Research Strategy

6.1 Specific Aims

The primary goal of the proposed K01 is to provide real-time, open-access, population-level surveillance for HIV trends in order to support actionable evidence-based decision making. Despite a significant investment in big data analytics across public health, new strategies to track critical HIV trends, namely uptake of pre-exposure prophylaxis (PrEP), are still being developed. This is partially due to the fact that analyzing such data requires quantitative methods that are uncommon in public health, but are routinely utilized in data science. **My research proposes an accessible, cost-effective method of monitoring the uptake of PrEP in near real-time by mining internet query data**. PrEP use is critical to the prevention and control of HIV[38] as it reduces the risk of contracting HIV through sex by up to 99% and reduces the risk of contracting HIV through injection drug use by up to 74%[139]. However, there is a paucity of timely and accessible data on PrEP uptake. For instance, the most current data for PrEP uptake accessible through AIDSvu.org as of July, 2020 is from 2018[2]. Part of the reason why information on PrEP is not accessible in a timely manner is because of the proprietary nature of PrEP data, which relies on the willingness of pharmaceutical companies to share information that could negatively impact their business models (e.g., Gilead's Truvada and Descovy). Given the inability to freely access data on PrEP in real time, it is critical that we turn to public, open-source alternatives to monitor PrEP uptake that can identity where the gaps in uptake are the largest (Aim 1), develop techniques that utilize already released pharmaceutical data to develop validated proxy measures for uptake of PrEP (Aim 2), and evaluate PrEP uptake generated by HIV policy at the local and national level (Aim 3).

Aim 1: Discover Data-Driven Insights through Mining Internet Search Histories Related to PrEP: We will ascertain how searches for PrEP have changed in the United States by utilizing internet search histories. We will investigate internet search queries indicating interest in PrEP over time, along with variations in internet search queries across states and metropolitan area. In addition, will also investigate anomalous variations in PrEP internet searches. **Objective 1:** Investigate potential seasonal and temporal periodicities in internet searches for PrEP to identify recurring patterns in population demand. **Objective 2:** Detect anomalous spikes in PrEP searches to discover unknown drivers of demand. **Objective 3:** Describe the geographic variation in internet searches for PrEP at the state and metropolitan level.

Aim 2: Identify the Subset of Searches that Mirror Population PrEP Utilization to Use as a Real-Time Proxy In The Future: We will create a valid proxy for measuring uptake of PrEP by utilizing PrEP related internet search queries. This includes identifying a criterion to validate against by combining data on PrEP usage for selecting candidate proxy queries, and building validation models to produce real time estimates of PrEP uptake. **Objective 4:** Use principles of data science to select from a large list of queries the subset that most accurately forecast future population trends in PreP utilization nationally. **Objective 5:** Expand national forecasts to include more targeted geographic units, including states and metropolitan areas. **Objective 6:** Share these results so HIV researchers and decision makers have data on PrEP utilization years in advance of what is publicly available.

Aim 3: Evaluate Communication Campaigns, Policies, and Social Disparities that May Impact PrEP Utilization: We will use proxy models created during this study to quantify how changes in key HIV policy affect the uptake of PrEP. National and statewide changes to HIV policy will be evaluated for their ability to foster PrEP uptake. Policies will be compared to each other for their ability to affect change based on their content. **Objective 7:** Evaluate PrEP uptake generated by public policy changes. **Objective 8:** Evaluate health disparities in PrEP uptake. These aims leverage my background in epidemiology in order to investigate the uptake of PrEP in a significant and innovative way. The PrEP surveillance proposed in the study will enable researchers and policy makers to access affordable data-driven information in near real-time that will allow for better decision making.

6.2 Significance: While the use of big data analytics in public health has been increasing, HIV investigators have been slow to adopt surveillance strategies utilizing these techniques[40]. As a result, there is currently no publicly available system to monitor public interest in PrEP in near real time. Surveys and in-depth interviews continue to be the primary sources of knowledge for understanding PrEP related behaviors. These sources, however, are subject to well-known limitations, such as respondents' reluctance to participate, lag time between questionnaire design and implementation, data collection and data availability, and intermittent coverage of important topics[41]. Scaling up their use is not always feasible given both costs and concerns about overburdening the population with too many surveys[42]. My proposed research overcomes the shortfalls of traditional HIV surveillance by providing accessible and affordable information on PrEP trends in near-real time. The ability to improve on traditional survey-based surveillance with real-time data from internet search histories will allow for higher quality monitoring that more accurately informs efforts to promote PrEP[2243].

