TOWARDS ENDING THE HIV EPIDEMIC ONE COUNTY AT A TIME AN EVALUATION FRAMEWORK FOR SUBCOUNTY-LEVEL USAGE OF PRE-EXPOSURE PROPHYLAXIS FOR EXPOSING DISPARITIES AND GUIDING TARGETED INTERVENTIONS

by

Sagar Satyanarayana

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Infrastructure and Environmental Systems

Charlotte

2021

Approved by:

_____
Dr. Gabriel Zenarosa

_____
Dr. Daniel Janies

_____
Dr. Patrick Robinson

_____
Dr. Guanglin Xu

_____
Dr. Michael Dulin

ABSTRACT

SAGAR SATYANARAYANA. TOWARDS ENDING THE HIV EPIDEMIC ONE COUNTY AT A TIME
An Evaluation Framework for Subcounty-level Usage of Pre-exposure Prophylaxis for Exposing Disparities and Guiding Targeted Interventions. (Under the direction of DR. GABRIEL ZENAROSA)

About 1.1 million Americans were living with HIV in 2019 with total lifetime cost to treat a single person with an HIV infection estimated to be around $501,000. Huge disparities in new HIV incidences exists between different geographic and demographic groups, with Southern US accounting for about 52% and men accounting for 79.2% of all new cases in the country. In February 2019, "Ending the HIV Epidemic: A Plan for the United States" was proposed, with pre-exposure prophylaxis (PrEP) being a major component of the prevention strategy. PrEP, a pill taken daily by mouth, contains antiretroviral drugs and is highly effective in preventing acquisition of HIV. Despite the fact that PrEP coverage in the US improved from 9% in 2016 to 18% in 2018, huge disparities in PrEP prescriptions exists in different geographic, racial/ethnic and age groups. Existing metrics to measure PrEP coverage like the PrEP-to-need ratio, defined as the ratio of number of patients with at least one day of PrEP prescription in a year to new HIV cases overestimates PrEP coverage. Moreover, previous studies conducted at the national and state levels often fail to capture disparities in PrEP use within the county and cannot be used by county public health officials to conduct targeted interventions.

In this dissertation I develop an evaluation framework for HIV prevention using novel metric bounds encompassing PrEP patient, and pills count for measuring PrEP usage at subcounty level, as well as an evaluation framework to quantify the effects of Public Health Interventions (PHIs) on PrEP usage. Pharmacy claims data for PrEP along with HIV incidences and census data for Mecklenburg County from 2013–2019 will be used in the analysis. The following specific aims are followed to accomplish this

objective: **Aim 1:** Aggregate ZIP codes to avoid potential patient re-identification, **Aim 2:** Establish a novel bounds for the likely PrEP-to-need ratio in different geographic and demographic groups, and **Aim 3:** Evaluate the influence of G2Z-MC on monthly PrEP users.

The 29 ZIP codes in Mecklenburg County were aggregated into 13 geographically adjacent ZIP code groups to avoid the risk of potential patient re-identification. From 2013 to 2019, there were 2,045 PrEP patients and 466,525 PrEP pills dispensed in Mecklenburg county. The population-adjusted PrEP patients per 100,000 increased from 5.64 to 106.39 and pills dispensed increased from 3,609 to 187,050. The overall [dose adjusted PrEP-to-need ratio (daPnR), PrEP-to-need ratio (PnR)] range increased from [0.0578, 0.3275] to [2.1176, 4.9628]. The [daPnR, PnR] ratio range increased from [0.053, 0.5] to [0.717, 2.09] for females and from [0.059, 0.285] to [2.429, 5.601] for male. Patients aged $\geq$ 45 years had a notable increase in [daPnR, PnR] ratio range from [0.079, 0.458] to [3.56, 6.649]. The ZIP code group containing main campuses of the two largest Mecklenburg County hospitals and a specialty pharmacy had a notable increase in [daPnR, PnR] ratio range from [0.275, 2.5] to [10.236, 22.67]. The G2Z-MC intervention had a significant gradual effect of about nine PrEP patients every month.

PnR and daPnR ratios increased from 2013 to 2019. However, female patients, aged $\leq$ 24 years, or belonging to certain ZIP code groups are underserved. Our results indicate the need for focused efforts to make PrEP more accessible for underserved populations. The results of this dissertation quantified and identified opportunities for improvement in PrEP use within Mecklenburg County in different geographic and demographic groups. This study also establishes a framework to evaluate the effects of Public Health Interventions (PHIs) on PrEP usage in Mecklenburg County. This work can be extended to other counties and will also provide a foundation to conduct similar studies for other emerging infectious diseases.

DEDICATION

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, Shashikala and Tadagada Lakshmipathaiah Satyanarayana whose words of encouragement and push for tenacity ring in my ears.

ACKNOWLEDGEMENTS

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Dr. Gabriel Zenarosa, whose expertise was invaluable in formulating the research questions and methodology. Your unwavering support and guidance throughout my work and my time here at UNC Charlotte pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to thank my committee members, Dr. Patrick Robinson, Dr. Daniel Janies, Dr. Guanglin Xu, and Dr. Michael Dulin for their valuable guidance throughout. You provided me with the tools that I needed to choose the right direction and successfully complete my dissertation.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support of my friends, who provided stimulating discussions and much needed distractions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

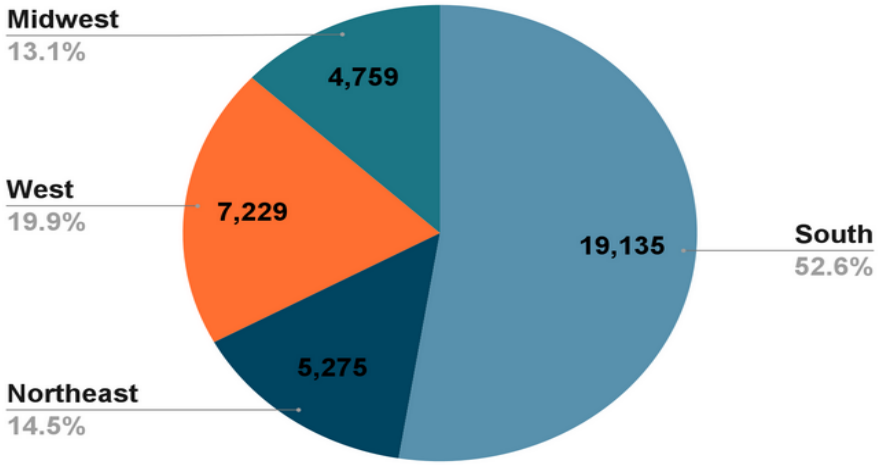
CHAPTER 1: INTRODUCTION

At the end of 2019, about 1.1 million Americans live with HIV [1, 2]. Since 2013, the number of new HIV infections each year in the United States (US) has remained stagnant at around 38,000 [1, 3], with an estimated lifetime cost of $501,000 to treat a single patient with HIV [4, 5]. In addition, huge disparities in HIV incidences and prevalence exist in different geographic and demographic subgroups, with the Southern US accounting for 52.6% of new infections and about 79.2% of new infections in men in 2019 [1]. Figure 1.1 and Figure 1.2 show the HIV incidences and prevalence in the US by region, sex, and age groups respectively.

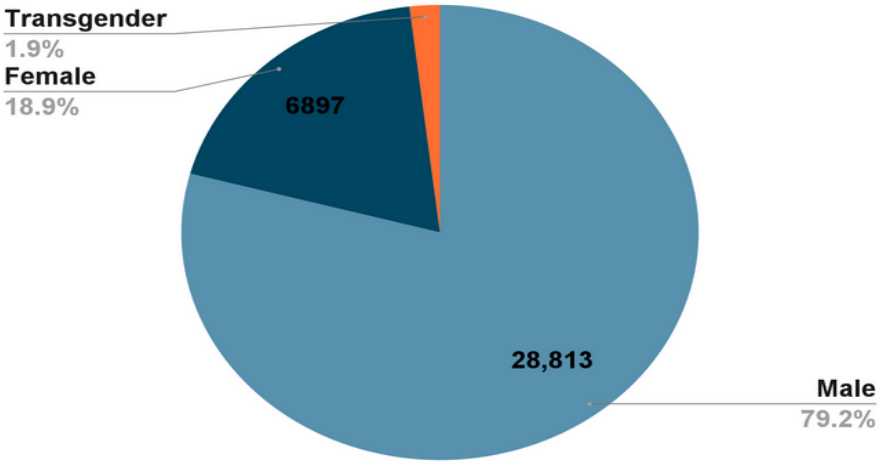
In 2019, a new federal initiative, "Ending the HIV Epidemic: A Plan for America" was proposed with an overall objective to reduce new HIV infections by 90% by 2030 [2, 6]. The plan calls for (i) diagnosing 95% of persons living with HIV and 95% of those diagnosed to have suppressed viral loads; and, (ii) 50% of all persons at risk of contracting HIV to be being prescribed PrEP. The plan calls for intensified efforts to diagnose, treat, and prevent HIV infections in the US, with pre-exposure prophylaxis (PrEP) being a major component of the prevention strategy [3].

PrEP, a pill taken daily by mouth, contains antiretroviral drugs and is highly effective in preventing the acquisition of HIV [7, 8, 9, 10, 11]. The Food and Drug Administration (FDA) currently approves of two drugs to be used for PrEP [7]. The combination of tenofovir disoproxil fumarate (TDF) and emtricitabine (FTC) was approved for PrEP by the FDA in 2012, commonly known by the brand name Truvada® [12, 13]. In 2019, the FDA approved a second drug for PrEP known by the brand name Descovy®, a combination of tenofovir alafenamide (TAF) with FTC [14, 15].

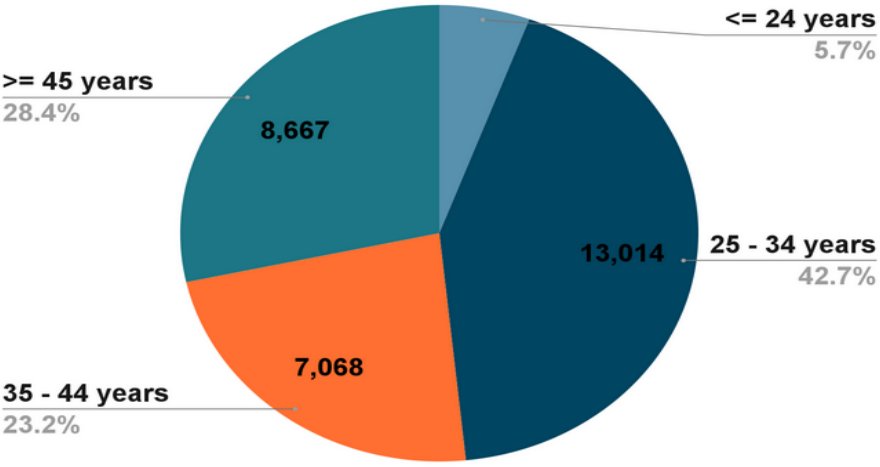
The focus of the "Ending the HIV Epidemic: A Plan for America" on using PrEP as



HIV incidences by region

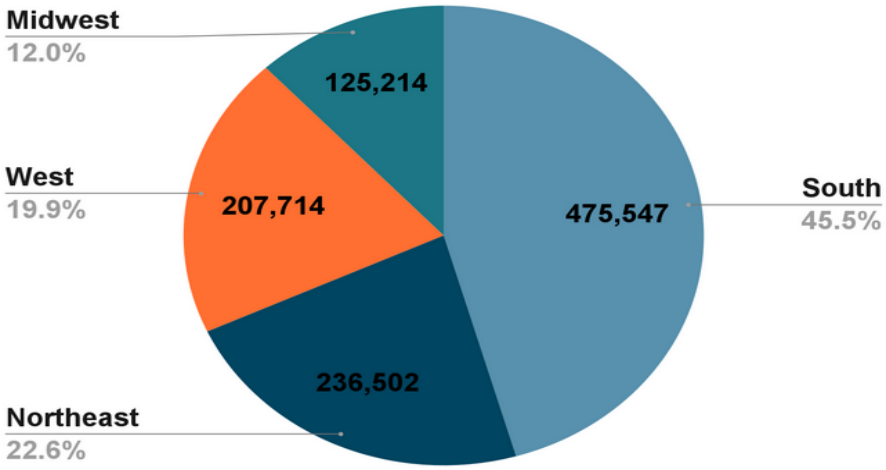


HIV incidences by sex

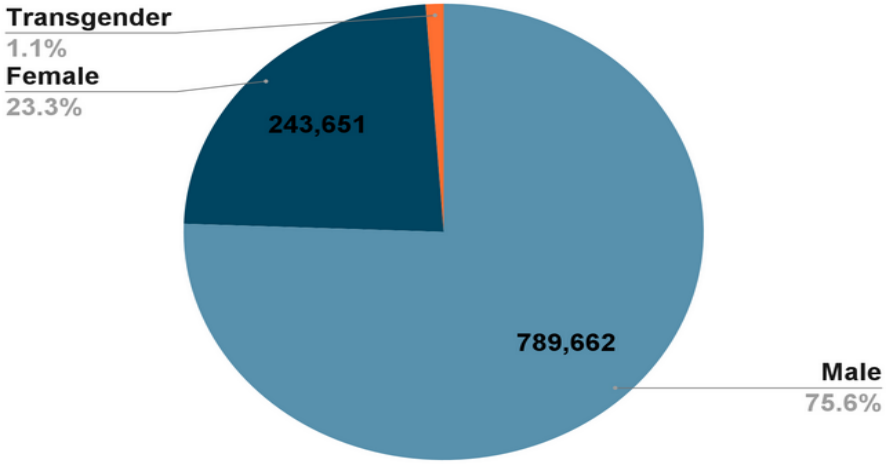


HIV incidences by age group

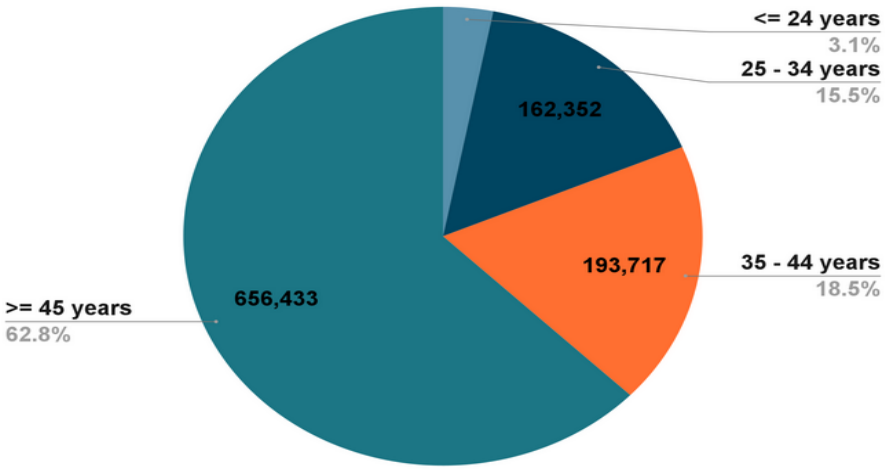
Figure 1.1: HIV incidences in 2019 by region, sex, and age group in the US



