEP to stay HIV negative. In the negative tweets we found concerns of whether Truvada can protect against drug resistant strains of HIV (example negative tweets shown in table 2). Other concerns suggested that over prescription of Truvada could give individuals a false sense of security, and lead to a rise of non-HIV sexually transmitted diseases. These results show that patients, and the public at large need to be educated on the specific risks of drug resistant strains.

The issue of individuals gaining a false sense of security towards other sexually transmitted diseases is being addressed though clinical education for individuals that are prescribed Truvada, though these results show that this information is lacking in some individuals, presumably not taking Truvada. This implies that the dissemination of Truvada educational information through a platform like Twitter may improve overall PrEP-related knowledge and address some of the concerns and stigmas currently associated with PrEP. Together these approaches and results show that we can take raw text and metadata and extract keywords, hastags, temporal trends, and sentiment information. Doc2Vec and sentiment classification allow the researcher to extract a set

Table 2: Negative sentiment tweets.

Category	Text
General	"Also, how f***ing vile of Hillary to say. Reagan did f***ing NOTHING during the AIDS epidemic until it was too late. What a stupid old hag."
General	"I wonder why he beat her a** when she was tryna leave like she wasn't gone be running back when she found out she had HIV & nobody want her"
General	"Aaannd. Hillary Clinton breathes a sigh of relief that Twitter has left its outrage of her AIDS comments behind to tend to Drumpf debacle."
PrEP specific	"RT gaston_croupier #Truvada patent's not expired yet but it is sold online as a generic drug? There's something rotten in internet #PrEP h"
PrEP specific	"Equality_MI Syph & Hep C have gone up 550% in Gay Men bc many feel tht bc they're on PrEP, they don't need condoms. HIV isn't the only STI."
PrEP specific	"Xaviom8 in interviews he says he was adherent. strain was highly resistant, and Truvada wouldn't have blocked it anyways. PrEP didn't fail."
Truvada specific	"not surprised at all that someone got HIV on truvada. people get pregnant on birth control. tomato-condoms are still important-tomahto"
Truvada specific	"Now reading that truvada does not protect against certain strains of the HIV virus. Yet people want to take that risk.."
Truvada specific	"I think I have conjunctivitis unless truvada cured it overnight cuz im not feeling as horrible today as last night"

of the N most highly relevant tweets from a large corpus that can be easily human-readable.

Though our previous research on data mining social media has uncovered important tweets, keywords, sentiments and hashtags related to the Twitter discussion of PrEP, many important questions remain poorly understood. One important issue is determining why patients stop adhering to their PrEP medication. While our LDA results uncovered "stigma" and other related keywords, and some of the critical tweets we identified described uncertainty in the efficacy of PrEP, this question still remains to be fully answered. One of the challenges with using Twitter as a data source is that we cant verify personal information for the people authoring the tweets. For example we don't know if they are taking Truvada, or whether they are HIV negative or positive, or other important details of their medical status. We also don't know if misinformation or excessive negativity is being spread by uninformed individuals, or by nefarious individuals. By incorporating medical data we could potentially identify direct connections between sentiment and written opinions with more concrete medical outcomes, though such connections would fall under The Health Insurance Portability and Accountability Act of 1996 (HIPAA) and require patient consent and/or medical board approval. As of the time this previous work was carried out we were unable to access such data.

1.2 Proposed work

There are many ways that we can continue to analyze social media data to identify specific issues impeding HIV prevention efforts. One extension we propose is to investigate additional diseases or behaviors that could correlate with risk of developing HIV. One of the overall goals of biosurveilance is to predict disease outbreaks before they happen in order to intervene preventatively. An outbreak in Scott County Indiana in early 2016 was thought to have been caused by drug usage. In a town of 4,000 people 135 people were diagnosed with HIV, and about 80% of those diagnosed were codiagnosed with hepatitis C. In theory some online social activity may have been able to indirectly identify this outbreak before it happened[8]. In order to identify and predict these hot-spots we can make use of the geolocation metadata, and also do network analyses to identify subgroups, and how information and via inference, social interactions propagate through these subgroups. Some first steps along these lines have already been made by our collaborators [26], who used drug-related terms to predict HIV prevalence at the county level.