Search query monitoring is a logical starting point for impactful PrEP research because the data are public, easily accessible, and there is a significant body of existing research focused on applying internet search histories to PrEP research. **My research will build on the utility of search query monitoring to mine internet search histories for data-driven insights into public interest in PrEP (Aim 1).** Since 2010 multiple studies have confirmed the effectiveness of PrEP for preventing HIV infection[444523]. The number of PrEP users has increased dramatically since the FDA approved the use of Tenofovir/Emtricitabine (Truvada) for prophylaxis use against HIV infection[46], increasing by 880% from 2012 to 2016. This translates into over 77,000 PrEP users in 2016. However, the CDC has noted that the total number of people prescribed PrEP represent approximately 7% of the estimated 1.1 million persons who had indications for use[47]. It is highly likely that socioeconomic factors associated with PrEP use are not distributed equally given that the number of white men prescribed PrEP is six times higher than the number of black men. Only 2% of women with PrEP indications are prescribed PrEP despite women and black men accounting for approximately 40% of individuals with PrEP indications[48]. **My research on PrEP surveillance will mirror PrEP utilization in populations of socioeconomic interest at the metropolitan, state, and national levels by combining PrEP prescription data and internet search histories to improve HIV prevention years in advance of what is publicly available (Aim 2).** In turn, our models will be able to provide accurate and timely information on PrEP uptake that will allow health policy officials to make decisions that are more informed. This is particularly significant given that the U.S. Department of Health and Human Services (HHS) has made preventing HIV infections a cornerstone of the Ending the HIV Epidemic Initiative.[38] By allowing for the surveillance of PrEP trends in near-real time, it will be possible to better assess the level of PrEP uptake generated by health campaigns that target populations with higher indications for PrEP use. For example, our models will be able to quantify levels of interest in PrEP use generated when California passed Senate Bill 159 (SB159) which allowed pharmacists to initiate and provide PrEP without a prescription, by measuring uptake of PrEP before the enactment of the bill, and comparing it to PrEP uptake after the bill was passed. In addition to assessing the impact of past health policies on PrEP interest, our proposed research will also be able to monitor the impact of future changes to PrEP policies, such as widely anticipated results of the HPTN 083 & 084 trials on long lasting injectable PrEP[49], using similar methods. **My research will provide new tools for program managers to evaluate their PrEP promotion efforts by measuring the amount of public interest for PrEP generated during their health campaigns (Aim 3).** All three of my proposed aims will bring significant advances to the surveillance of PrEP trends. The aforementioned aims rely heavily on data science, analytics, and knowledge of infectious disease that I as an epidemiologist am uniquely positioned to leverage. As a result, I can better mine these data while contributing to the canon of PrEP information that will lead to actionable insights for promoting PrEP usage.

Innovation: Our study will demonstrate a novel approach to PrEP surveillance by capturing trends on PrEP use to improve HIV prevention and control in near-real time. Each aim proposes a unique application of data science to the surveillance of PrEP. Aim 1 investigates the pattern of internet searches for PrEP by building upon similar techniques that have been used to investigate infectious disease trends[22]. However, our aim is uniquely different from previous studies because we are investigating a proactive behavior aimed at the prevention of illness, rather than a reactionary behavior individuals engage in when they are ill[50]. Aim 2 of my proposal brings a unique perspective to the investigation of PrEP trends by combining traditional survey data with emerging data strategies. This innovative methodology will allow health policy makers to incorporate additional socioeconomic and demographic factors into their analysis. Lastly, Aim 3 will allow us to quantify public interest in PrEP generated by health communication campaigns and changes in PrEP policy. This would give policy makers and program managers a unique tool to help assess how much attention their efforts are generating.

6.4 Approach:

6.4.1 Data Acquisition: This proposal will make use of three sources of publicly available data. Each set of data will provide unique insights into evaluating and forecasting PrEP trends.

PrEP Prescriptions: Data on PrEP prescriptions used in this study originates from AIDSVu through a unique data sharing agreement with Symphony Health and Gilead[51]. PrEP prescription data represent electronic, patient-level prescription data from an overall sample that represents more than 54,000 pharmacies, 1,500 hospitals, 800 outpatient facilities, and 80,000 physician practices across the U.S. Prescriptions for TDF/FTC that were not made for PrEP related reasons (ex. HIV treatment, post-exposure prophylaxis, and chronic hepatitis B management) were removed through an algorithm validated by Gilead.[52] An individual was considered a PrEP user if their prescription was for a minimum duration of 30 days. Quarterly data is available beginning from 2012 at the metropolitan, state, and national levels.