HIV prevalence by region



HIV prevalence by sex



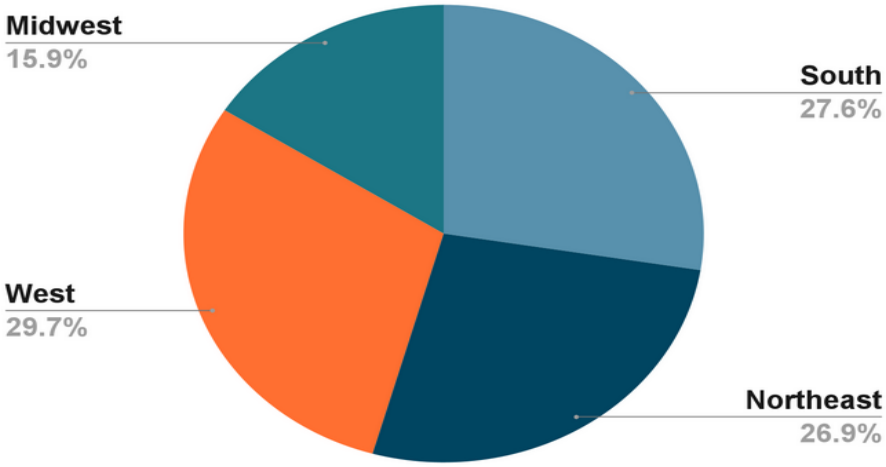
HIV prevalence by age group

Figure 1.2: HIV prevalence in 2019 by region, sex, and age group in the US

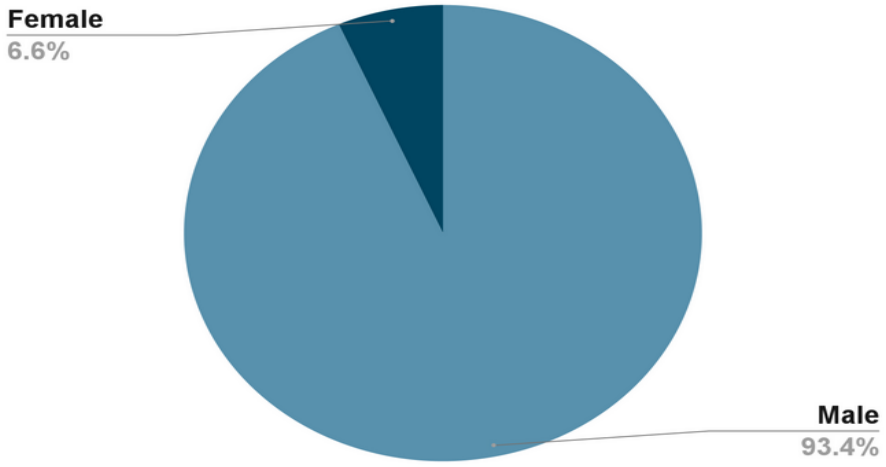
a preventive strategy is logical, given its potential. Around 37.6% of new infections are transmitted by patients who are not yet diagnosed with HIV, 42.6% from diagnosed HIV patients not receiving medical care, and 19.8% from HIV patients receiving medical care but are not yet virally suppressed to prevent transmission [16, 17]. A 2016 study indicated that the lifetime risk for an American to be diagnosed with HIV was one in 106, and the lifetime risk of HIV diagnosis in North Carolina was one in 100 [18]. According to CDC's criteria for PrEP based on risk score indices [19], about 1.2 million people in the US had indications for PrEP [20]. The "Ending the HIV Epidemic: A Plan for America" calls for 50% of people with indications for PrEP to be prescribed PrEP by 2030. Only 18.1% of people with indications for PrEP were using PrEP in the US in 2018 [3]. This makes it imperative to increase the use of PrEP as a preventive strategy, as daily adherence to PrEP has proven to be highly effective in preventing new HIV infections [8, 9, 11, 18, 21, 22, 23, 24, 25].

PrEP coverage, calculated as the number of persons prescribed PrEP divided by the estimated number of persons who had indications for PrEP, improved from 9% in 2016 to 18% in 2018. However, disparities in PrEP prescriptions exist in different geographic, racial/ethnic, and age groups [3, 7]. The Figure 1.3 shows the disparities in PrEP use, with only about 27.6% of PrEP being used in the southern US, which has nearly half of all new incidences in the country. A vast disparity between sexes also exists, with about 93.4% of men using PrEP compared to only 6.6% women using PrEP. We see age group 25–34 years having the highest PrEP users with about 39.8%, closely followed by age groups 35–44 years, and age group $\geq$45 years at 24.5% and 24.4% with age group $\leq$24 years having the least number of PrEP users at 11.3%. Hence, monitoring trends in PrEP use corresponding to the epidemiological need (HIV incidence) at different geographic and demographic groups can guide interventions to ensure PrEP is provided for those who need it most.

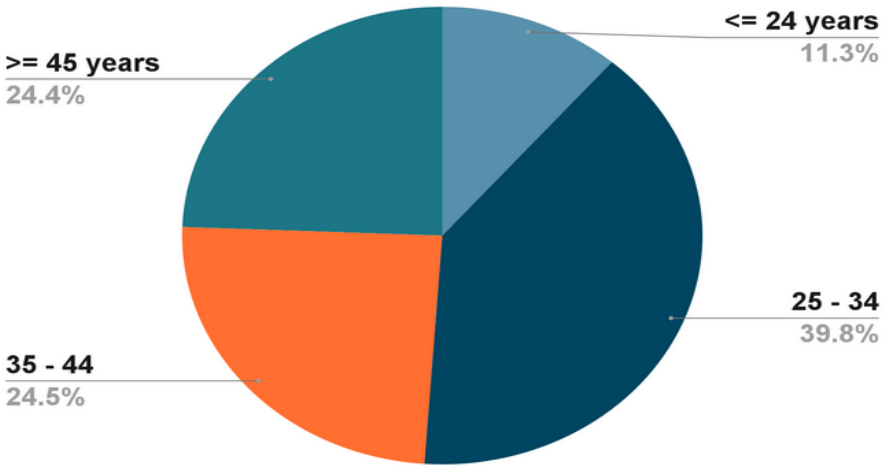
The PrEP-to-need ratio (PnR), introduced by Siegler et al., is a metric to measure



PrEP coverage by region



PrEP coverage by sex



PrEP coverage by age group

Figure 1.3: PrEP coverage in 2018 by region, sex, and age group in the US

PrEP usage and is used to elucidate disparities in PrEP provision across geographic areas and demographic groups [26, 27]. PnR is defined as the number of PrEP users divided by the number of new HIV diagnoses and compares the relative level of PrEP provision to the epidemiological need (HIV incidence). The number of PrEP users used to compute PnR is defined as the number of persons with at least one day of PrEP prescription in the year. The PnR can sometimes overestimate the PrEP coverage as PrEP needs to be taken throughout a year by a single patient for the patient to be protected in that year. However, due to changes in risk, inconvenience, aversion to taking a preventive drug, and limited access to PrEP, PrEP is usually taken by people only when they are at risk (e.g., sexually active). This poses a challenge in determining the exact PnR to protect an individual entirely according to their risk of HIV acquisition.

Various studies are conducted to monitor PrEP usage at the state, jurisdiction, and county levels but none at the subcounty level [26, 27, 28, 29, 30]. Subcounty studies can expose disparities at an actionable level by the county's Department of Public Health. This study proposes a method for monitoring PrEP usage at subcounty (ZIP code) levels in Mecklenburg County in North Carolina, which includes the City of Charlotte.

The Charlotte metropolitan area is ranked 28^{th} among 107 metropolitan areas in the nation for new HIV infections in 2018 [1]. Mecklenburg County is experiencing a persistent epidemic of new HIV infections, with increasing rates in some sub-populations and geographic regions within the county. In 2019, there were 6,665 people diagnosed with HIV living in Mecklenburg County with 270 new cases in the year [31]. A comprehensive countywide HIV plan, "Getting to Zero – Mecklenburg County" (G2Z-MC), was initiated in June 2018 to reduce the rate of new infections in Mecklenburg County. One key strategy in the G2Z-MC plan is to increase the use of PrEP among persons at increased risk of acquiring HIV infection. The plan includes

staged approaches intended to increase the use of PrEP in Mecklenburg County over the next two years, including the establishment of a pilot program to provide PrEP to uninsured patients. The G2Z-MC plan is expected to increase PrEP use in Mecklenburg County [32]. However, an effective and sensitive method of monitoring changes in PrEP uptake is needed to substantiate and quantify the plan's impact.

This dissertation aims to develop an evaluation framework for HIV prevention using a novel metric (bounds) encompassing PrEP patient counts and days adjusted pill counts for measuring PrEP usage at the subcounty level. This work is responsive to G2Z-MC and can monitor and assess its influence on HIV prevention using PrEP. We propose the following specific aims to accomplish these objectives.

- **Aim 1: Aggregate ZIP codes to avoid potential patient re-identification**

  We develop an optimization framework to aggregate ZIP codes containing potentially re-identifiable patient information in their corresponding demographic groups in the pharmacy claims dataset. ZIP codes will be aggregated into ZIP code groups with an optimally maximum number of groupings to avoid information loss while preventing patient re-identification. This step is required for the pharmaceutical data provider to release data used in subsequent analyses. We will restrict groupings to contain contiguous ZIP codes to afford geographically targeted intervention planning.

- **Aim 2: Establish novel bounds for the likely PrEP-to-need ratio**

  Quantify PrEP coverage within the county from filled prescriptions of TDF+FTC and TAF+FTC by deriving a method to identify PrEP regimens from prescription histories of patients. The identified PrEP regimens will be used to calculate the bounds for the likely PnR in a year in different geographic and demographic groups. Pharmacy claims data for TDF+FTC and TAF+FTC prescriptions within the county are obtained from PRA Health Sciences. The need is deter-

mined by the number of new cases of HIV in a year and is obtained from the North Carolina Division of Public Health.

- **Aim 3: Evaluate the influence of G2Z-MC using PnR bounds**

  Segmented linear regression analysis will be conducted on the number of monthly PrEP patients to quantify the immediate and gradual effects of G2Z-MC intervention on the number of PrEP users. This analysis is used to quantify the effectiveness of the G2Z-MC to provide vital information regarding the intervention's effectiveness for public health decision-makers in the county.

CHAPTER 2: DATA SOURCES

2.1 Pharmacy claims data

We obtain pharmacy claims data for TDF+FTC and TAF+FTC pills dispensed from 2012 and 2019 in Mecklenburg County pharmacies from the Pharmaceutical Research Associates Health Sciences (PRA). PRA specializes in collecting medical information nationally and has access to about 92% coverage of pharmacy claims for TDF+FTC and TAF+FTC in the US and its territories [33]. The dataset obtained also contains patient demographic information, including sex, age, insurance payer type, and the ZIP code of the dispensing pharmacy but not of the patient's residence. Patient residence ZIP codes are available in other medical datasets, however, the coverage for those is only around 60% [33].

Before the data release, PRA provided re-identifiable fields that needed merging. We use an aggregation algorithm to optimally merge the ZIP codes containing any fields with re-identifiable information with adjacent ZIP codes to form contingent ZIP code groups, as detailed in Chapter 3.

After de-identifying ZIP codes containing potentially re-identifiable information, we received the dataset, which encompassed patients' scripts for TDF+FTC and TAF+FTC, either with or without other antiretrovirals dispensed in Mecklenburg County from 2012 to 2019. The dataset contains eleven fields which are described below in Table 2.1. This dataset will then be used to report trends in TDF+FTC and TAF+FTC used for PrEP in different geographic and demographic groups across Mecklenburg County from 2013 to 2019 in Chapter 4.

Table 2.1: Columns in patient scripts file containing pharmacy claims for TDF+FTC and TAF+FTC obtained from Pharmaceutical Research Associates Health Sciences (PRA)

Column name	Description
patient_id	unique identification code for each patient
drug_id	unique identification number for each drug type
orig_drug_name	name of the drug, includes individual names for the ARVs
drug_name	name of the drug (either as TDF+FTC, TAF+FTC or ARV)
orig_rx_fill_date	original fill date
days_supply	number of days of the prescription filled (Pills count)
orig_rx_ds_end_date	original end date
rx_fill_date	adjusted fill date
rx_ds_end_date	adjusted end date
age_group	age group of patients broken into: ($\leq$24 years, 25–34 years, 35–44 years, and $\geq$45 years)
pharmacy_zip_rollup_group	ZIP code of the pharmacy from where the prescriptions werre filled
patient_gender	sex of patients as either male or female
rx_payer_group	insurance payer type of the patients, broken into: ('Commercial / Self Pay', 'Medicare / Medicaid', 'PAP', 'Unknown')

2.2 HIV incidence data

We obtain the number of HIV incidences per year in Mecklenburg County in 2013–2019 from the Enhanced HIV/AIDS Reporting System (eHARS). eHARS is a browser-based, CDC-developed application to assist health departments with reporting, data management, analysis, and data transfer to CDC [34]. HIV incidences data for Mecklenburg County is derived and provided by the North Carolina Division of Public Health, Epidemiology branch. The dataset contains HIV incidences by different geographic location (ZIP code) and demographics groups (sex and age groups). This data will be used as the epidemiological need to determine trends in PrEP coverage corresponding to its need across different geographic and demographic groups across the county, as detailed in Chapter 4.

2.3 Census data for Mecklenburg County

We obtain census data for the population in Mecklenburg County in each ZIP code by sex and age group from the Applied Geographic Solutions (AGS). This dataset includes the census population in 2010 and the estimated population in 2019. We web-scraped 2000 census population data from an open-source website (`https://www.ZIP-codes.com/county/nc-mecklenburg.asp`). We use these three data points to interpolate the annual population from 2013 to 2019 in different ZIP code groups, sexes, and age groups. Using three data points to interpolate instead of two will better capture the trend in the data because it can account for the curvature, which is impossible with two data points. We interpolate the yearly population in different ZIP codes by sexes, and age groups using cubic splines. The annual interpolated population in 2013–2019 is used as the denominator to calculate the population-adjusted rate of PrEP in different geographic regions and demographic groups within the county, as detailed in Chapter 4.

We also web-scraped latitude and longitude coordinates of the ZIP code centers from the same above website. These latitude and longitude coordinates will assist in computing the distances between the ZIP codes in Chapter 3 to aggregate ZIP codes into de-identified ZIP code groups.

CHAPTER 3: AGGREGATION OF STUDY POPULATION WITH SIZE AND GEOGRAPHIC CONSTRAINTS

3.1 Background

The digitalization of health data has generated large volumes of medical information. The benefits of sharing this information are vast, ranging from being able to easily communicate and confirm published research results, facilitating additional novel analysis on the same datasets, making the data available for instruction and education, and reducing duplicate data-collection efforts [35, 36, 37, 38, 39, 40]. Due to these benefits, dissemination, and sharing of patient-specific medical information is gaining popularity rapidly; microscopically detailed information is becoming the new reporting norm in place of aggregate statistics. Additionally, the call to make data gathered using public funds accessible has increased the need for sharing medical information [36, 41]. While dissemination and sharing of medical information can support healthcare research, this accessibility also poses new challenges in protecting privacy.

The Health Insurance Portability and Accountability Act (HIPAA) was established to protect privacy in health information held or transmitted nationally. HIPAA considers several elements as personal identifiers and terms such data as Protected Health Information (PHI). HIPAA prohibits PHI-sharing without consent or having the data de-identified either by expert determination or safe harbor methods [42]. Although there is no universal method or a threshold of re-identification risk of de-identified records, several guidelines and methods are proposed and practiced [42]. Studies have shown that, even when datasets do not contain personal identifiers like names or phone numbers, quasi-identifiers can be used for re-identification. Quasi-identifiers

are defined by Dalenius et al. [43] as any combination of characteristics on which linking can be enforced with other publicly available sources of data to re-identify patients, including ZIP code, age, and sex. To safeguard privacy in such datasets, it is common practice to use k-anonymization [44].

A k-anonymized dataset has a property that each record is similar to at least $k-1$ records on the quasi-identifiers—that is, each quasi-identifier must have at least k counts (chosen by the data holder) of individuals in them [44]. k-anonymization significantly reduces the re-identification risk of individuals by $1/k$ [45]. HIPAA does not consider a k-anonymized dataset as PHI and, in turn, is not subject to HIPAA privacy rules [42]. Exploratory data analysis that divides the study population into smaller subgroups to derive deeper insights is increasingly becoming popular. These small subgroups make it difficult to k-anonymize datasets with low counts; some quasi-identifiers may not have the required number of records in them for k-anonymization.