In order to extend this work, we could use the presence of other disease keywords known to coinfect with HIV as a basis for predicting HIV prevalence. This proposal is facilitated since we already have a Twitter corpus that contains tweets mentioning HIV, hepatitis B and C and about 30 other infectious diseases. Searching the HIV literature shows that the use of Twitter mentions of Hepatitis C (codiagnosed at a rate of 80% in Scott County[1]) and/or other STDs to predict HIV prevalence are lacking. Investigation of the Hepatitis-HIV connection is also particularly important because it allows us to test the concern that we identified in our previous work, that overuse, or misuse of PrEP could lead to increased prevalence of Hepatitis C. One way to test and measure this criticism of PrEP, would be to try to predict Hepatitis C prevalence using mentions of PrEP on Twitter at a county level. If this prediction succeeds, then we have evidence that confirms and quantifies this criticism, and can seek to address how PrEP is administered to prevent coinfection with Hepatitis, however if this prediction fails, we have shown that PrEP is indeed not leading to increased levels of Hepatitis, and thus we will have strengthened the case for PrEP adoption. Quantification of the predictability of the Hepatitis C connection would help potential patients and clinicians determine the benefits and risks associated with using PrEP as the primary HIV protection method.

Finally, we would like to mention that our quantitative approaches and computational pipelines for mining qualitative sentiments surrounding disease treatment provide an important contribution by themselves to the larger data science community. In our previous work we have shown a specific application where we mine social sentiment to identify what is working and what the challenges are for PrEP, but a similar framework could easily be taken and applied to improve the social barriers surrounding some other disease treatment like cancer and chemotherapy. Pharmaceutical companies, academic researchers, and hospitals can use our open source code with minimal modification to monitor their disease and treatment of interest to monitor and improve the outcomes and happiness of their patients. We anticipate that the computational approaches produced during the course of the proposed work will also demonstrate useful methods that can be applied to other areas of public health research.

2 Combining Social Media and Phylogenetic information to infer HIV outbreak dynamics

Our previous work mentioned in section 1 of this proposal relied exclusively on social media data, which prevented us from quantifying direct medical outcomes of how sentiment affects PrEP usage and HIV protection. Specifically, we uncovered concerns of drug resistant strains of HIV, which PrEP would not provide protection for, and concerns about the spread of other diseases, such as Hepatitus C, as a result of improper use of PrEP. Previous work on small clinical cohorts in San Francisco, 2 Hepatitis C infections among 485 PrEP patients, has quantified the incidence rate at 0.7 per 100 patient years[23]. Work based on simulation has predicted that use of PrEP may lead to an increase in drug resistant HIV by 9-40%[2]. In order to test and quantify these concerns, we need medical data, and in order to be most relevant to our local medical region, that data should come from the north Georgia medical system.

Though we are not aware, from searching the medical and scientific literature, of genomic and epidemiological data being collected from HIV patients in the north Georgia area, we do know of a study conducted in San Diego[16] and Chicago[19], that used HIV genomic sequence data from 478 infected individuals to infer a phylogenetic infection network. These researchers also associated epidemiological data, such as age, sex, risk factors, cell count, and viral load with the phylogenetic infection networks, see figure2. By examining the genomic information, these researchers were able to identify the incidence of drug resistant forms, and how HIV subtypes spread in the local area over the collection period. The SanDiego study examined the affect of anti-retroviral therapy (ART) on the propensity to generate an infection, finding significantly less transmission when ART was started within the first 12 months. The Chicago study found that HIV transmission happened sporadically throughout the city with no correlation to the individual's region of residence implying that transmission encounters happen far from home. Together, these two studies give information on the factors and aspects driving HIV transmission, and the efficacy of certain treatments, though since these studies were performed prior to Truvada's FDA approval in 2012 (1996-2011 and 2005-2011 respectively), there was no investigation of PrEP.

In both the Chicago and San Diego studies, the pol region of HIV-1 was sequenced as a routine assessment of potential drug resistance. To generate the pylogenetic networks, these sequences were aligned using multiple sequence alignment, and while different pylogentics software was used in the two papers, both papers used a cutoff of 1.5% genetic sequence distance to determine an edge connection between individuals. Date of infection was able to determine, in many cases, a directionality for edges in the transmission network.

2.1 Proposed work

To our knowledge, no one has yet combined genomic, epidemiological and social media data to infer and quantify the direct medical outcomes of PrEP usage or link social media sentiments surrounding PrEP to direct HIV medical outcomes. Furthermore, we are not aware of any complex HIV network studies being conducted on the population of the northern Georgia. Thus to better understand the effects of PrEP, and quantify the direct risks and concerns surrounding PrEP in the north Georgia region, we propose a project to combine genetic, and epidemiological data with social media data to determine the interactions between PrEP usage, PrEP social media sentiment, and incidence of drug resistant HIV and Hepatitis C.