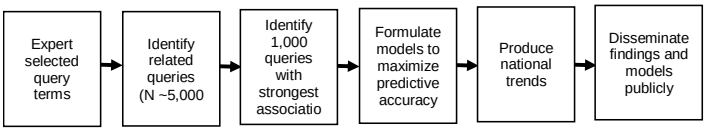
Medical Expenditure Panel Survey (MEPS): Widely considered one of the most complete sources of data on healthcare and health insurance coverage. Each year approximately 15,000 households, representing approximately 33,000 individuals are sampled for this survey. MEPS is nationally representative of the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers. Household Component data set consists of six files, which describe the demographics and characteristics of the survey population, and eight event-level files that capture all interactions with the U.S. medical system[53].

Google Trends API: We will use the private Google Trends Application Programming Interface (API) that is available to Dr. Ayers (co-mentor) to collect internet search query data from the Google platform. This automated programmable interface will allow us to perform thousands of searches at a time on demand. Information available through this API includes the volume of searches for each term, the number of searches per unit of time (day, week, month, years), and the geographic location of the searches (country, region, state, city, metropolitan area). Data on the volume of searches per day during the time period of the study will be downloaded. Trends in searches are calculated by measuring the fraction of searches that include the terms (or categories) in question in a user-chosen geography at a particular time relative to the total number of searches at that time. Search volume data for this proposal will be represented as a query fraction of the proportion of searches of a specific search term relative to all searches measured per 10 million searches. For example, if the search term "HIV test" had a query fraction of 30, it would represent 30 searches for the term "HIV test" per 10 million searches.

6.4.2 Data Preparation:

Determining Candidate Queries: The entirety of Google search queries starting from 2010, the date of the first successful PrEP trial CAPRISA 004[54], will be sampled to identify the subset of searches that produces the most valid PrEP trends. The first step in this process will be to select candidate search terms relevant to PrEP (*i.e.*, "pre-exposure prophylaxis"). This index will be compiled based on a literature review and consultation with my mentors who have proven to be PrEP experts (Dr. Little & Dr. Kates). Subsequent query terms beyond the initial search term will be identified from Google's search query archive (*i.e.*, PI and mentors select the first term "PrEP", the next most related term is "buy PrEP", whose next most related term is "by PrEP at CVS", etc.). Mentor Dr. John Ayers has made frequent use of this system to identify related search terms[5556]. The query fraction for each term will be downloaded through the time period of the study. The resulting pool of thousands of queries will be initially processed by correlating each trend with PrEP prescription data. The 1,000 queries exhibiting the strongest correlation retained for model building, replicating methods Dr. Ayers used for his Google Flu Trends revision[22] **(Figure 1).**

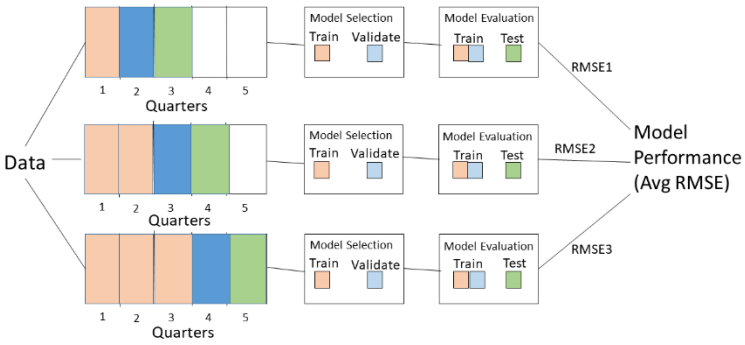
Figure 1: Candidate Query Selection:



Predictive/Forecasting Model Building**.** I will create models predicting PrEP uptake using data science techniques that I will learn as part of my training goals (such as boosted regression, support vector machines regression, and lasso regression)[5758]. I will particularly focus on developing Auto Regressive Integrated Moving Average (ARIMA) models using quarterly PrEP prescription data to predict PrEP uptake. ARIMA model construction will follow the Box-Jenkins procedure[59]. Autocorrelation plots will be used to determine the autocorrelation function and partial autocorrelation functions for the autoregressive (AR) and moving average (MA) model parameters. The stationarity of the data will be evaluated using augmented dickey fuller (ADF) and Kwiatkowski-Phillips-Shin (KPSS) tests. In the case of non-stationarity, differencing and data transformations will be applied when necessary. Seasonal periodicity of the data will be investigated by using a continuous wavelet transform to isolate the daily, weekly, monthly, and seasonal components of the time series and incorporated into seasonal ARIMA models if seasonality detected[60]. Log-likelihood, Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) will determine goodness of model fit. Forecasting predictions will be done in quarterly increments. While there are numerous difficulties when using traditional