Traditionally, such problems are overcome by concealing quasi-identifier values (*suppression*), replacing those values with a more general one (*generalization*), or a combination of both until all quasi-identifiers have the required number of records [44]. Although both the methods achieve k-anonymity, they may lead to the loss of information that could otherwise be used [46]. Several other *perturbation* methods like *condensation*,[47] *data scrambling and swapping*,[48] or *adding noise*[49] are sometimes used to achieve k-anonymity. While these methods maintain the overall statistical property, they may distort data leading to less accurate results [46].

To provide geographically and demographically actionable information (i.e., for targeted interventions), we required prescription records of TDF+FTC and TAF+FTC with minimally anonymized ZIP codes and patient demographics. While the required data are free of any personal identifiers, they contain quasi-identifiers. Small numbers of PrEP prescriptions in some ZIP codes necessitated k-anonymization to protect patient privacy.

Our data provider, Pharmaceutical Research Associates (PRA) [50] Health Sciences requires requires 5-anonymized datasets for release [33]. Because demographic quasi-identifiers are established in other datasets with which our prescription data can be linked (e.g., census data for population-adjusted prescription rates), we addressed the need for k-anonymization by minimally aggregating ZIP codes. Anonymity approaches like suppression and perturbation are not useful in our case because suppressing or perturbing the ZIP codes loses geographic information. Moreover, naïvely generalizing the ZIP codes using their initial three or four digits generates large or non-contiguous geographic areas, which prevents geographically targeted interventions for G2Z. Thus, a custom generalization approach is required.

This chapter details the algorithm for k-anonymizing the data while finding the largest numbers of groupings of contiguous ZIP codes using an optimization model. Our optimization model aggregates ZIP codes with potentially re-identifiable low counts (i.e., less than five patients) to achieve the required k-anonymization (i.e., $k = 5$). The optimal solution obtained from this model k-anonymizes the prescription dataset with minimum aggregations leading to a maximum number of groups. To evaluate the performance of our optimization model, we compare its resulting ZIP code groupings to a heuristic. We note that while we demonstrate our optimization approach on prescription data for TDF+FTC and TAF+FTC use for PrEP in Mecklenburg County, it can be applied to other k-anonymization settings.

The chapter is organized as follows. First, we summarize the prescription data requiring k-anonymization, with which we formulate a model to optimally find the largest number of ZIP code groups achieving the required k-anonymization. We then present a heuristic for finding ZIP code groups with which to compare to our model. We show that our model outperforms the heuristic. Finally, we discuss possible extensions to our model for other analyses.

3.2 Methods

Our goal is to 5-anonymize PRA Health Sciences prescription data of TDF/FTC and TAF/FTC by minimally aggregating contiguous ZIP codes along with the following patient demographic quasi-identifiers:

- Sex: female and male;
- Age group: $\leq$24 years, 25–34 years, 35–44 years, and $\geq$45 years;
- Insurance type: Patient Assistance Program (PAP), Medicare & Medicaid, commercial & self-pay, and unknown.

Mecklenburg County, at the time of writing, has 83 ZIP codes (including post office boxes and corporate mail services) which are contained within 29 ZIP codes of polygonal regions. Table 3.1 summarizes the data and shows the need for 5-anonymization (i.e., 59 of 290 entries possibly are re-identifiable).

Since our evaluation study deals with reporting the number of patients within each ZIP code in the county, we had to ensure all 29 ZIP codes either had zero (i.e., no re-identifiable patients) or at least five patients across all 10 quasi-identifier categories—that is, in each ZIP code, there must either be zero or at least five patients in both sexes, all age groups, and all insurance types. Out of 29 ZIP codes in the county, only ZIP code 28262 had the required number of patients in all quasi-identifier categories. Henceforth, we use the terms 5-anonymized and 5-identifiable to refer to ZIP codes (or groups of them) having met and unmet the required number of patients for 5-anonymization, respectively.

We formulate a zero-one (or binary) linear program (BLP) to aggregate ZIP codes containing 5-identifiable patients in any quasi-identifier subgroup into ZIP code groups satisfying 5-anonymization. This can be achieved either by aggregating 5-identifiable ZIP codes with a 5-anonymized ZIP code or by aggregating two 5-identifiable ZIP

Table 3.1: ZIP codes in Mecklenburg County across all 10 quasi-identifier categories, where ○ represents fields having 5-identifiable patients (i.e., 59 of them) and ● represents fields having 5-anonymized patients (i.e., 231 of them). Actual numbers are obscured to protect against possible patient re-identification

Zip code	Female	Male	≤24 years	25–34 years	35-44 years	≥45 years	PAP	Medicare & Medicaid	Commercial & Self-pay	Unknown
28031	○	●	●	○	●	●	●	●	●	●
28036	○	●	●	●	○	●	●	○	●	●
28078	●	●	●	●	●	○	●	●	●	○
28211	●	●	●	●	●	○	●	●	●	○
28212	●	●	○	●	●	●	●	●	●	○
28262	●	●	●	●	●	●	●	●	●	●
28269	○	●	●	○	●	●	●	●	○	●
28205	●	●	○	○	●	●	●	●	○	●
28213	○	●	●	○	●	●	●	●	●	○
28215	●	●	○	○	●	●	●	●	●	○
28134	●	○	●	●	○	●	●	○	●	●
28273	●	●	●	●	●	○	●	●	●	●
28202	●	●	●	●	●	○	●	●	●	○
28206	●	○	●	●	○	●	○	●	●	○
28204	●	●	○	●	●	●	●	●	●	○
28207	○	●	●	○	●	●	●	○	●	●
28270	●	●	●	○	●	●	●	●	●	○
28277	●	●	●	●	●	○	●	○	●	●
28208	●	●	○	●	●	●	●	●	●	○
28278	●	●	●	●	○	●	●	●	●	●
28209	●	●	●	●	●	○	●	●	○	●
28210	○	●	○	●	●	●	●	○	●	●
28226	●	●	○	●	●	●	●	○	○	●
28214	○	●	●	○	●	●	●	○	●	●
28216	●	●	○	●	●	●	●	●	●	●
28203	●	●	●	●	●	●	●	●	●	○
28217	●	●	○	●	●	●	●	●	●	●
28105	●	●	●	●	●	○	●	○	○	●
28227	●	●	●	●	●	●	●	●	●	○

codes that satisfy 5-anonymization. We seek the largest number of ZIP code groups that satisfy the 5-anonymization in every group.

3.2.1 Zero-one (or binary) linear program (BLP)

3.2.1.1 Parameters:

Let $n = 29$ denote the number of ZIP codes and $m = 10$ denote the quasi-identifier categories for which we must achieve k-anonymization, where $k = 5$. We define two index sets $I \equiv \{1, \ldots, n\}$ and $Q \equiv \{1, \ldots, m\}$. Let v_i^q denote the number of patients in ZIP code $i \in I$ for quasi-identifier category $q \in Q$. Additionally, let the distance between the geographic centers of two ZIP codes $i \in I$ and $j \in I$ be denoted by $d_{i,j}$, which is calculated using the Haversine formula: [51, 52]

$$d_{i,j} = 2r \sin^{-1} \sqrt{h},$$

where $r = 6371$ is the radius of the earth in kilometers,

$$h = \sin^2 \left(\frac{\phi_i - \phi_j}{2} \right) + \cos(\phi_i) \cos(\phi_j) \sin^2 \left(\frac{\lambda_i - \lambda_j}{2} \right),$$

and (ϕ_i, λ_i) and (ϕ_j, λ_j) are the (latitudes, longitudes) in radians of the centers of ZIP codes i and j, respectively.

Furthermore, we manually compute an adjacency matrix such that $a_{i,j} \in \{0, 1\}$, for all $i \in I$ and $j \in I$, indicate if ZIP codes i and j are geographically adjacent to each other. That is if ZIP codes $i \in I$ and $j \in I$ are adjacent then $a_{i,j} = 1$ and if $i \in I$ and $j \in I$ are not adjacent then $a_{i,j} = 0$. We will restrict ZIP code groups to only consist of adjacent ZIP codes, which formally requires $a_{i,i} = 1$, for all $i \in I$.

3.2.1.2 Decision Variables:

The optimization model uses four decision variables, explained below:

- $x_i^g \in \{0, 1\}$. This decision variable indicates whether or not the ZIP code $i \in I$ is

assigned to group g. When the ZIP code $i \in I$ is assigned to group g, x_i^g must be equal to one and when ZIP code $i \in I$ is not assigned to group g, x_i^g must be equal to zero.

- $y^g \in \{0, 1\}$. This decision variable indicates whether or not group g contains at least one ZIP code in it. That is y^g must be one when group g is assigned at least one ZIP code and must be zero when group g is not assigned any ZIP code.

- $z^g \in \{0, 1\}$. This decision variable indicates whether group g contains at least two ZIP codes. When group g contains at least two ZIP codes, z^g will be equal to one, and if group g contains less than two ZIP codes, z^g is equal to zero.

- $x_{i,j}^g \in \{0, 1\}$. This decision variable indicates where ZIP codes $i \in I$ and $j \in I$ are both assigned to a single group g. When both ZIP codes $i \in I$ and $j \in I$ are assigned to group g, then $x_{i,j}^g$ must be one and $x_{i,j}^g$ must be zero when either or both ZIP codes $i \in I$ and $j \in I$ are not assigned to group g.

3.2.1.3 Objective Function:

Our overall objective is to minimize the total pairwise distances among the geographic centers of the ZIP codes in each group that satisfy (through the constraints below) a k-anonymized dataset. This affords intervention efforts to service groups that are not geographically distant from each other. Moreover, this objective keeps ZIP code groups as small as possible since it minimizes the sum of distances between ZIP codes in the same group—that is, any additional ZIP code added to the group will increase the value of the objective function. Formally, our objective function is expressed as:

$$\min \ \frac{1}{2} \cdot \sum_{i \in I} \sum_{j \in I} d_{i,j} \cdot \sum_{g \in I} x_{i,j}^g.$$

3.2.1.4 Constraints:

Six constraints detailed below are imposed to make sure aggregations are done propitious to the study.

1. The first constraint below makes sure all n ZIP codes must be assigned to a group, and no ZIP code is assigned to more than one group. This constraint makes sure that no ZIP code remains unassigned and the same ZIP code is not assigned to more than one group.

$$\sum_{g \in I} x_i^g = 1, \qquad \forall i \in I.$$

2. To appropriately indicate if group $g \in I$ has at least one ZIP code assigned to it (i.e., y^g), we use the following constraints:

$$\sum_{i \in I} x_i^g \geq y^g, \qquad \forall g \in I,$$
$$\sum_{i \in I} x_i^g \leq n \cdot y^g, \qquad \forall g \in I.$$

3. To indicate if group $g \in I$ has at least two ZIP codes assigned to it (i.e., z^g), we use the following constraints:

$$\sum_{i \in I} x_i^g - y^g \geq z^g, \qquad \forall g \in I,$$
$$\sum_{i \in I} x_i^g - y^g \leq n \cdot z^g, \qquad \forall g \in I.$$

4. To ensure every group formed by the aggregations achieves k-anonymization of patients (i.e., $k = 5$), we use the following constraint. This constraint ensures that the number of patients for each quasi-identifier category in each group is either

equal to zero or greater than or equal to k.

$$\left(\sum_{i\in I} v_i^q \cdot x_i^g\right)^2 \geq k \sum_{i\in I} v_i^q \cdot x_i^g, \qquad \forall q \in Q, g \in I$$

$$\sum_{i\in I}\sum_{j\in I} v_i^q \cdot v_j^q \cdot x_{i,j}^g \geq k \sum_{i\in I} v_i^q \cdot x_i^g, \qquad \forall q \in Q, g \in I.$$

5. To ensure only geographically adjacent ZIP codes are aggregated, we use the following constraint. This constraint makes sure that all the ZIP codes in a group are adjacent to at least one other ZIP code in the same group, which results in aggregation of only contiguous ZIP codes to a particular group.

$$\sum_{j\in I} a_{i,j} \cdot x_{i,j}^g \geq x_i^g \cdot z^g, \qquad \forall i \in I, \forall g \in I$$

6. To linearize the decision variable $x_{i,j}^g = x_i^g \cdot x_j^g$, we use the following constraints:

$$x_{i,j}^g \leq x_i^g, \qquad \forall i \in I, \forall j \in I, \forall g \in I,$$

$$x_{i,j}^g \leq x_j^g, \qquad \forall i \in I, \forall j \in I, \forall g \in I,$$

$$x_{i,j}^g \geq x_i^g + x_j^g - 1, \qquad \forall i \in I, \forall j \in I, \forall g \in I.$$

3.2.1.5 Implementation and Hyper-parameter Tuning:

We formulated and solved our BLP on Python[53] using a Gurobi solver.[54] We set some Gurobi tuning parameters to improve computation times. First, we prioritize finding optimal solutions (i.e., via MIPFocus = 2) [55]. Second, we set branching priorities to the maximum values for the decision variables, x_i^g, and minimum values for the other (dependent) variables, $x_{i,j}^g$, y^g, and z^g, for all $i \in I$, $j \in I$, and $g \in I$.

3.2.2 Heuristic aggregation algorithm

To evaluate the performance of the optimization model, we developed a heuristic approach with which to compare the resulting ZIP code aggregations. The heuristic uses a myopic, multi-step k-anonymization per quasi-identifier. It k-anonymizes first by sex, followed by age group, then by insurance type. The algorithm uses the following steps per quasi-identifier:

1. Each ZIP code containing zero or at least k patients in a considered quasi-identifier form their own group.

2. Any ZIP code not in a group (i.e., those with one to four patients) are added into the group containing the ZIP code with the smallest (Haversine) distance to it, as well as containing at least one adjacent ZIP code.

3. We check all formed groups to see if any large group can be broken down into smaller groups. Within each group, any disjoint subgroups all-satisfying k-anonymization are re-formed.

After Step 3, newly aggregated groups are considered as new ZIP codes in Step 1.

3.3 Results

Our BLP aggregated the 29 ZIP codes into 13 ZIP code groups of geographically adjacent ZIP codes. In contrast, the heuristic approach aggregated the 29 ZIP codes into nine ZIP code groups. Figure 3.1 shows the ZIP codes and the k-anonymized groups resulting from both algorithms. Moreover, our BLP yielded more groups than using the generalization technique via the first-three and first-four digits of the ZIP codes, which result in three (k-anonymized) and nine (not all k-anonymized) groups, respectively.

Maximizing the number of ZIP code groups affords more granular monitoring of changes in PrEP use. For example, ZIP codes 28203 and 28217 are grouped into

ZIP code group 11 using the optimization model since they both together satisfy the threshold constraint. In contrast, the heuristic algorithm has grouped these ZIP codes into ZIP code group 7 along with three other ZIP codes already in group 7. This demonstrates the heuristic algorithm yields suboptimal solutions. Furthermore, using our BLP affords the sharing of k-anonymized data to afford to conduct analyses at more granular levels.