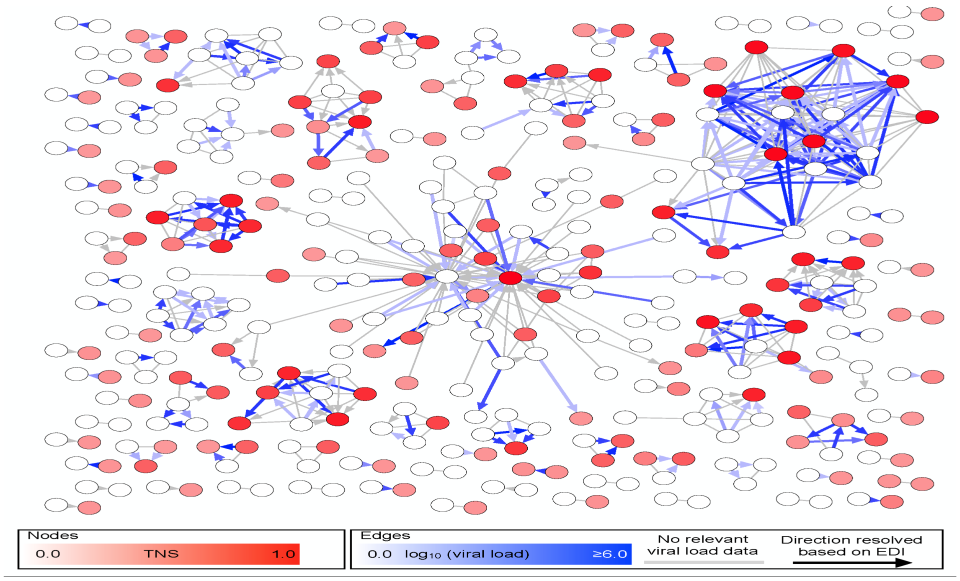

Figure 2: Phylogenetic network from San Diego study[16].

We will acquire anonymized sequence and epidemiological data for HIV from the north Georgia region, and use standard phylogentic tools to quantify the presence of drug resistant strains and construct a phylogenetic network using methods described in [16]. We will gather the epidemiological and network details, including geolocation trends, individuals gender, race, risk factors and cell counts. A simple linear correlation of all of these variables with keywords from social media, and the corresponding p-value for these correlations, binned by geographical area, would allow us to determine the social media keywords most strongly associated with each of network and epidemiological variables.

One potential issue in this proposal, is how to make connections between individuals' medical information and social media. This is made especially hard since to make such a connection would violate ethical concerns, and also HIPPA privacy laws. How researchers have gotten around this concern is to bin social media and medical patients into geographical regions and then ask what are the differences or correlations between the regions. For example in the Chicago study[19] the researchers separated the city into regions to compare and determine at-risk locations, see figure3. This requires that our data be geotagged in some way. Conveniently the Twitter social media that we have is tagged down to the latitude and longitude coordinates, allowing us to bin by county, by city or by state depending on the precision of the geolocation data attached to the medical data.

In addition to linear correlations by geographical area, we also want to propose some more complicated network methods. One such method, would be to use the infection transmission graph to perform spectral clustering on the geographical regions (we use connections between individuals in different regions to infer connections between geographical regions). This produces a clustering of regions by connectivity. We then ask whether these clustered regions are similar by their social media activity. Hypothesis testing by comparing this clustering to a randomized clustering (performed by simulation) would tell us whether or not the connectivity between regions affects their medical or social media attributes. Intra-regional connectivity is another variable that can be correlated with social media terms in the simple linear correlation proposed above. If successful,

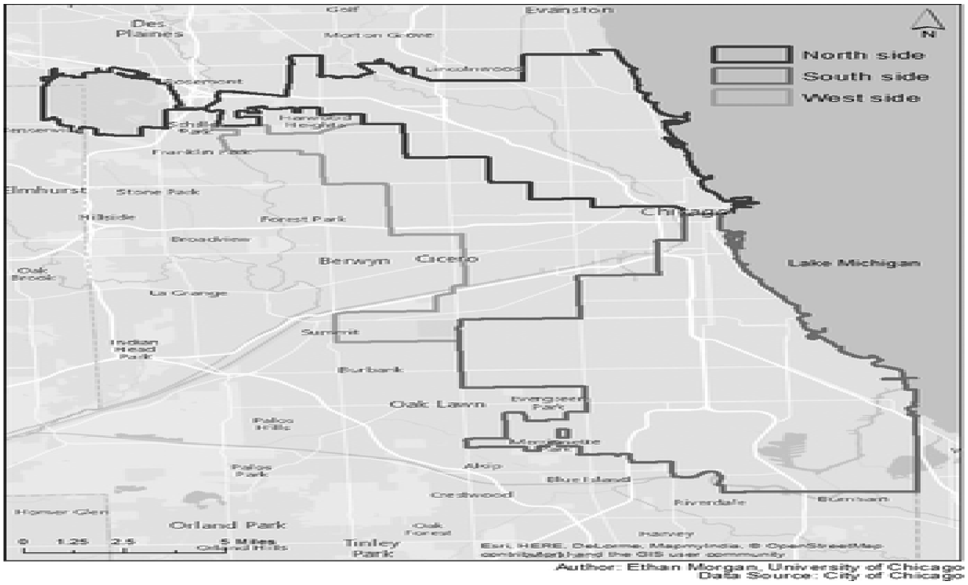

Figure 3: Regions of Chicago from [19].

