Project Summary

Overview

In this document the PIs describe an approach to identify trends in disease treatment by applying existing and novel machine learning techniques. Data sources will include both direct clinical sources and also pertinent text and metadata acquired from users of social media. High level information, analytical tools, and data resources produced through these analyses will be disseminated back into both clinical and social media venues to improve patient's direct medical outcomes and address social concerns. 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 is referd to as Pre-Exposure Prophylaxis (PrEP) and uses the small molecule nucleotide reverse transcriptase inhibitor trade named Truvada.

Intellectual Merit

The information produced through data mining and machine learning will be useful to provide clinicians and public health professionals feedback concerning successes and challenges associated with disease treatments. Pertinent tools and summary datasets will also be produced and made available to researchers. In addition novel applications of machine learning and data science methods will be developed as necessary, and shared with the data science community to assist and inspire related applications in other domain areas.

Broader Impacts

The proposed research will contribute to an increased scientific literacy through collaborations and dissemination of our findings through social media. This research will also lead to improvements of the health of all Americans, with in some cases such as PrEP, special emphasis on underrepresented groups such as the LGBT community. This work also has the potential to engage collaborations and partnerships between the social media industry, medical providers and pharmaceutical companies, and academia as we come together to improve public health.

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 show demonstrated efficacy at preventing HIV for HIV negative individuals in serodiscordant relationships[**?**].

Inital studies of PrEP have shown that it is highly effective[**?**], 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 understanding, social stigma, and uncertain insurance status have been identified as challenges preventing continued adoption[**?**]. Also, since Truvada is an oral NRTI, it must be taken daily. Cases when patients do not adhere to their full prescription have led to loss of viral protection, some patients and clinicians worry that lack of adherence could lead to increased risk of infection with drug resistant strains[**?**].

In order to identify factors that lead to the direct mitigation of HIV infection, we would like to be able to access direct medical data, however doing so can be difficult due to privacy regulations. Additionally, medical data often does not contain direct unfiltered opinions and feedback which capture patient and public perception. For this reason we propose the use of social media as one of our principle sources of data. 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. We have used many of these metadata attributes in previous analyses and will continue to do so in future analyses. Twitter also features a convenient well documented and fee API and has over 300 million monthly active users.

In past analyses 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 relevent to the PrEP conversation on Twitter including unexpected political conections "NancyReagan" and popular hashtags associated with PrEP that might not have been known to researchers in advance such as "#whereisprep". In addition Doc2Vec allows us 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 structurally important tweets that can be human-read without reading the full dataset. Reading these tweets uncovered important blog articles describing some of the fears of drug resistant forms of HIV that could result from misuse of PrEP medications. The linked blog articles that we found also 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. 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.

Using an open dataset of tweets which were labeled with binary sentiment labels, we performed a sentiment analysis. This 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 added into the loop to quickly read the top positive and negative tweets to get a sense of the successes and issues present in public perception. In the positive tweets we found hyperlinks to blogs with positive firsthand accounts from individuals successfully using PrEP to stay HIV negative. We also found concerns of whether Truvada can protect against drug resistant strains of HIV. 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 of the N most highly relevant tweets from a large corpus that can be easily human-readable.

- Twitter data acquisition, filtering
- Embedding analyses
- Document / Topic modeling results
- Sentiment Classification / Analysis

2 Arvind's secion (to be decided)

3 Broad application of biosurveillance and social media patient feedback

Broader Impacts

@articleliu2014early, title=Early experiences implementing pre-exposure prophylaxis (PrEP) for HIV prevention in San Francisco, author=Liu, Albert and Cohen, Stephanie and Follansbee, Stephen and Cohan, Deborah and Weber, Shannon and Sachdev, Darpun and Buchbinder, Susan, journal=PLoS Med, volume=11, number=3, pages=e1001613, year=2014, publisher=Public Library of Science

@articlecalabrese2015stigma, title=How stigma surrounding the use of HIV preexposure prophylaxis undermines prevention and pleasure: a call to destigmatize Truvada Whores, author=Calabrese, Sarah K and Underhill, Kristen, journal=American journal of public health, volume=105, number=10, pages=1960–1964, year=2015, publisher=American Public Health Association

@articlearnold2012qualitative, title=A qualitative study of provider thoughts on implementing pre-exposure prophylaxis (PrEP) in clinical settings to prevent HIV infection, author=Arnold, Emily A and Hazelton, Patrick and Lane, Tim and Christopoulos, Katerina A and Galindo, Gabriel R and Steward, Wayne T and Morin, Stephen F, journal=PloS one, volume=7, number=7, pages=e40603, year=2012, publisher=Public Library of Science

@articlegolub2013efficacy, title=From efficacy to effectiveness: facilitators and barriers to PrEP acceptability and motivations for adherence among MSM and transgender women in New York City, author=Golub, Sarit A and Gamarel, Kristi E and Rendina, H Jonathon and Surace, Anthony and Lelutiu-Weinberger, Corina L, journal=AIDS patient care and STDs, volume=27, number=4, pages=248–254, year=2013, publisher=Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA

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)