Table 3.2: Aggregated ZIP code groups from the optimization model and the heuristic algorithm

Zip code	Group number assigned by Optimization model	Group number assigned by Heuristic Algorithm
28031	0	0
28036	0	0
28078	0	0
28211	1	5
28212	1	8
28262	2	1
28269	2	6
28205	3	2
28213	3	2
28215	3	8
28134	4	5
28273	4	3
28202	5	7
28206	5	1
28204	6	7
28207	6	2
28270	7	5

Table 3.2: Aggregated ZIP code groups from the optimization model and the heuristic algorithm

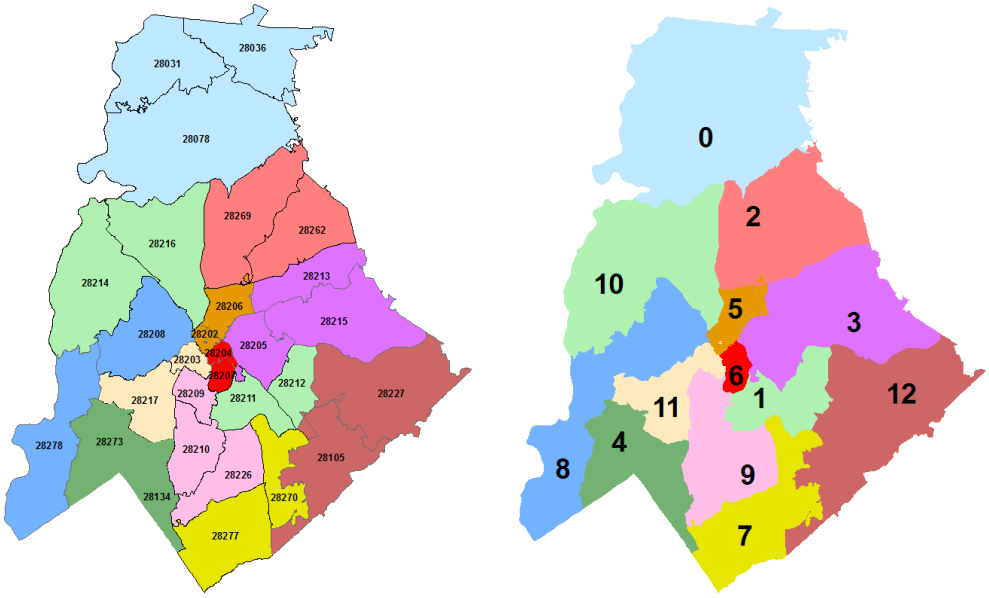
Zip code	Group number assigned by Optimization model	Group number assigned by Heuristic Algorithm
28277	7	8
28208	8	7
28278	8	4
28209	9	7
28210	9	3
28226	9	5
28214	10	4
28216	10	6
28203	11	7
28217	11	7
28105	12	5
28227	12	9

3.4 Conclusion

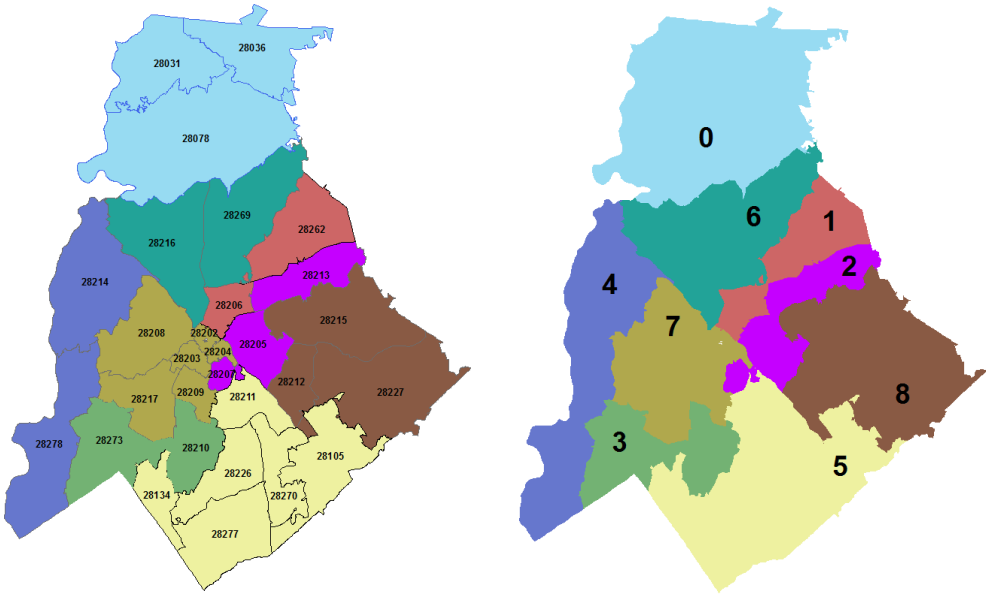
De-identification of PHI has vast potential in advancing medical research through HIPAA-compliant data-sharing. One of the widely practiced methods to afford this is k-anonymization, which we used to afford PRA Health Science's release of TDF+FTC and TAF+FTC prescription data for our ongoing study for evaluating PrEP use in Mecklenburg County. We formulated and used a BLP to maximize the number of ZIP code groups while satisfying k-anonymization across quasi-identifiers of patient sex, age group, and insurance type. Our BLP generated 13 ZIP code groups of 5-anonymized patient prescription records, which outperforms heuristic and other naïve generalization approaches, thus affording more granular analyses.

Our BLP can be used as an alternative method to k-anonymize geographic areas in studies involving PHI, allowing the sharing of de-identified PHI where the traditional techniques of suppression, naïve generalization, and perturbation are unsuitable. Our model can be extended to use different proximity (distance) metrics, such as economic, housing, demographic characteristics, and/or other health indicators in the objective function for aggregating ZIP codes similar in these dimensions. This can allow us to see patterns in changing PrEP prescription, incidences of sexually transmitted infections (STIs), or other epidemiological characteristics in different ZIP code groups. For example, we could use a ZIP code's economic characteristics as a proximity metric to aggregate ZIP codes and compare the difference in PrEP prescriptions among these economically similar ZIP code groups. This allows public health departments to devise targeted interventions in economically weak ZIP codes.

We also can use our BLP to analyze clusters of geospatially similar patterns in PrEP prescriptions, STI incidences, or other epidemiological characteristics. These other indicators can be used as values for imposing threshold constraints to expose clusters with unusually high or low incidences.



All 29 ZIP codes ZIP code groups by BLP



All 29 ZIP codes ZIP code groups by heuristic

Figure 3.1: All 29 ZIP codes in Mecklenburg County (left) and de-identified ZIP code groups (right) by our BLP and heuristic approach, where boundaries are separated by different colors

CHAPTER 4: BOUNDS FOR THE LIKELY PREP-TO-NEED RATIO

4.1 Background

Previous studies on PrEP usage are conducted at the national, state, or county-wide levels but none at the sub-county (or ZIP code) level. These studies fail to capture the trends in PrEP usage within a county and, in turn, cannot provide insights for subsequent actions by county public health officials. This chapter quantifies PrEP usage within the county (at the ZIP code group level) in various demographic groups using the pharmacy claims data. This affords MCPH to use the PrEP metrics to devise targeted interventions for those underserved.

TDF+FTC and TAF+FTC currently are the only FDA-approved drugs for PrEP. However, they may or may not be taken with other antiretrovirals (ARVs), for use as HIV treatment, post-exposure prophylaxis (PEP), and chronic hepatitis B off-label management [28]. This makes it important to identify the reason why a patient used a particular TDF+FTC and TAF+FTC episode. The pharmacy claims dataset from PRA did not contain any diagnosis codes for patients, making it impossible to identify the reason TDF+FTC and TAF+FTC was used directly from the dataset. This calls for a need to identify and differentiate episodes of TDF+FTC and TAF+FTC used for PrEP from other uses. In this chapter, we also derive a rule-based algorithm based on the prescription histories of patients in the pharmacy claims dataset in order to identify the reason for using TDF+FTC and TAF+FTC for PrEP, PEP, HIV treatment or unknown use.

Traditional metrics like the PrEP prevalence defined as PrEP users per population and the PrEP-to-need ratio (PnR) defined as the ratio of PrEP users to new HIV incidences [26] can sometimes probably overestimate PrEP coverage. These metrics

consider any patient with a partial PrEP prescription in a year as a PrEP user for that year. This overestimates the PrEP coverage as PrEP should be taken by the patient throughout the year for it to provide sufficient protection for the entire year. Recently, PrEP is being used on-demand to offer patient protection around the times when the patient is at risk for acquiring HIV and not throughout the entire year (not an FDA approved usage). Hence, in this chapter we establish a lower and upper bounds for the *likely* PnR and demonstrate its use in Mecklenburg County, NC.

4.2 Methods

The Office of Research Protections and Integrity of the University of North Carolina at Charlotte exempted our study from Internal Review Board (IRB) review and allowed us to use the data obtained from Pharmaceutical Health Associates (PRA) Health Sciences and perform analyses on the 2012–2019 extract of the pharmaceutical claims data for Mecklenburg County involving TDF+FTC and TAF+FTC whether or not accompanied by ARVs.

4.2.1 Population, data, and study design

The study population used in the development of this algorithm consists of patients having been prescribed the drugs of interest, TDF+FTC and TAF+FTC, with or without any accompanying ARVs from 2012 to 2019 within Mecklenburg County, NC. The pharmacy claims dataset obtained from PRA Health Sciences contained 3,027 unique patients having filled prescriptions of TDF+FTC and TAF+FTC in Mecklenburg County from 2012 to 2019 (§2.1). There are a total of 1,023,961 pills (daily-doses) of TDF+FTC and 26,340 pills of TAF+FTC dispensed in Mecklenburg County, NC as recorded in the dataset. Prescription records for TDF+FTC and TAF+FTC started in January 1, 2012 and October 1, 2019, respectively. The recorded supply of TDF+FTC and TAF+FTC pills lasted until June 19, 2020 and May 19, 2020, respectively; however, we right-censor supplies after December 31, 2019

since we conduct analyses on a yearly basis. There were 40 unique ARV combinations; those that overlapped were merged together as single ARV episodes.

To determine PrEP use of TDF+FTC and TAF+FTC, we developed an algorithm for identifying PrEP regimens. To report PrEP coverage, we calculated the population-adjusted PrEP patients and bounds for the likely PnR. Population-adjusted PrEP patients are calculated as the number of PrEP patients per 100,000 people (in a demographic of interest) in Mecklenburg County, NC as shown in Equation 4.1.

$$\begin{matrix}\text{Population-adjusted PrEP} \\ \text{patients per 100,000}\end{matrix} = \frac{\text{Number of unique PrEP users in a year}}{\text{Population in the year}} * 100000. \tag{4.1}$$

To establish the upper bound for the likely PnR, we use the PnR which is an overestimate of PrEP coverage, and we use the PnR metric derived by Siegler et al., [26] for this. It is calculated as the ratio of number of unique patients on PrEP over the number of new HIV infections in that year shown in Equation (4.2).

$$\text{PrEP-to-need ratio (PnR)} = \frac{\text{Number of unique PrEP patients in the year}}{\text{Number of new HIV incidences in the year}}. \tag{4.2}$$

The lower bound will be a conservative estimate of PrEP coverage. We define a novel metric called the dose-adjusted PrEP-to-need ratio (daPnR), which requires a dose-adjusted PrEP patient protected for the full year, calculated by dividing the number of pills dispensed for PrEP in a year by the number of days in that year. Thus, the daPnR will be the ratio of dose-adjusted PrEP patients to the number of new HIV infections in a year shown in Equation (4.3).

$$\text{dose-adjusted PrEP-to-need ratio (daPnR)} = \frac{\frac{\text{Number of PrEP pills dispensed in the year}}{\text{Number of days in the year}}}{\text{Number of new HIV incidences in the year}}. \quad (4.3)$$

4.2.2 Regimen classification

4.2.2.1 Decision rules

A medical doctor with expertise in the different uses of TDF+FTC/TAF+FTC (ARV treatment, PrEP and PEP usage scenarios) clinically assessed the entire dataset and flagged each prescription episode as either for PrEP, PEP, HIV or unknown use. We encoded those assessments as rules to automatically flag the prescription episodes, which was subsequently reviewed for clinical validity. An iterative process comparing clinical and algorithmic assessments was carried out to derive programming rules for automatically flagging the prescription episodes. We finalized 11 rules to encompass all cases and reliably identify each prescription episode detailed in Table 4.1.

4.2.2.2 Description of the algorithm

We extract each patient's prescription histories individually; this contains the pharmacy claims history for TDF+FTC, TAF+FTC, and any ARVs for the patient from 2012 to 2019. We use the pharmacy claims data in 2012 to have sufficient history of the patient for the identification algorithm. However, we do not report the PrEP patient and pill counts in 2012 because we do not have sufficient history of the patients before 2012 to confidently identify PrEP episodes from the others. The pharmacy claims histories for each unique patient are followed from their first prescription date to the last, sorted by the drug name (ARV before TAF+FTC and then TDF+FTC). As explained below, the set of 11 decision rules (Table 4.1) was applied to each prescription episode.

Two similar prescription episodes with a day gap are considered a single continuing

Table 4.1: Decision rules applied for classifying patient's prescription episodes as either PrEP, PEP, HIV treatment or unknown use

Rule #	Decision Rule
R1	Merge same-drug regimen episodes having one-day gaps as a single episode.
R2	Combine concurrent or overlapping ARV episodes as a single ARV episode, and absorb any overlapping TDF+FTC/TAF+FTC (partial) pills with the ARV episode (but not the non-overlapping partial episodes).
R3a	Classify any ARV episodes of more than 36 pills as HIV treatment; also, classify all subsequent episodes as HIV treatment.
R3b1	Classify any ARV episodes of between 21 and 36 pills (inclusive) as PEP regimens, except if the next episode is within 21 days.
R3b2	Classify any ARV episodes of between 21 and 36 pills (inclusive) as HIV treatment only if the next ARV episode is within 21 days; also, classify all subsequent episodes as HIV treatment.
R3c	Classify any ARV episodes of less than 21 pills as unknown use.
R4a	Classify any TDF+FTC/TAF+FTC monotheraphy episodes of at least 21 pills as PrEP regimens.
R4b	Classify any TDF+FTC/TAF+FTC (partial) episodes of less than 21 pills as unknown use.
R5a	For any HIV treatment ending in TDF+FTC/TAF+FTC monotherapy episodes of 60 or more pills (i.e., previously marked as HIV treatment by R3a or R3b2), reclassify them as PrEP regimens, if the ARV episodes triggering R3a or R3b2 consists of at most 120 pills.
R5b	Classify any TDF+FTC/TAF+FTC episodes satisfying Rule R5a and having less than 21 pills as unknown use.
R6	Classify all left-censored TDF+FTC/TAF+FTC episodes preceding an HIV treatment as HIV treatment.

episode and are merged using Rule R1. The end date in the new merged episode will be the end date of the last episode following the day gap, and the pills used in the entire episode will be the sum of the days covered by the merged episodes.

Any concurrent or overlapping ARV (i.e., combination ARV) prescription episodes, regardless of the ARV type used, are considered a single ARV episode and are merged using Rule R2. In such a case, the end date of the merged episode is the end date of the last running episode, and the number of pills in the episode is the sum of the days covered by the merged episodes. In addition, suppose an episode of TDF+FTC/TAF+FTC is overlapping with an episode of other ARVs. In that case, these overlapping episodes of TDF+FTC/TAF+FTC are considered part of

the combination therapy with other ARVs. We merge these overlapping pills of TDF+FTC/TAF+FTC with the ARV as a single ARV combination therapy episode. The new end date in the merged episode is the end date of the last episode, and the pills dispensed is the sum of the days covered by the merged episodes. Any other episodes of TDF+FTC/TAF+FTC not completely overlapping with ARVs are considered as a separate TDF+FTC/TAF+FTC monotherapy episodes.

Next, decision Rule R3a is applied for all ARV episodes to check if a patient has used these for possible HIV treatment. We consider a single continuous ARV episode of more than 36 days as an HIV treatment episode. Suppose a patient was found to be prescribed an ARV episode of more than 36 days, the corresponding episode is classified as an HIV treatment episode, and the patient is considered HIV positive. All following episodes, including ARVs and TDF+FTC/TAF+FTC for the patient from this time on, will be classified as for HIV treatment.

If decision Rule R3a was not satisfied, i.e., for ARV episodes of less than 36 days. Decision Rule R3b1 will be applied to check if the ARV episode is for PEP. The episode will be classified as PEP if the ARV was prescribed for between 21 and 36 days (inclusive) with no evidence of previous HIV treatment as identified by Rule R3a in that patient. Once identified as a PEP episode, by the decision Rule R3b1, the patient's previous ARV prescription episode will be examined using Rule R3b2. If the previous ARV episode of that patient is within 21 days, we consider both these episodes as HIV treatment episodes. These two ARV episodes and all subsequent episodes will be flagged as for HIV treatment according to Rule R3b2. If not, the two ARV episodes with more than 21 days between them will be flagged as PEP episodes.