this research would demonstrate an underlying mechanism, the network interactions, affecting HIV transmission. It has been hypothesized and many simulations have been done predicting the affect of giving transmission hub, high risk individuals PrEP in an effort to prevent HIV for the overall region and this network analysis could provide an more precise empirical result to calibrate those simulations.

The primary concern that would threaten this plan would be the inability to acquire data, either because it has not been collected in our geographical region, clinicians are unwilling to collaborate, or some regulatory issue prevents us from gaining access to this data. If this happens, and we are unable to get recent HIV sequence data that would match our social media data by collection dates and locations (our PrEP-related social media data was collected nationwide November 2015-ongoing), one solution would be to request social media data from a date range and location for which HIV genetic and epidemiological data is available.

In the scenario that we are still unable to acquire medical and social media data from the same set of dates and geographical location, another thing that we can do is attempt to acquire prescription data instead. Our collaborators have mentioned before that they may be able to acquire prescription data that contains anonymized information on individuals purchasing various prescription drugs. This would not allow us to directly infer transmission networks, or quantify the rate of drug resistant forms of HIV, but it would provide some medical data to associate with social media data. For example we could use the prevalence of HAART therapy as a proxy for drug resistant HIV prevalence, other ART as a proxy for general HIV, and Hepatitis drugs as a proxy for hepatitis. We could link the levels of these prescriptions, with levels of Truvada prescription and social media sentiment keywords over the time period.

3 Determining associations between personality and drug perception on social media

Our initial PrEP research mentioned in section 1 of this proposal, and work by other researchers, has demonstrated that social media can be used to monitor and identify issues of drug adherence, usage concerns, and attitudes of patients, potential patients, and other interested parties using social media data. These subjective thoughts and attitudes are important since while they do not capture direct medical outcomes, if patients, potential patients and other interested parties aren't subjectively satisfied with their treatments, this provides an important avenue to guide improvements to the treatment. Personality may also provide some explanation for adoption and adherence to prescriptions, which greatly influences medical outcomes.

In order to further categorize individuals attitudes and perceptions in social media data, researchers have made use of quantitative psychology profiles. The most standard profile, the big five personality score[10], has demonstrated an ability to predict a variety of attitudes and behaviors, and has since been accepted as one of the best currently available model of personality. The big five profile is composed of extraversion, a measure of sociability, agreeableness, a measure of social trust, conscientiousness, a measure of impulse control, neuroticism, a measure of emotional instability, and openness, a measure of adventurousness.

Researchers have used the big five personality profile in conjunction with social media to study addiction[3], and predict suicide rates[14]. In the context of prescription drugs, one study found that extraversion correlated negatively with levels of drug adherence of antidepressants using medical surveys[7], though this study did not use social media. The study found that extraversion was found to be a more significant predictor than severity of depression symptoms. Other work has used the big five personality to predict drug and alcohol abuse[18] using data collected from surveys posted on social media. This study found that extraversion and conscientiousness were associated with increased drug and alcohol usage.

Recently, studies have used big five personality information in social media data to predict HIV prevalence in a regional geographical area using social media data [13]. In this research, the county-level HIV prediction model showed that contentiousness, or future-orientation, associated negatively with HIV prevalence, demonstrating an indirect connection between risk taking language posted online and risk of HIV infection.

Despite these previous efforts, little of the prevailing research has been conducted from the perspective of trying to identify concerns towards prescription drugs in order to understand and improve patient treatment perceptions. Thus the goal proposed in this section aims to build on this research area by using Twitter data to determine associations between a set of popular medications including antidepressent drugs, pain drugs, and STD treatment and prevention drugs. By combining personality and sentiment information, this will allow us to determine aspects of personality that are favorable or unfavorable to disease treatments. We have chosen to focus on these specific drugs in order to build on established connections between mental health drugs and personality types, and also to continue to investigate psychological motivations related to PrEP adoption and adherence.

3.1 Previous work

We acquired keywords associated with personality types from previous researchers [20]. These researchers associated words with personality scores from 75,000 facebook users who took a big five personality test, and published the resulting keywords correlated with each of the five personality categories. For sentiment data, we used a dataset provided by Sanders Analytics that contained tweets that were human-labeled as positive or negative sentiment. We also acquired a set of tweets containing at least one popular psychiatric, pain, or HIV related prescription drug in the top 100 drugs by sales.