model cross-validation with time series data, models will be trained using a method called “day forward-chaining” (sometimes referred to as rolling-origin evaluation[61]). Using this method, we successively consider each quarter as the test set and assign all previous data into the training set. As an example, if our dataset has five quarters, then we would produce three different training and test splits (**Figure 2**). This method produces many different train/test splits and the error (such as root mean squared error (RMSE), mean absolute error (MAE) and adjusted R^2) on each split is averaged in order to compute a robust estimate of the model error. Overall, our model fitting will allow us to (a) estimate how well internet search queries can predict PrEP uptake and (b) the queries that are most predictive of PrEP prescriptions. This last question is substantive in that it can also inform potential revision to the query terms being selected.

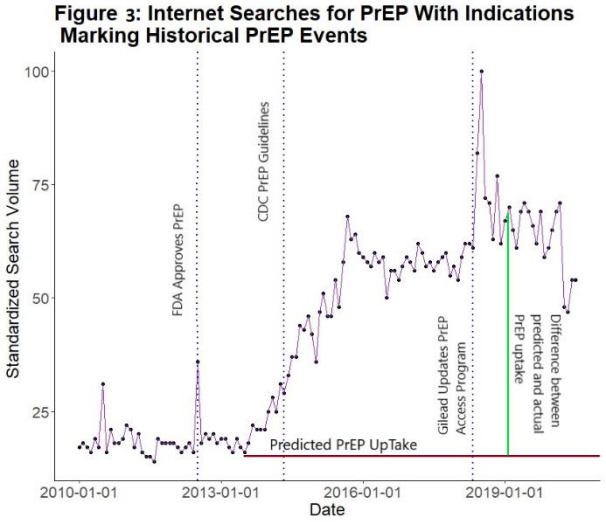
Figure 2: Day Forward-Chaining Cross-Validation



Aim 1: Discover Data-Driven Insights Through Mining Internet Search Histories Related to PrEP: This aim focuses on applying time series methodologies to the analysis of internet search histories data in order to produce population trends for PrEP usage. **Objective 1:** Investigate potential seasonal and temporal periodicities in internet searches for PrEP to identify recurring patterns in population demand. We will describe the overall trend in PrEP data nationally, statewide, and within metropolitan areas by fitting various regression models and through Mann-Kendall tests for trends. Using the techniques we developed in the model building section of our approach, we will investigate seasonality/ periodicity by using a continuous wavelet transform to isolate the weekly, monthly, and seasonal components of the time series data[60]. Wavelet transformation is preferred over other statistical methods because it is assumption free. The resulting series will then be intuitively compared as ratios of searches across time periods after accounting for mean centering. **Objective 2:** Detect anomalous spikes in searches for PrEP to discover unknown drivers of population demand. We will use various anomaly detection approaches, such as the Seasonal Hybrid Extreme Studentized Deviate (SHESD)[62] or the Inter-Quartile Range method, to detect outlying days, weeks, or months when PrEP searches are unusually high or low. Both methods will be conducted using the “AnomalyDetection” R package. **Objective 3:** Describe the geographic variation in internet searches for PrEP at the state and metropolitan level. Pearson correlations and Mantel tests comparing the correlation between national search trends for PrEP to search trends for each state and metropolitan are will be calculated. The correlation between PrEP search trends for each metropolitan areas will be used to investigate if the proximity of metropolitan areas affects the similarity of search trends.

Aim 2: Identify the Subset of Searches that Mirror Population PrEP Utilization to Use as a Real-Time Proxy In The Future: This aim will construct proxy models of PrEP uptake by utilizing internet search histories, PrEP pharmaceutical data, and MEPS data to forecast PrEP use. **Objective 4:** Use principles of data science to select from a large list of queries the subset that most accurately forecast future population trends in PrEP utilization nationally. Using techniques described in the model building section of our approach, quarterly prescription data for PrEP from 2010 to 2018, along with sociodemographic and insurance data from MEPS during the same time-period, will be used to create PrEP utilization models. **Objective 5:** Expand national forecasts to include more targeted geographic units, including states and metropolitan areas. Forecasting models described in our approach will be used to forecast PrEP uptake at the state and metropolitan level. Pearson correlations and Mantel tests comparing models predicting PrEP utilization for each state and metropolitan area will be calculated. The correlations between forecasted PrEP use between metropolitan areas will be calculated in order to investigate if the proximity between metropolitan areas affects the similarity of search trends. **Objective 6:** Share prediction models so HIV researchers and decision makers have data on PrEP utilization years in advance of what is publicly available. We will make our validated models publicly available through the creation of PrEPTrends.org in order to demonstrate the use of these data. We will replicate the simple website designs implemented in the past for Google Flu Trends[63], showing a time series for predicted PrEP uptake. The free to use webpage will be designed to be functional on desktop and mobile devices. Complete and unrestricted data availability will be maintained and users will be able to download the entire PrEPTrends.org database. As a result, data on PrEP utilization will become publicly available.