An ARV episode not satisfying all three decision rules R3a, R3b1, or R3b2—those short ARV episodes prescribed for less than 21 days—are flagged as for unknown use according to Rule R3c. Any ARV episode for unknown use continuing beyond December 31, 2020 are right-censored.

Any prescription episode that is not an ARV—that is, for TDF+FTC/TAF+FTC episodes—rule R4a is checked: any TDF+FTC/TAF+FTC episodes of at least 21 days will be classified as a PrEP episode, provided that the patient has no history of previous HIV treatment regimen identified by rules R3a or R3b2.

TDF+FTC/TAF+FTC episodes not satisfying Rule R4a, will be flagged as right-censored episodes if the prescription carries on after 2020. If the TDF+FTC/TAF+FTC episode is for less than 21 days without right-censoring, these episodes will be classified as for unknown use episodes according to Rule R4b.

Next, we check for TDF+FTC/TAF+FTC monotherapy episodes (TDF+FTC or TAF+FTC in the absence of any other retroviral) of 60 days or more in patients previously flagged as HIV-positive by rules R3a or R3b2. If the previous HIV treatment episode in the patient was limited to fewer than 120 pills of ARV, we suspect these HIV treatment episodes were possibly used for consecutive PEP instead and re-flag them as PEP episodes. All following TDF+FTC/TAF+FTC monotherapy episodes in such a patient will be classified as PrEP episodes according to Rule R5a or as for unknown use according to Rule R5b.

In the last step, Rule R6, we check for left-censored ARVs. Suppose a patient has an ARV prescribed that is censored before 2012. If the next (uncensored) prescription episode is for HIV treatment, we suspect that the left-censored episode of ARV may be for HIV treatment since we do not have sufficient prescription history for that particular patient before 2012. A flowchart for the identification algorithm is shown in Figure 4.3.

Figure 4.1 and Figure 4.2 illustrate two hypothetical patients and how the 11 decision rules are applied in varying cases. In both figures, the first panels show the drug episodes before applying the rules; the second panels show how rules R1 and R2 apply; the third panels show how rules R3* and R4* apply; and the fourth panel (in Figure 4.2) shows how rules R5* and R6 apply.

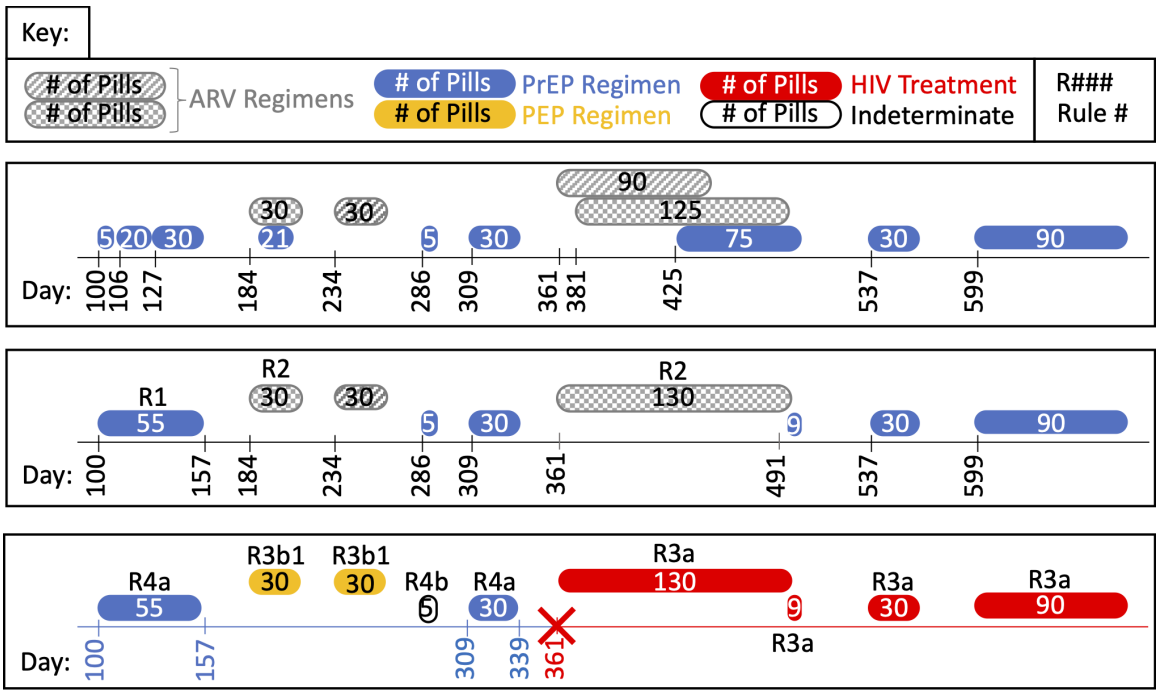


Figure 4.1: Illustration PrEP identification algorithm for a patient with HIV treatment episode (Patient 1)

For Patient 1 (in Figure 4.1), the three separate episodes of TDF+FTC/TAF+FTC of 5, 20, and 30 pills starting from Day 100 have a day gap between them. Rule R1 merges these three episodes into one single episode with a start date as Day 100, an end date as Day 157, with the number of pills still at 55. Next, the TDF+FTC/TAF+FTC episode starting at Day 189 is merged with the ARV episode starting on Day 184 according to Rule R2. Similarly, the two ARV and a TDF+FTC/TAF+FTC episodes starting on Day 361, Day 381, and Day 425 are merged according to Rule R2. Since the TDF+FTC/TAF+FTC episode starting on Day 425 ends nine days after the overlapping ARV episode, we consider these nine days as TDF+FTC/TAF+FTC monotherapy days. The TDF+FTC/TAF+FTC episode of five days starting from Day 286 is not considered for PrEP but rather for unknown use according to Rule R4b. The last panel in Figure 4.1 shows the final assessment made by the algorithm. The two TDF+FTC/TAF+FTC episodes starting from Day 100 and Day 309 are

assessed as PrEP episodes of 55 and 30 pills ending on Day 157 and Day 339, respectively. The ARV episode starting from Day 361 is assessed to be for HIV treatment according to Rule R3a. All the subsequent episodes from Day 361 (including TDF+FTC/TAF+FTC episodes) are considered for HIV treatment and are discarded.

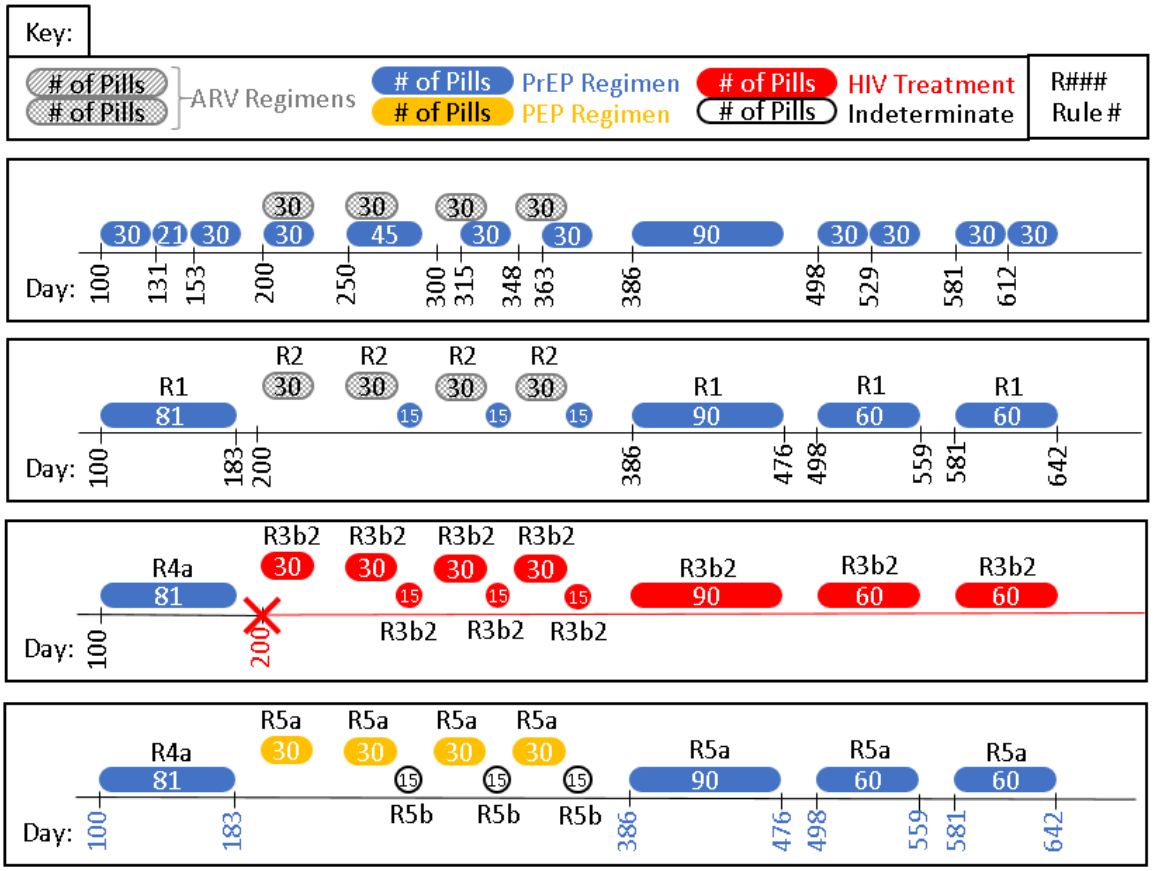


Figure 4.2: Illustration of PrEP identification algorithm for a patient with PEP episodes (Patient 2)

In Patient 2 (in Figure 4.2), the three TDF+FTC/TAF+FTC episodes starting from Day 100 with a day gap between them are merged as a single episode of 81 pills ending on Day 183 according to Rule R1. The overlapping TDF+FTC/TAF+FTC episodes from Day 200 are merged with the corresponding ARV episodes according to Rule R2. The third panel in Figure 4.2 shows how the algorithm first assessed

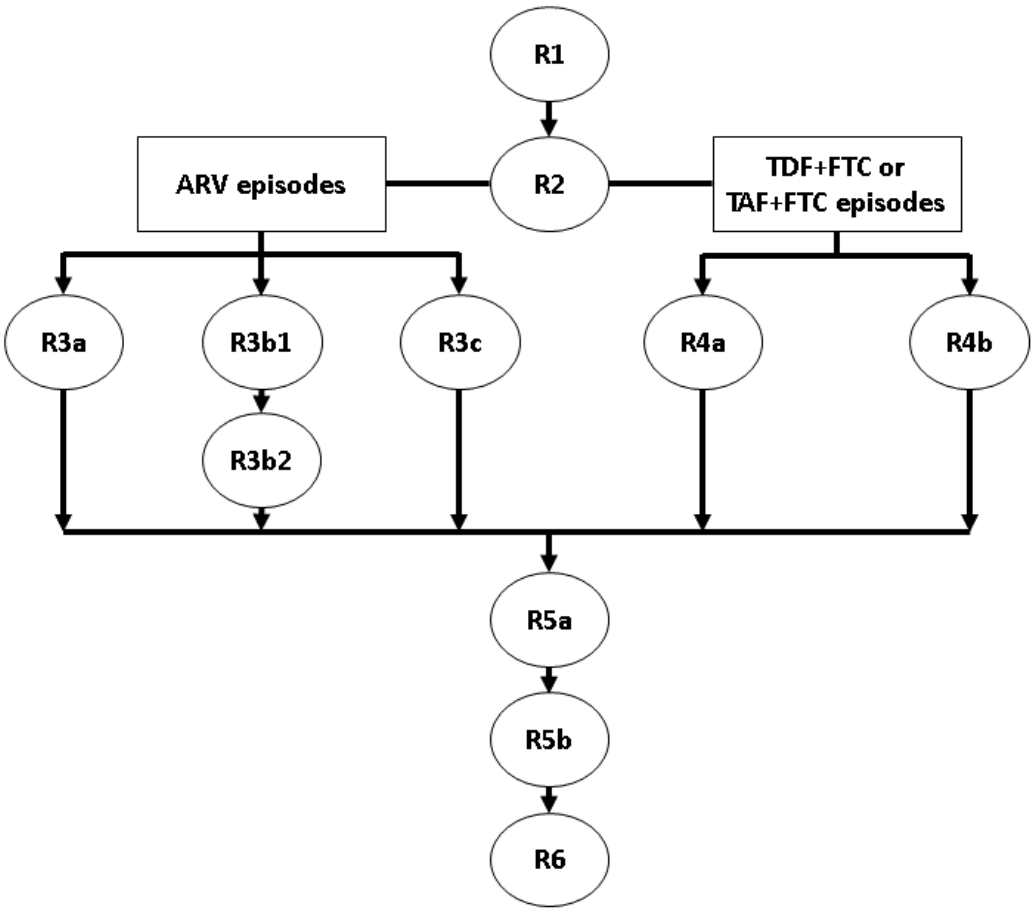


Figure 4.3: PrEP identification algorithm flowchart

the ARVs from Day 200 for HIV treatment according to Rule R3b2 because of the first two consecutive ARV episodes of 21 pills or more within 21 days apart, and it and classified all subsequent TDF+FTC/TAF+FTC and ARV episodes as HIV treatment. However, because the HIV treatment ends with at 210 ($\geq$ 60) pills of TDF+FTC/TAF+FTC monotherapy (i.e., in the first and second panels in Figure 4.2) and the patient was classified as HIV positive because of (at most) 120 pills of ARV, we apply Rule R5a and reconsider the ARV pills as PEP regimens and the subsequent TDF+FTC/TAF+FTC pills as PrEP regimens (i.e., in the last panel). The last panel in Figure 4.2 shows the final assessment made by the algorithm. Patient 2 was assessed to have four episodes of PrEP prescriptions: an 81-pills episode

starting from Day 100 to Day 183 according to Rule R4a and a 90-pills episode from Day 386 to Day 476, a 61-days episode from Day 498 to Day 559, and a 60-pills episode from Day 581 to Day 642 as identified by Rule R5a. The four ARV episodes between Day 200 and Day 393 are assessed as PEP episodes according to Rule R5a. The three 15 pills episodes of TDF+FTC/TAF+FTC between Day 315 and Day 378 are flagged as for unknown use according to Rule R5b and are not considered PrEP episodes.

4.3 Results

4.3.1 Patients and dosing of PrEP medications

The PRA pharmacy claims data contained 1,050,301 pills of TDF+FTC/TAF+FTC dispensed for 3,027 patients in Mecklenburg County, NC between 2012 and 2019. The PrEP regimen identification algorithm identified 540,003 pills of TDF+FTC/TAF+FTC were used for HIV treatment, 23,406 pills were used for PEP, and 2,637 pills were for *unknown* use. A total of 938 patients were identified as using either TDF+FTC or TAF+FTC for reasons other than PrEP. Regimen histories in 2012 were left-censored to provide us enough information about a patient's prescription history to confidently identify the reason for regimen use. Any regimen continuing after 2019 was also right-censored and left out of the subsequent analyses. Around 44 patients were using about 17,730 PrEP pills outside 2013–2019. Thus, we identified 2,045 patients were using PrEP, and around 466,525 pills were dispensed in Mecklenburg County from 2013–2019. The PrEP patients and pills included and excluded in the study are shown in Figure 4.4.