We trained a word2vec model on the sentiment and prescription drugs, inferred sentiment labels for the prescription drug tweets, and separated the prescription drug tweets into positive and negative sentiment corpses. Then we constructed heatmaps showing the word2vec similarity between the prescription drug names and psychology keywords. For each of the 5 psychology categories, we used five keywords most positively associated with that category. We constructed heatmaps showing the overall word2vec similarity between keywords, and a heatmap showing the word2vec similarity from the positive corpus minus the similarity from the negative corpus. In both cases the rows (prescription drugs), were clustered using hierarchical clustering.

For the psychiatric drugs, we found that extraversion, agreeableness and openness terms tended to be associated with most of the psychiatric drugs in the positive sentiment tweets relative to the negative sentiment tweets, while neurotocism tended to have the opposite pattern4. This makes sense based on the relative sentiments broadly associated with these psychological categories. In the same heatmap, we found that Pristiq and Strattera, two Serotoninnorepinephrine reuptake inhibitor (SNRI) class drugs[12], clustered closely together, despite commonly being prescribed for different conditions, depression and attention deficit hyperactivity disorder (ADHD)respectively. The other ADHD drug in our dataset, Vyvanase, is an amphetamine class drug and clustered far from Strattera, perhaps demonstrating that these drugs have very different sentiments in the social media discourse. Our data show some evidence that of the two drugs, Strattera is the one with more positive sentiments, especially in the categories of extroversion, agreeableness and openness. Ongoing and future work can uncover further relationships.

The HIV treatment related drugs show a similar high level pattern, with all three drugs disproportionately positive in relation to extraversion, agreeableness and openness terms5. This provides further evidence for a high-level trend relating drug sentiments and these broad psychological categories. The clustering pattern shows Truvada and Atripla, nucleotide reverse transcriptase inhibitors (NRTI's) clustering together, and Norvir, a drug containing a protease inhibitor clustering farther away. Norvir seems to have a higher overall positive sentiment than the other two HIV drugs. This trend cannot be explained by sentiment concerning side effects, since as a highly active antiretroviral theropy (HAART) Norvir has more sever side effects at typical doses than the other two drugs, and is administered to combat later stages of AIDs infection. It is possible that the increased positive sentiment for Norvir is related to its role as an affordable highly effective drug treatment option. Historically, Norvir has had one of the largest impacts to dramatically reduce deaths from HIV[11].

We have also done similar work examining the interactions between personality and sentiment towards pain related drugs, but we have omitted it here in order to keep this proposal concise.

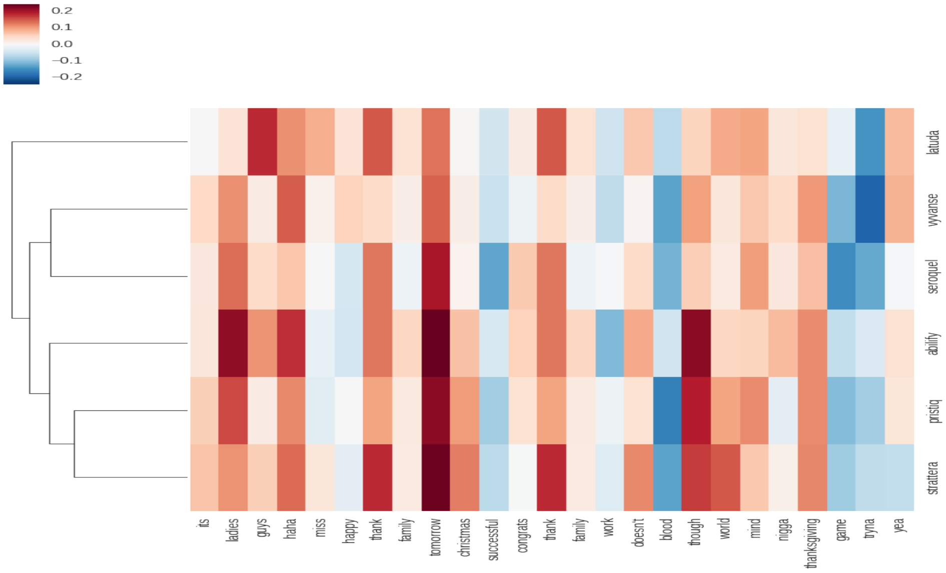

Figure 4: Relative similarity for psychiatric drug and personality keywords for positive sentiment tweets relative to negative sentiment tweets. The top 5 personality keywords associated with each of the following big 5 categories are shown in order: extraversion, agreeableness, conscientiousness, openness, neuroticism.