Aim 3: Evaluate Communication Campaigns, Policies, and Social Disparities that May Impact PrEP Utilization: This aim will quantify how changes in key HIV policy affect the uptake of PrEP. **Objective 7:** Evaluate PrEP uptake generated by Public Policy Changes. Forecasting models outlined in our proposal's approach will be used to quantify the effect that national and statewide PrEP policies have on PrEP uptake. Previous studies have used internet search queries to assess changes in public interest generated by government policies such as the Affordable Care Act during it's first enrollment period[64] and the effect air pollution regulations have on automobile purchases[65]. We will identify major national public policy changes regarding PrEP **(Figure 3)**, such as when the U.S Food and Drug Administration approved Truvada for PrEP use in 2012 and when the U.S Department of Health and Human Services announced the Ending the HIV Epidemic (EHE) Initiative in 2019[66]. Using the models we outline in our approach section, we will predict the expected volume of PrEP uptake that would have happened assuming that each particular PrEP policy did not occur. We will then compare this expected volume of PrEP uptake to the actual change in volume of PrEP uptake that resulted after the new PrEP policy was enacted **(Figure 3)**. The difference in the volume of PrEP uptake that we expect to happen without changes in PrEP policy will be compared the actual change in PrEP uptake at 1 month, 6 months, 1 year, and 5 years after each policy is enacted.



Objective 8: Evaluate Health Disparities in PrEP uptake. National, statewide, and metropolitan changes in PrEP uptake based on health insurance and socioeconomic indicators from MEPS data will be investigated. Factors affecting the accessibility and quality of care, such as the percentage of the population with a usual source of care, persons with difficulty accessing needed care, and patient-reported quality of doctor's visit will be explored for their association with PrEP uptake. The association between PrEP uptake and the distribution of race and ethnicity will also be investigated. Changes in state-level estimates of household medical utilization and expenditures will also explored for their affect on PrEP uptake. States that have participated in Medicaid expansion will be compared to states that have not participated in Medicaid expansion. The effect of a state's distribution of different forms of health insurance on PrEP uptake will also be investigated.

Data Privacy: The data used in this study have been de-identified are publicly available to researchers and health professionals. As a results, my analysis will not contain personal identifiers that could compromise an individual's privacy. I will follow all data privacy protocols outlined by the Agency for Healthcare Research and Quality, Data and Safety Monitoring Policy[67]. This research is in line with previous research deemed review exempt by the UCSD IRBs. However, I will submit a formal application to UCSD's IRBs to obtain a letter of exemption for this project.

Limitations: It is possible that an individual could search for information on PrEP after filling a prescription. It is also possible that individuals who are curious about PrEP but have no intention of filling a prescription would search for it. While spurious internet searches for PrEP are a possibility, it is generally accepted that internet search activity has potential predictive power for activities that an individual would want to research before they take action[16]. It is possible that the results of our research could be subject to ecological fallacy. The PrEP prescription data used in this study are aggregated at the national, state, and metropolitan levels. This leaves our research open to bias in interpreting individual behavior based on the group to which those individuals belong. It is possible that individuals who are the most likely to need access to PrEP are unable to access the internet. Roughly 10% of Americans did not have access to the internet in 2019[68]. Internet access is strongly associated with SES factors, where 15% of African Americans and 27% of individuals 65 and older do not have access to the internet. While SES does affect who can access the internet, we believe that this will not greatly affect our results given that the majority of PrEP users are younger than 49 years[69] and only 3% of individuals 49 years old or younger did not have access to the internet[68]. Lastly, PrEP prescription data does not include information from closed healthcare systems that do not make their data available to AIDSvu.

7. Future Directions: I believe the training I receive through this mentored development award will provide me with the data science skills in time series analysis necessary to expand my research into other areas of HIV treatment and prevention. In Year 4 I will submit a UCSD CFAR developmental grant that will expand on my proposed research by investigating the utility of using internet search histories to predict HIV risk behaviors. In

Year 5 I will submit and R01 that will seek to develop methodologies that validate predictive models explaining HIV risk behaviors.