4.3.2 Annual PrEP use in the entire county

There was an increasing trend annually in both the number of PrEP patients and the number of PrEP pills in the county from 2013–2019. The population-adjusted PrEP patients per 100,000 people in the entire county increased from 5.64 in 2013 to

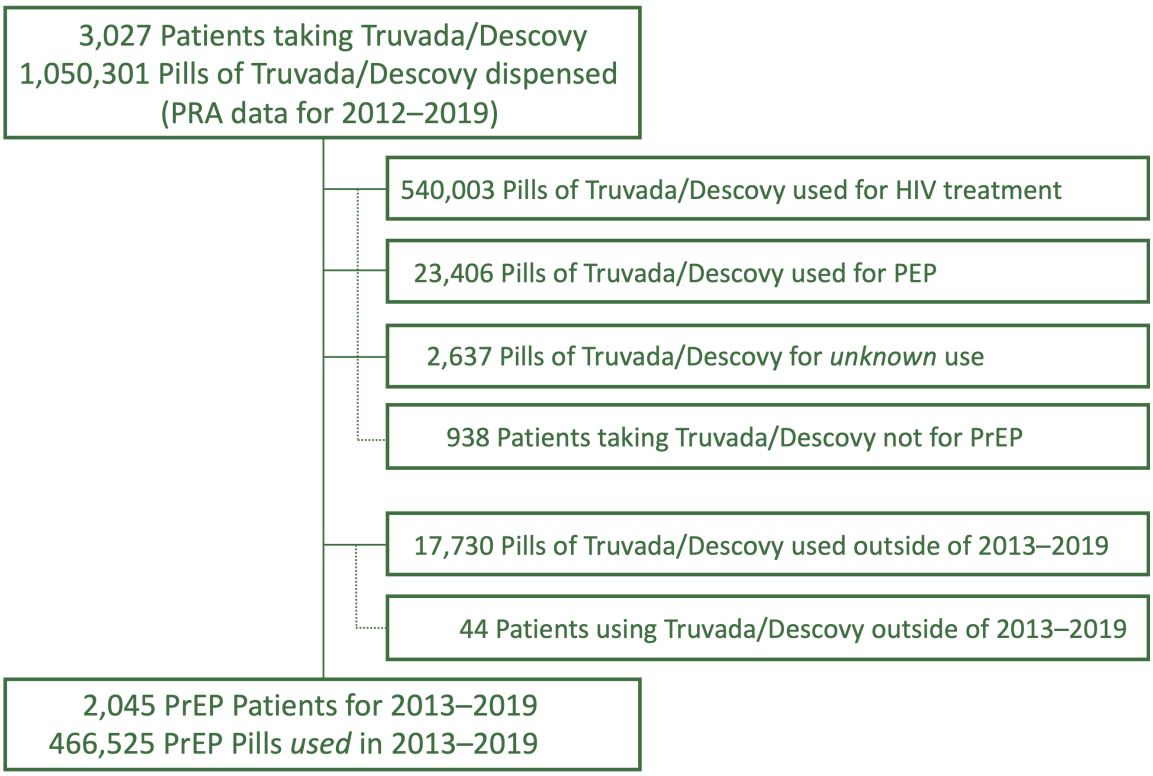


Figure 4.4: Patients and dosing of PrEP medications

106.39 in 2019 as shown in Figure 4.5. The number of PrEP pills dispensed in the county also increased from 3,609 in 2013 to 187,050 in 2019 as shown in Figure 4.6. The PnR (i.e., the upper bound) increased from 0.32749 in 2013 to 4.96281 in 2019. The daPnR (i.e., the lower bound) increased from 0.058 in 2013 to 2.118 in 2019. The PnR bounds for the entire county from 2013 to 2019 are shown in Figure 4.7.

4.3.3 Annual PrEP use by sex

PrEP use in both sexes increased annually from 2013 to 2019. However, we see huge disparities in PrEP use among men and women. The population-adjusted male PrEP patients per 100,000 increased from 8.14 to 204.06, and the population-adjusted women PrEP patients per 100,000 increased from 3.31 in 2013 to 15.72 in 2019. The annual trend in the number of PrEP pills also increased throughout the study period in both sexes. The number of PrEP pills used by men increased from 2,953 in 2013 to

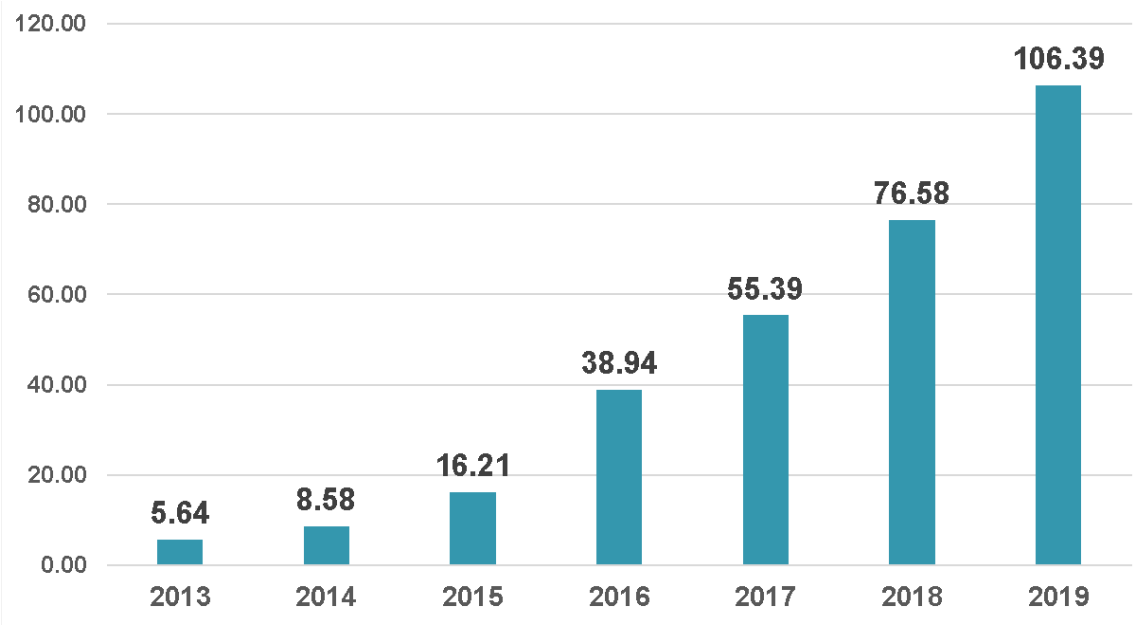


Figure 4.5: Population-adjusted PrEP patients per 100,000 in Mecklenburg County from 2013 to 2019

175,542 in 2019. At the same time, the number of PrEP pills used by women increased from 656 in 2013 to 11,508 in 2019. Figure 4.8 and Figure 4.9 show the population-adjusted PrEP patients and PrEP pills dispensed annually from 2013–2019 in men and women across Mecklenburg County.

The PnR (i.e., the upper bound), increased from 0.285 in 2013 to 5.601 in 2019 in men. In comparison, it increased from 0.5 in 2013 to 2.1 in 2019 in women. The daPnR (i.e., the lower bound), increased from 0.059 to 2.429 in men and 0.05 to 0.72 in women from 2013 to 2019. The plot of the PnR bounds by sex from 2013 to 2019 is shown in Figure 4.10.

4.3.4 Annual PrEP use by age groups

We do not have access to population data broken for age groups 25–34 years and 35–44 years. Hence, we report population-adjusted PrEP patients broken down into three age groups ($\leq$ 24 years, 25–44 years, and $\geq$45 years) and report the PrEP pills dispensed and PnR bounds for four age groups ($\leq$ 24 years, 25–34 years, 35–44

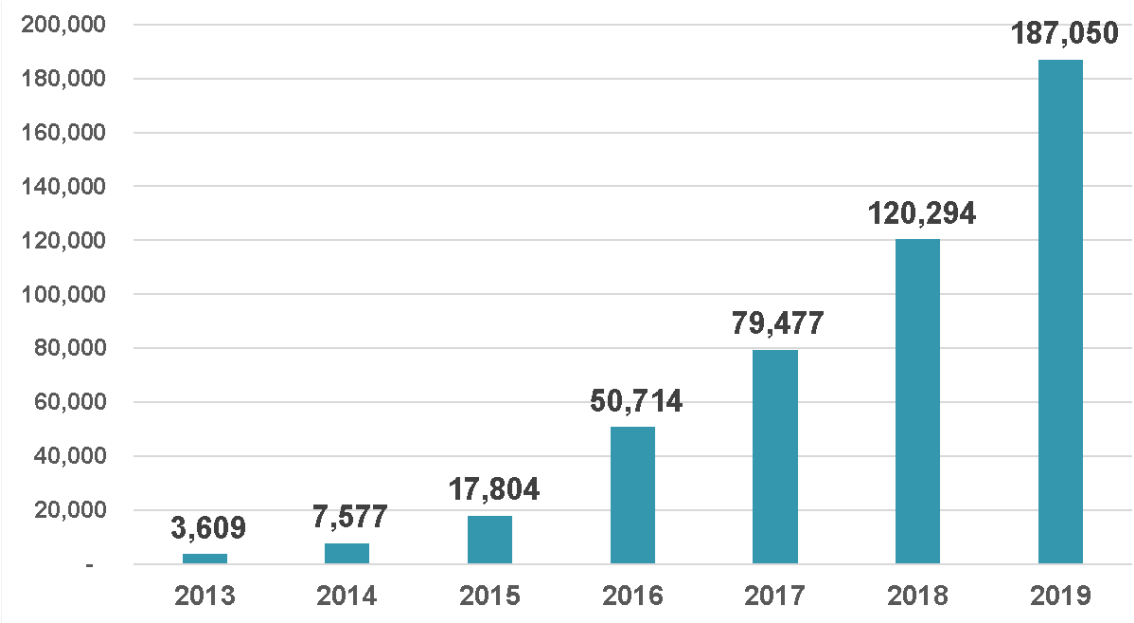


Figure 4.6: PrEP pills dispensed in Mecklenburg County from 2013 to 2019

years, and ≥45 years). The population-adjusted PrEP patients per 100,000 were the highest in the age-group 25–44 years in all the years from 2013–2019. The population-adjusted PrEP patients in the age-group 25–44 years increased from 30 in 2013 to 784 in 2019. From 2013 to 2017, the age group with the second-highest PrEP coverage per 100,000 is ≥45 years, with the population-adjusted PrEP patients of 22 in 2013 and 385 in 2017. In 2018 and 2019, the age group with the second-highest PrEP coverage per 100,000 is ≤ 24 years, with 70 and 136 PrEP patients per 100,000, respectively. The population-adjusted PrEP patients per 100,000 in different age groups across the county from 2013–2019 is shown in Figure 4.11.

We see an increase in PrEP pills in all age groups every year throughout the study period shown in Figure 4.12. Age group 35–44 years have the highest number of pills dispensed for PrEP from 2013 to 2015. From 2016 to 2019 the age group 25–34 years has the highest number of PrEP pills dispensed. The PnR bounds for all age groups from 2013 to 2019 is shown in Figure 4.13, where we see an annual rise in both the upper and lower bound in all the age groups throughout the study period.

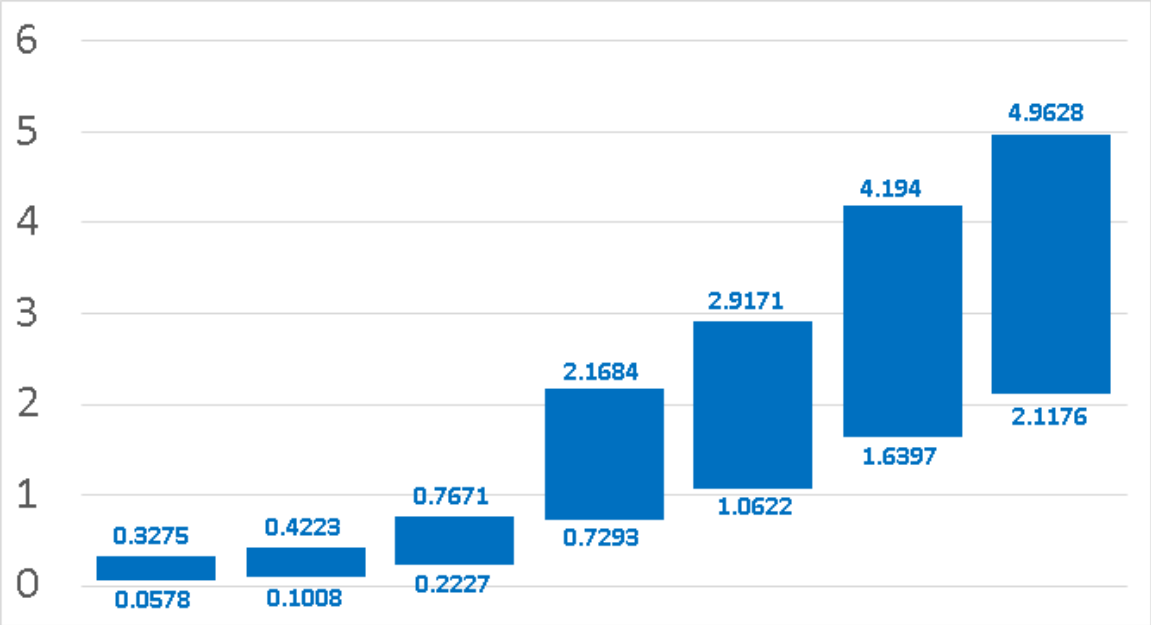


Figure 4.7: PrEP-to-need ratio bounds in Mecklenburg County from 2013 to 2019

4.3.5 Annual PrEP use by ZIP code groups

The population-adjusted PrEP patients per 100,000 people across different geographic areas (13 ZIP code groups) across the county is shown in Table 4.2. The annual trend in the population-adjusted PrEP patients increased in all the ZIP code groups throughout 2013 to 2019. ZIP code groups 5, 6, and 11 are the top three groups with the highest PrEP patients per 100,000 people in 2019 at 276.33, 717.07, and 257.93, respectively. ZIP code groups 7, 9, and 12 are the last three in population-adjusted PrEP patients in 2019 at 49.62, 48.53, and 52.99, respectively.

The number of PrEP pills dispensed every year from 2013 to 2019 in all 13 ZIP code groups is shown in Table 4.3. Zipcode group 3 has the highest number of pills filled in 2019 at 46,133, and ZIP code group 1 has the least number of pills dispensed at 5,835 pills in 2019.

The PnR (i.e., the upper bound) for all 13 ZIP code groups is shown in Table 4.4. ZIP code group 6 has the highest PnR in 2019 at 22.67, and ZIP code group 8 has the least PnR in 2019 at 1.74. The daPnR (i.e., the lower bound) for all ZIP code

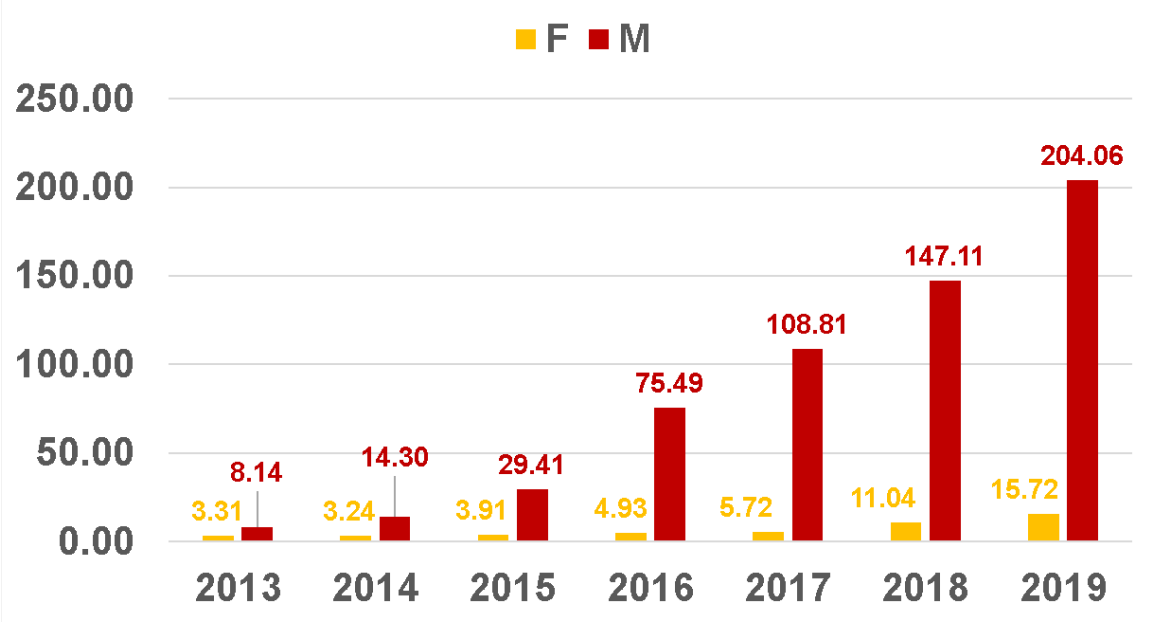


Figure 4.8: Population-adjusted PrEP patients per 100,000 by sex in Mecklenburg County from 2013 to 2019

groups, is shown in Table 4.5. Here again, ZIP code group 6 has the highest daPnR of 10.236, and ZIP code group 8 has the least daPnR of 0.546 in 2019.