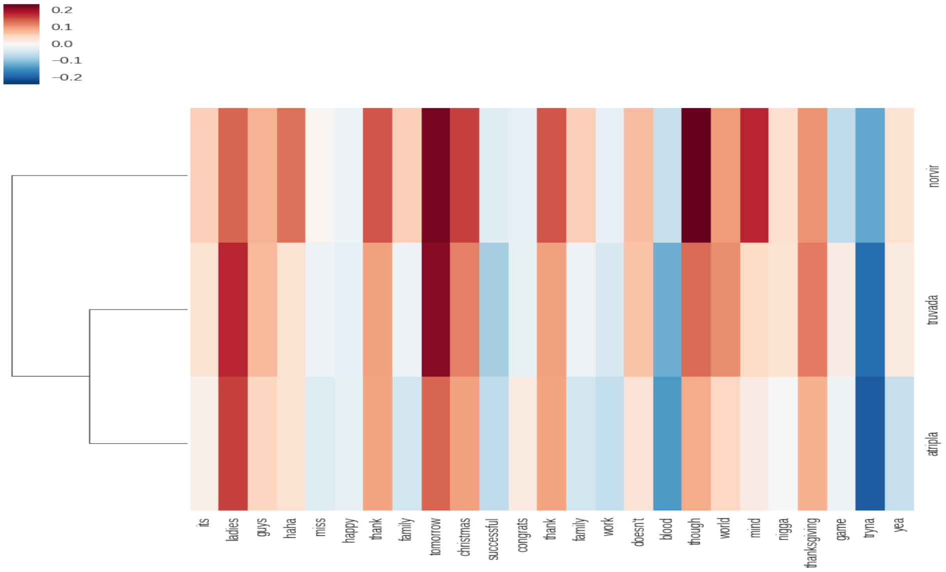

Figure 5: Relative similarity for HIV-related drug and personality keywords for positive sentiment tweets relative to negative sentiment tweets. The top 5 personality keywords associated with each of the following big 5 categories are shown in order: extraversion, agreeableness, conscientiousness, openness, neuroticism.

3.2 Proposed work

Further work will focus on two priorities, validating the high-level sentiment results, and identifying finer details explaining high level patterns. In order to address the issue of validation, we propose performing a similar analysis to the one described in our previous work on other social media data. We have access to a large dataset containing millions of Reddit comments and Reddit has been used in several medically related datamining analyses performed by other researchers[6]. Reddit represents a similar social media platform to Twitter, and while it lacks certain metadata, it has a much larger maximum character limitation. This may allow text mined from Reddit to contain more context than Twitter derived text, either way, if an analysis performed on Reddit data provides similar results, this gives us confidence in our Twitter results.

The incorporation of Reddit data for validation could fail, notably due to lack of mentions of the relevant pharmaceutical drugs. In exploratory research months ago, we found Reddit data to be relatively lacking in HIV related terms relative to Twitter. In this case, we could come up with a statistical way of validating the broad patterns shown in our heatmaps above. We could use a do statistical hypothesis testing. Though word2vec has no generally accepted statistical hypothesis testing associated with it, we could perform a sort of statistical hypothesis testing via simulation, to provide corresponding p-values for each value in the heatmap. We could also provide a p-value to the overall interaction of an average of the 5 psychology keywords in each category. Significant p-values would give us some measure of confidence in the high level interactions we are seeing.

The second priority that we have is delving deeper into the details underlying the psychology-drug sentiments we are seeing. Here we may be able to have a human read the small number of tweets that mention both a drug and a psychological keyword of interest. Alternatively we would use a sorting procedure based on doc2vec, as described in our previous work in section 1, to identify the most positive and most negative tweets associated with a given disease-psychological keyword of interest. Investigation of these specific tweets, and the metadata associated with them may be able to explain some of the factors underlying the high-level trends we uncovered in our preliminary results.

If successful, our results will show how psychology influences drug sentiment, which in turn will give us some information on how to improve drug perception. Improved drug perception will allow clinicians and public health professionals to better educate potential patients to the benefits and concerns associated with a drug, leading to better adherence, adoption, and eventually better health outcomes.

References Cited

[1] Indiana county with massive hiv outbreak missed advanced warning sign of new hepatitis c cases. `https://www.poz.com/article/indiana-county-massive-hiv-outbreak-missed-advanced-warning-sign-new-hepatitis-c-cases`. Accessed: 2017-01-17.

[2] Ume L Abbas, Gregory Hood, Arthur W Wetzel, and John W Mellors. Factors influencing the emergence and spread of hiv drug resistance arising from rollout of antiretroviral pre-exposure prophylaxis (prep). *PLoS One*, 6(4):e18165, 2011.

[3] Cecilie Schou Andreassen, Mark D Griffiths, Siri Renate Gjertsen, Elfrid Krossbakken, Siri Kvam, and Ståle Pallesen. The relationships between behavioral addictions and the five-factor model of personality. *Journal of Behavioral Addictions*, 2(2):90–99, 2013.

[4] Emily A Arnold, Patrick Hazelton, Tim Lane, Katerina A Christopoulos, Gabriel R Galindo, Wayne T Steward, and Stephen F Morin. A qualitative study of provider thoughts on implementing pre-exposure prophylaxis (prep) in clinical settings to prevent hiv infection. *PloS one*, 7(7):e40603, 2012.

[5] Sarah K Calabrese and Kristen Underhill. How stigma surrounding the use of hiv preexposure prophylaxis undermines prevention and pleasure: a call to destigmatize truvada whores. *American journal of public health*, 105(10):1960–1964, 2015.

[6] Annie T Chen, Shu-Hong Zhu, and Mike Conway. Combining text mining and data visualization techniques to understand consumer experiences of electronic cigarettes and hookah in online forums. *Online journal of public health informatics*, 7(1), 2015.

[7] Nicole L Cohen, Erin C Ross, R Michael Bagby, Peter Farvolden, and Sidney H Kennedy. The 5-factor model of personality and antidepressant medication compliance. *The Canadian Journal of Psychiatry*, 49(2):106–113, 2004.

[8] Caitlin Conrad, Heather M Bradley, Dita Broz, Swamy Buddha, Erika L Chapman, Romeo R Galang, Daniel Hillman, John Hon, Karen W Hoover, Monita R Patel, et al. Community outbreak of hiv infection linked to injection drug use of oxymorphoneindiana, 2015. *MMWR Morb Mortal Wkly Rep*, 64(16):443–444, 2015.

[9] Sarit A Golub, Kristi E Gamarel, H Jonathon Rendina, Anthony Surace, and Corina L Lelutiu-Weinberger. From efficacy to effectiveness: facilitators and barriers to prep acceptability and motivations for adherence among msm and transgender women in new york city. *AIDS patient care and STDs*, 27(4):248–254, 2013.

[10] Samuel D Gosling, Peter J Rentfrow, and William B Swann. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

[11] Robert S Hogg, Michael V O'Shaughnessy, Nada Gataric, Benita Yip, Kevin Craib, Martin T Schechter, and Julio SG Montaner. Decline in deaths from aids due to new antiretrovirals. *The Lancet*, 349(9061):1294, 1997.

[12] Robert H Howland. Potential adverse effects of discontinuing psychotropic drugs–part 2: Antidepressant drugs. *Journal of psychosocial nursing and mental health services*, 48(7):9–12, 2010.

[13] Molly E Ireland, H Andrew Schwartz, Qijia Chen, Lyle H Ungar, and Dolores Albarracín. Future-oriented tweets predict lower county-level hiv prevalence in the united states. *Health Psychology*, 34(S):1252, 2015.

[14] Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. Tracking suicide risk factors through twitter in the us. *Crisis*, 2014.

[15] Margaret A Lampe, Dawn K Smith, Gillian JE Anderson, Ashley E Edwards, and Steven R Nesheim. Achieving safe conception in hiv-discordant couples: the potential role of oral preexposure prophylaxis (prep) in the united states. *American Journal of Obstetrics and Gynecology*, 204(6):488–e1, 2011.

[16] Susan J Little, Sergei L Kosakovsky Pond, Christy M Anderson, Jason A Young, Joel O Wertheim, Sanjay R Mehta, Susanne May, and Davey M Smith. Using hiv networks to inform real time prevention interventions. *PloS one*, 9(6):e98443, 2014.

[17] Albert Liu, Stephanie Cohen, Stephen Follansbee, Deborah Cohan, Shannon Weber, Darpun Sachdev, and Susan Buchbinder. Early experiences implementing pre-exposure prophylaxis (prep) for hiv prevention in san francisco. *PLoS Med*, 11(3):e1001613, 2014.

[18] Nicholas A Livingston, Kathryn M Oost, Nicholas C Heck, and Bryan N Cochran. The role of personality in predicting drug and alcohol use among sexual minorities. *Psychology of Addictive Behaviors*, 29(2):414, 2015.

[19] Ethan Morgan, Alexandra M Oster, Stephanie Townsell, Donna Peace, Nanette Benbow, and John A Schneider. Hiv-1 infection and transmission networks of younger people in chicago, illinois, 2005-2011. *Public Health Reports*, 132(1):48–55, 2017.