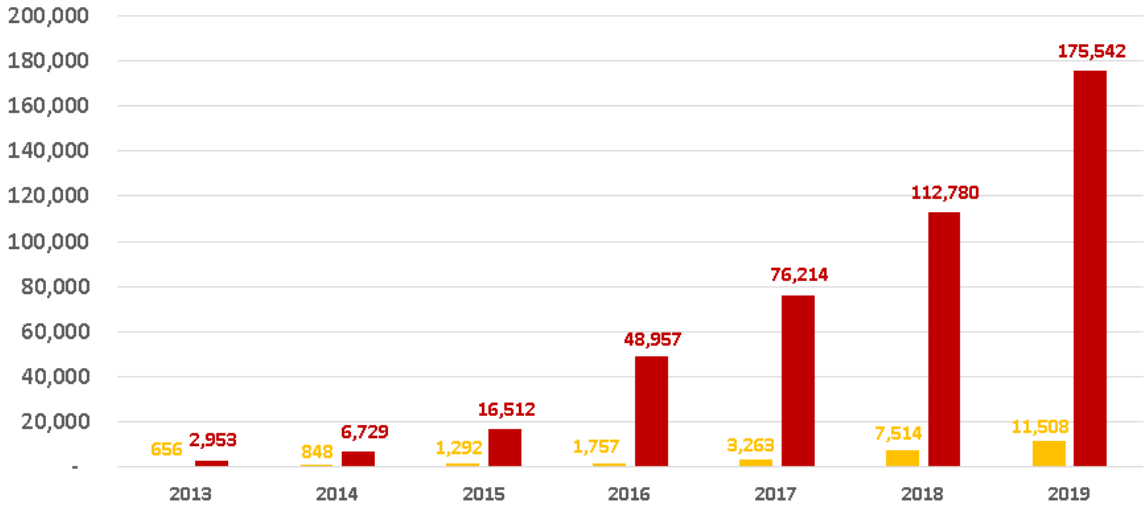


Figure 4.9: PrEP pills dispensed in Mecklenburg County from 2013 to 2019 by sex

4.4 Conclusion

We analyzed trends in PrEP use within Mecklenburg County, NC, for the first time using pharmacy claims data for TDF+FTC and TAF+FTC. This quantifies the PrEP use within the county from 2013 to 2019, allowing public health planners to see changes in trends in PrEP usage over the years in the county. The analyses conducted in different demographic (sex and age) and geographic (ZIP code groups) groups allow Mecklenburg County Public Health Department to devise demographically and geographically targeted interventions to increase PrEP usage in underserved groups.

We see the contrast between PrEP patients and PrEP pills in some of the groups. For example, we see that PrEP patients are highest in the 25–34 years group, but the group greater than 45 years uses the most PrEP pills. This could indicate stricter adherence to medication by age group greater than 45 years compared to the 25–34 years group, which could be further examined by looking at the $\frac{\text{PrEP pills}}{\text{PrEP patients}}$ ratio.

PrEP usage is analyzed via metrics, such as the PrEP prevalence (i.e., population-adjusted PrEP patients) and PrEP-to-need bounds, which can be used to contrast PrEP coverage with respect to their metric differences, such as population and new

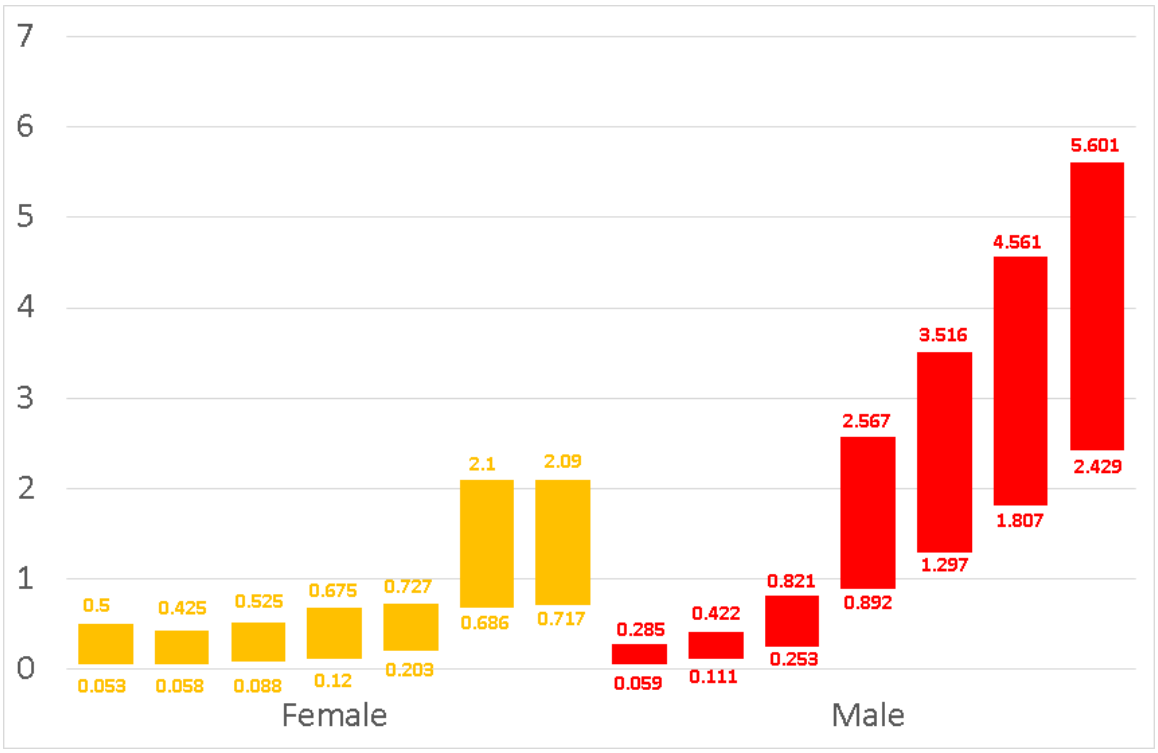


Figure 4.10: PrEP-to-need ratio bounds in Mecklenburg County from 2013 to 2019 by sex

HIV incidences, respectively. For example, ZIP code group 5 had high PrEP prevalence but mediocre PnR bounds; this indicates that this group had a relatively higher number of PrEP patients with respect to their population than most other groups, however the PrEP patients compared to the need is middling among all groups. This implies that the population-adjusted need is relatively higher than its population-adjusted PrEP patients. Conversely, ZIP code group 0 had a relatively low PrEP prevalence but high PnR bounds; this indicates ZIP code group 0, even with low PrEP prevalence, is outperforming most other groups in terms of PrEP-to-need.

We see considerably high number of population-adjusted PrEP users, PrEP pills dispensed, and PnR bounds in ZIP code group 6. We suspect this high number to be caused by the presence of two hospitals, one pharmacy, and less residents in the ZIP code group. Since, we use pharmacy ZIP codes and not the patient's residential

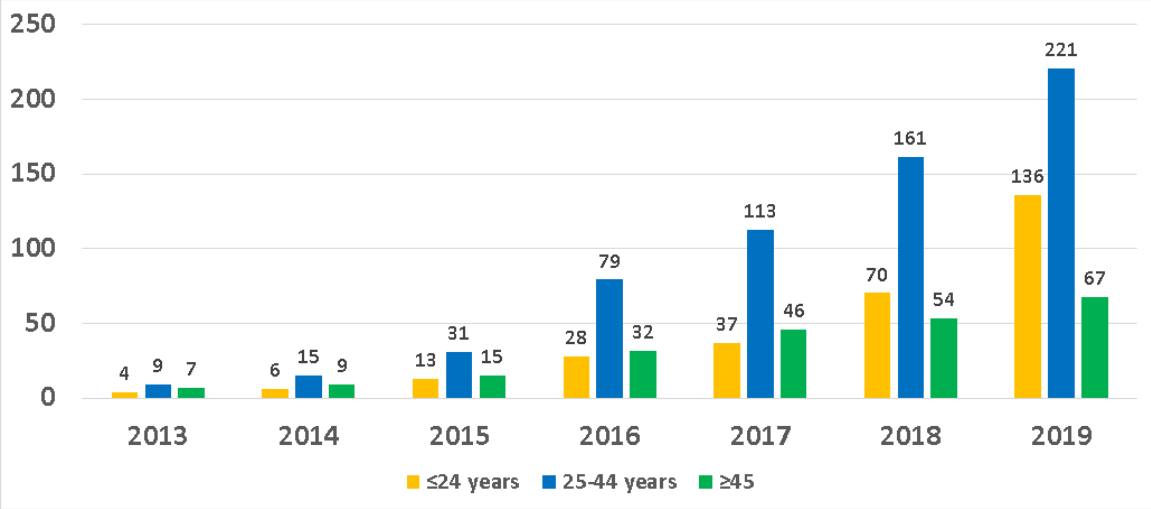


Figure 4.11: Population-adjusted PrEP patients in Mecklenburg County from 2013 to 2019 by age group

ZIP codes (not captured in the data), we cannot determine if these high counts in the group is due to the high number of pharmacies or PrEP users in the group. However, pharmacy ZIP codes is a reasonable approximation of residence as most people fill prescriptions close to their homes.

The PnR bounds established here can be used as a more comprehensive metric to monitor PrEP coverage. The PnR bounds can account for discontinuing prescriptions of PrEP and provides a range for PrEP needed to sufficiently protect individuals annually. This bound provides a sensitivity analysis of PrEP coverage in the county over a single metrics, such as PrEP prevalence and PnR.

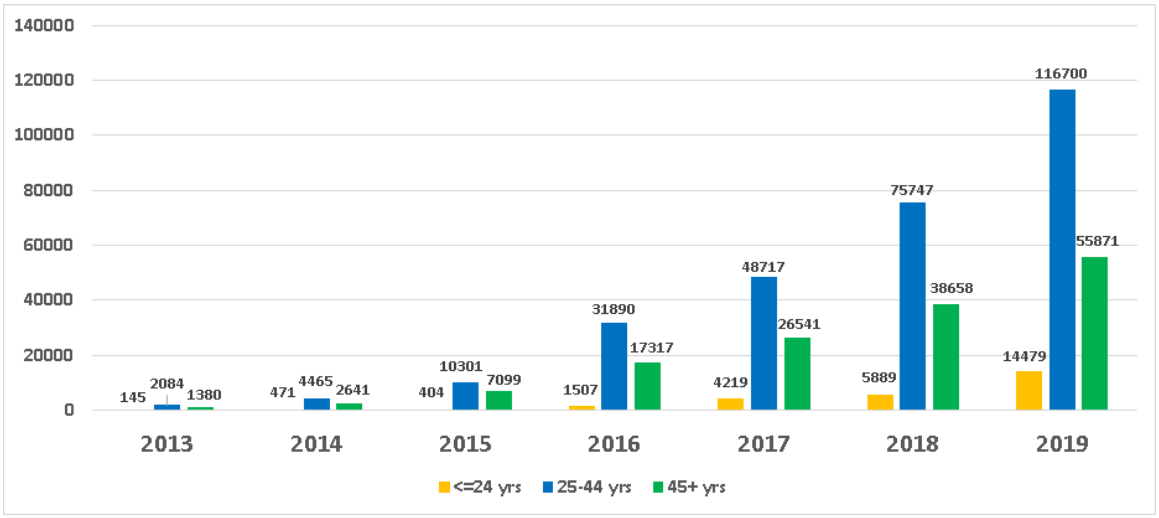


Figure 4.12: PrEP pills dispensed in Mecklenburg County by age group

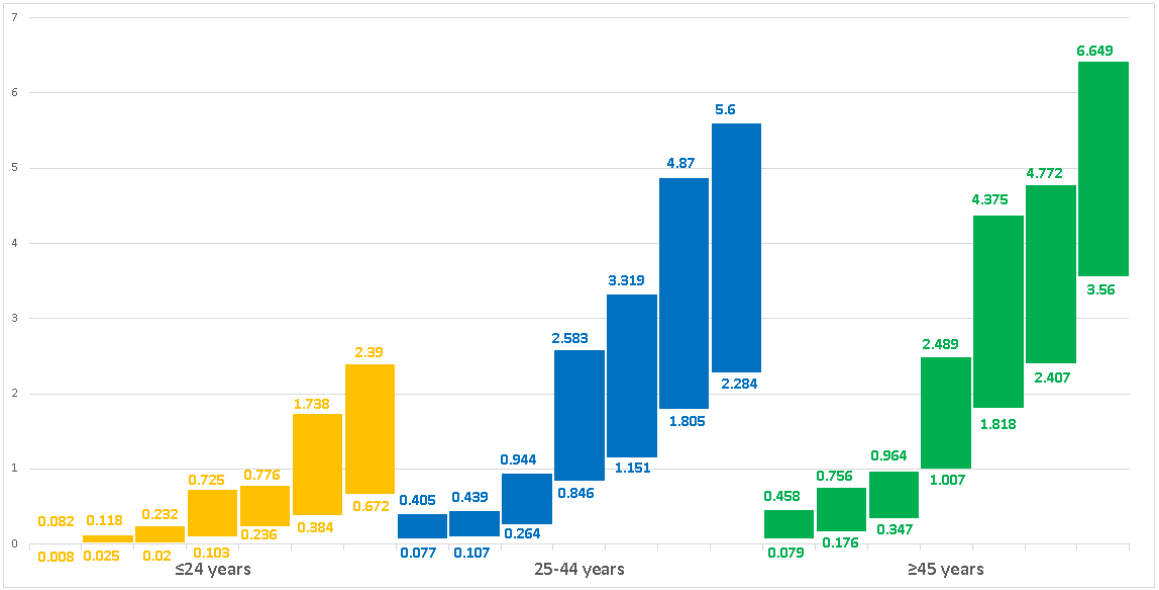


Figure 4.13: PrEP-to-need ratio bounds in Mecklenburg County from 2013 to 2019 by age group

Table 4.2: Population-adjusted PrEP patients by ZIP code groups

	2013	2014	2015	2016	2017	2018	2019
ZIP code group 0	4.94	9.65	9.44	23.12	32.71	61.73	87.02
ZIP code group 1	1.49	8.85	14.54	28.64	26.76	52.62	67.59
ZIP code group 2	9.77	8.69	19.59	33.47	39.55	44.72	58.49
ZIP code group 3	4.08	5.33	15.66	44.1	84.55	122.12	163.08
ZIP code group 4	11.32	8.82	10.75	37.77	47.18	96.37	144.64
ZIP code group 5	11.86	34.39	81.23	131.97	199.81	229.56	276.33
ZIP code group 6	32.96	12.79	37.15	167.48	213.43	427.71	717.07
ZIP code group 7	0	4.22	11.39	23.42	29.08	39.56	49.62
ZIP code group 8	1.74	5.04	6.5	9.42	18.19	46.82	75.36
ZIP code group 9	0	7.52	10.19	23.75	26.09	34.54	48.53
ZIP code group 10	4.59	4.5	6.61	32.48	40.49	38.86	55.79
ZIP code group 11	26.67	30.33	74.05	138.02	186.40	186.91	257.93
ZIP code group 12	4.16	6.13	3.01	25.6	42.56	43.74	52.99