[20] Hansen Andrew Schwartz, Johannes C Eichstaedt, Lukasz Dziurzynski, Margaret L Kern, Eduardo Blanco, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. Toward personality insights from language exploration in social media. In *AAAI Spring Symposium: Analyzing Microtext*, 2013.

[21] S Wade Taylor, Christina Psaros, David W Pantalone, Jake Tinsley, Steven A Elsesser, Kenneth H Mayer, and Steven A Safren. life-steps for prep adherence: Demonstration of a cbt-based intervention to increase adherence to preexposure prophylaxis (prep) medication among sexual-minority men at high risk for hiv acquisition. *Cognitive and Behavioral Practice*, 2016.

[22] Elisabeth Maria Van der Elst, Judie Mbogua, Don Operario, Gaudensia Mutua, Caroline Kuo, Peter Mugo, Jennifer Kanungi, Sagri Singh, Jessica Haberer, Frances Priddy, et al. High acceptability of hiv pre-exposure prophylaxis but challenges in adherence and use: qualitative insights from a phase i trial of intermittent and daily prep in at-risk populations in kenya. *AIDS and Behavior*, 17(6):2162–2172, 2013.

[23] Jonathan E Volk, Julia L Marcus, Tony Phengrasamy, and C Bradley Hare. Incident hepatitis c virus infections among users of hiv preexposure prophylaxis in a clinical practice setting. *Clinical Infectious Diseases*, 60(11):1728–1729, 2015.

[24] Norma C Ware, Monique A Wyatt, Jessica E Haberer, Jared M Baeten, Alexander Kintu, Christina Psaros, Steven Safren, Elioda Tumwesigye, Connie L Celum, and David R Bangsberg. What's love got to do with it? explaining adherence to oral antiretroviral pre-exposure prophylaxis (prep) for hiv serodiscordant couples. *Journal of acquired immune deficiency syndromes (1999)*, 59(5), 2012.

[25] Sean D Young and Devan Jaganath. Online social networking for hiv education and prevention: a mixed methods analysis. *Sexually transmitted diseases*, 40(2), 2013.

[26] Sean D Young, Caitlin Rivers, and Bryan Lewis. Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes. *Preventive medicine*, 63:112–115, 2014.

Biographical Sketch: Patrick Breen

(a) Professional Preparation

Bowdoin College BA Biochemistry and Mathematics, Brunswick Maine, 2013
University of Georgia PhD Bioinformatics, Athens Georgia, (attending)

(b) Appointments

Oak Ridge National Laboratory Fellowship (2016)
President of Bioinformatics Graduate Student Association (2015)
University of Georgia Presidential Graduate Fellowship (2013)

(c) Products

Mining Pre-Exposure Prophylaxis Trends in Social Media. Patrick Breen, Jane Kelly, Timothy Heckman, Shannon Quinn. DSAA2016.
P2Y6 receptor antagonist, MRS2578, inhibits neutrophil activation and aggregated NET formation induced by gout-associated monosodium urate crystals. Payel Sil ... Patrick Breen ...

(d) Synergistic Activities

Interacted with the local scientific community by judging at high school science fair (2014)

Data Management Plan

Multiple filtered streams of twitter data relevant to diseases studied will be acquired using Twitter's public streaming API. This data includes tweets related to Pre Exposure Prophylaxis, Truvada, HIV, and AIDS as well as other infectious diseases, and commonly used prescription medications. In addition to twitter data we will also acquire other social media data such as Reddit or Facebook, either from an open data repository, or in the case of Reddit through the public API. Direct medical data will be acquired from Oak Ridge National Laboratory and other sources and will be used in accordance with institutional and national laws and regulations (including HIPAA) governing appropriate use of medical data.

Data will be analyzed on local researcher's desktop machines and on research server clusters including the Quinn research group cluster, the Georgia Advanced Computing Resource Center, and Oak Ridge's institutional computing resources. Code developed for analyses will be made openly available on github.com under open source licenses.

Most of the primary information acquired through Twitter's API or direct medical information cannot be shared directly due to Twitter's terms of service and federal regulations such as HIPAA. However anonymized summary information can and will be disseminated in the form of research article publications, and perhaps also through blog articles and other non-technical mediums.

Collaborators and Other Affiliations Information

Collaborators and Co-Editors

Shannon Quinn (University of Georgia)
Jane Kelly (Georgia Department of Public Health)
Timothy Heckman (University of Georgia)
Arvind Ramanathan (Oak Ridge National Laboratory)

Graduate Advisors and Postdoctoral Sponsors

Shannon Quinn (University of Georgia)

Committee Members

Shannon Quinn (University of Georgia)
Jan Mrazek (University of Georgia)
Timothy Heckman (University of Georgia)
Keith Campbell (University of Georgia)