Table 4.3: PrEP pills dispensed by ZIP code groups

	2013	2014	2015	2016	2017	2018	2019
ZIP code group 0	217	1032	1808	3310	5008	9196	15687
ZIP code group 1	30	553	953	2415	2140	3106	5835
ZIP code group 2	873	1220	2272	5140	6741	7430	9490
ZIP code group 3	553	649	2577	8001	17966	30699	46133
ZIP code group 4	249	346	528	1522	2599	6977	11829
ZIP code group 5	136	783	2880	5745	9759	10590	14012
ZIP code group 6	201	150	391	3231	4739	9170	22417
ZIP code group 7	0	377	1687	2915	4049	5565	7871
ZIP code group 8	26	278	322	1083	1381	3267	6175
ZIP code group 9	0	497	579	2368	4062	5532	8884
ZIP code group 10	153	150	274	3389	5236	6131	8204
ZIP code group 11	796	951	3056	9946	11217	15080	20139
ZIP code group 12	375	591	477	1649	4580	7551	10374

Table 4.4: PrEP-to-need ratio by ZIP code groups

	2013	2014	2015	2016	2017	2018	2019
ZIP code group 0	0.83	1.25	5	4.17	6	9.86	19.8
ZIP code group 1	0.06	0.32	0.5	1.05	1.12	2.71	2.27
ZIP code group 2	0.65	0.43	0.88	3.08	1.45	2.39	3.17
ZIP code group 3	0.14	0.27	0.46	2.3	2.93	4.52	6.35
ZIP code group 4	0.71	0.4	0.56	3	2.88	4	6.17
ZIP code group 5	0.6	1	1.83	1.85	5.8	7.67	5.8
ZIP code group 6	2.5	0.67	2	9.33	18.5	19.25	22.67
ZIP code group 7	0	2	2.2	7.67	5.8	10	3.64
ZIP code group 8	0.05	0.09	0.16	0.24	0.67	1.23	1.74
ZIP code group 9	0	0.38	0.73	2.36	2.23	4.33	2.67
ZIP code group 10	0.29	0.15	0.25	1.36	1.52	1.54	2.08
ZIP code group 11	1.57	1.3	2.75	3.37	10	7.83	17.13
ZIP code group 12	0.24	0.55	0.21	2	3.38	3.54	3.56

Table 4.5: Dose-adjusted PrEP-to-need ratio by ZIP code groups

	2013	2014	2015	2016	2017	2018	2019
ZIP code group 0	0.099	0.353	2.477	1.507	2.287	3.599	8.596
ZIP code group 1	0.005	0.08	0.131	0.347	0.345	0.608	0.727
ZIP code group 2	0.141	0.145	0.239	1.08	0.56	0.885	1.13
ZIP code group 3	0.036	0.059	0.136	0.729	1.07	1.912	2.939
ZIP code group 4	0.097	0.095	0.161	0.693	0.89	1.593	2.701
ZIP code group 5	0.075	0.238	0.658	0.785	2.674	3.224	2.56
ZIP code group 6	0.275	0.137	0.357	2.943	6.492	6.281	10.236
ZIP code group 7	0	0.516	0.924	2.655	2.219	3.812	1.54
ZIP code group 8	0.004	0.022	0.035	0.118	0.21	0.344	0.546
ZIP code group 9	0	0.065	0.106	0.588	0.856	1.684	1.159
ZIP code group 10	0.03	0.016	0.031	0.421	0.574	0.7	0.864
ZIP code group 11	0.312	0.261	0.698	1.43	3.415	3.443	6.897
ZIP code group 12	0.06	0.147	0.093	0.347	0.965	1.591	1.776

CHAPTER 5: SEGMENTED REGRESSION ANALYSIS

5.1 Introduction

The most commonly considered approach to assess and measure the effectiveness of an intervention is randomized controlled trials (RCTs) [56, 57, 58]. However, conducting RCTs for all interventions may not be feasible [56]. RCTs can sometimes be prohibitively expensive to set up, and in addition, are susceptible to systematic error when generalizing results to "real world" settings leading to biased estimates [56]. In contrast, observational studies can address these shortcomings, however, the difficulty in establishing causation and the lack of control over confounding variables may lead to weaker insights [59].

Quasi-experimental study designs are proven to estimate the causal effects of an intervention using observational approaches. The quasi-experimental study designs are beneficial for analyzing observational data where full randomization, or a case-control design [60], is not affordable or possible. These designs make full use of the longitudinal nature of the data. Interrupted time series (ITS) analysis is one of the useful quasi-experimental designs to evaluate the longitudinal effects of interventions using regression modeling. Segmented regression analysis of interrupted time series data has proven to be able to quantify the effect of an intervention in statistical terms. This analysis also allows measuring the change in the outcome of interest due to an intervention immediately or over time [61].

In this chapter, we quantify the effect of the G2Z–MC intervention introduced in July 2018 on the number of unique PrEP patients in a month. The number of unique patients on PrEP in a month is defined as anyone with a prescription for PrEP as identified in Chapter 4 in that month. The plot of monthly PrEP patients from 2013

to 2020 is shown in Figure 5.1.

We analyze three interruption breakpoints in the monthly PrEP patients time series, namely:

1. **ramp-up interruption:** The first interruption breakpoint is in January 2016, referred to as the Ramp-up interruption. We see a slow rise in PrEP patients at the start from January 2013 to December 2015, which we suspect to be the initial ramp-up phase of users seen in most medications when first introduced into the marketplace.

2. **G2Z:** The second interruption breakpoint of interest in the time series is the Getting-to-Zero-MC (G2Z) intervention introduced in July 2018. The major objective of G2Z–MC is to increase PrEP usage in the county, and this point of interruption in the time series helps us determine the immediate and gradual effects caused by G2Z on PrEP usage in the county.

3. **COVID-19:** The third interruption breakpoint of interest is the restrictions introduced because of the COVID-19 epidemic in the county. We consider March 2020 as the interruption breakpoint since the COVID-19 cases were first recognized and Public Health Interventions (PHIs) were first introduced in Mecklenburg County beginning in March 2020. We see the trend in PrEP users decrease around and after the onset of the COVID-19 epidemic on visual inspection of the monthly PrEP patients plot in Figure 5.1. Moreover, studies have found that the COVID-19 pandemic disrupted access to and delivery of healthcare services, including HIV prevention [62, 63]. Hence, we use this interruption breakpoint to not have the effect of decreasing PrEP patients caused by an external factor like the COVID-19 pandemic impact our results quantifying the impact of G2Z.

In this chapter, we first conduct segmented regression analysis on the number of monthly PrEP patients in the county from 2013 to 2020 using a single segmented

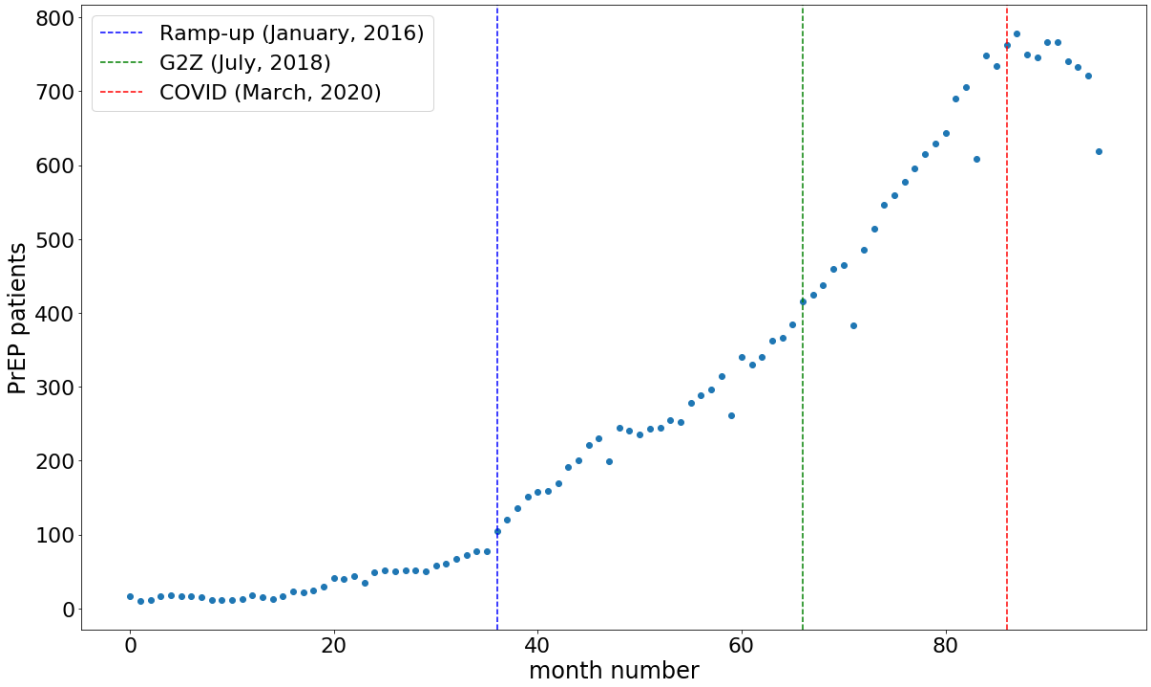


Figure 5.1: Monthly PrEP patients in Mecklenburg County from 2013 to 2020 with three interruption breakpoints points

regression model (Single model). The Single model is used to compute the immediate and gradual effects caused by the above three interruption breakpoints on the PrEP patient counts using just one model. We then use three pairwise segmented regression models to do the same in pairs of pre-interruption and post-interruption periods. We empirically show that both methods yield similar results. We then use the Single model to analyze the immediate and gradual effects of G2Z-MC to provide actionable nsights on its impact on PrEP usage in the county.

5.2 Methods

5.2.1 Segmented linear regression using a single model

5.2.1.1 Model formulation

The formulation for the full single segmented regression model with three interruption breakpoints and four segments can be written as shown in Equation (5.1).

$$
\begin{aligned}
Y_t = &\beta_0 + \beta_1 * (\text{months since 2013}) \\
&+ \beta_2 * \text{ramp-up} + \beta_3 * (\text{months since ramp-up}) * \text{ramp-up} \\
&+ \beta_4 * \text{G2Z} + \beta_5 * (\text{months since G2Z}) * \text{G2Z} \\
&+ \beta_6 * \text{COVID-19} + \beta_7 * (\text{months since COVID-19}) * \text{COVID-19}
\end{aligned} \tag{5.1}
$$

where,

$\boldsymbol{Y_t}$:	is the predicted number of monthly PrEP patients
months since 2013:	is the number of months since the start of the study
ramp-up:	is the indicator variable which is 0 before and 1 after ramp-up
months since ramp-up:	is the number of months since the ramp-up
G2Z:	is the indicator variable which is 0 before and 1 after G2Z
months since G2Z:	is the number of months since the G2Z
COVID-19:	is the indicator variable which is 0 before and 1 after COVID-19
months since COVID-19:	is the number of months since the COVID-19
$\boldsymbol{\beta_0}$:	is the baseline intercept (at January 2013)
$\boldsymbol{\beta_1}$:	is the baseline slope
$\boldsymbol{\beta_2}$:	is the change in intercept at ramp-up
$\boldsymbol{\beta_3}$:	is the change in slope after ramp-up
$\boldsymbol{\beta_4}$:	is the change in intercept at G2Z
$\boldsymbol{\beta_5}$:	is the change in slope after G2Z
$\boldsymbol{\beta_6}$:	is the change in intercept at COVID-19
$\boldsymbol{\beta_7}$:	is the change in slope after COVID-19.

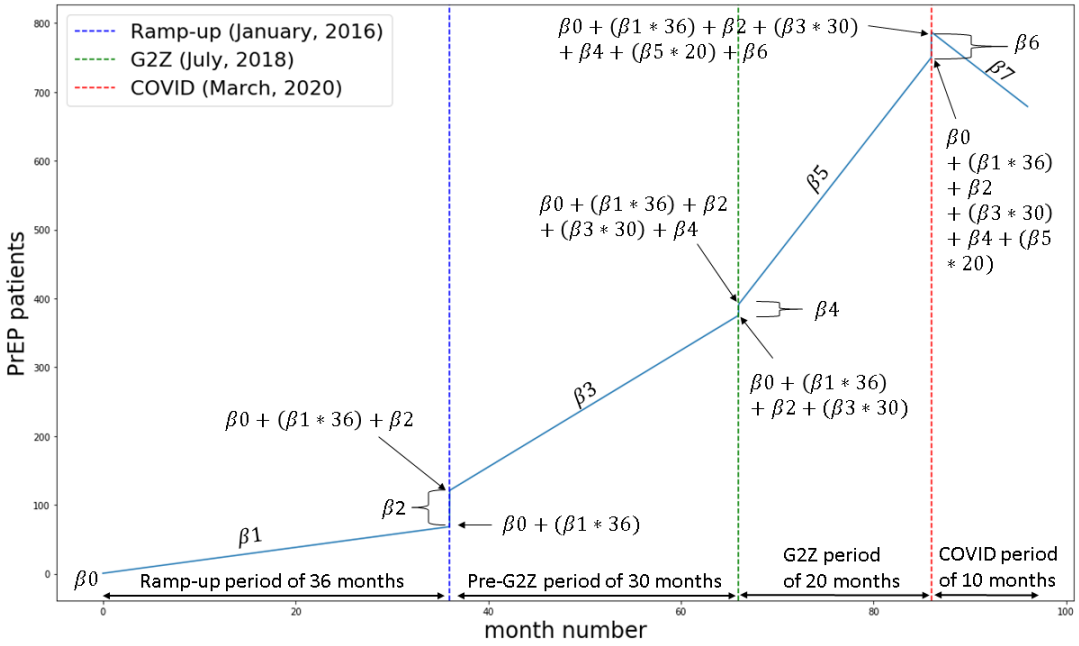


Figure 5.2: Illustration of the intercept and slope interpretation at three interruption breakpoints and four segments (periods) using a single segmented regression model

Ramp-up segment (period)

At the ramp-up period (January 2013 to December 2015), the indicator variables, ramp-up, G2Z, and COVID-19, are equal to zero. Hence Equation (5.1) will take the form of:

$$Y_t = \beta_0 + \beta_1 * (\text{months since 2013}), \tag{5.2}$$

where β_0 is the baseline intercept, and β_1 is the baseline slope in the ramp-up period. Hence, we can calculate the endpoint of the ramp-up line as $\beta_0 + (\beta_1 * 36)$, where 36 is the number of months in the ramp-up period. This endpoint gives the level of PrEP patients right before the ramp-up interruption (i.e., the intercept of the ramp-up regression line at the ramp-up interruption).

Pre-G2Z segment (period)

At the Pre-G2Z period, which is from January 2016 to June 2018, the indicator variable ramp-up is set to one, and the G2Z and COVID-19 variables are set to zero. Hence the regression Equation (5.1) takes the form of:

$$\begin{aligned} Y_t =& \beta_0 + \beta_1 * (\text{months since 2013}) \\ &+ \beta_2 * \text{ramp-up} + \beta_3 * (\text{months since ramp-up}) * \text{ramp-up}. \end{aligned} \tag{5.3}$$

The jump in the intercept at the ramp-up interruption (i.e., the immediate effect of the ramp-up interruption) will be given by β_2. Thus, this intercept can be calculated as $\beta_0 + (\beta_1 * 36) + \beta_2$, where $\beta_0 + (\beta_1 * 36)$ is the endpoint of the ramp-up period.

β_3 gives the change in slope after the ramp-up interruption, and β_1 gives the slope in the ramp-up period. Therefore, we can calculate the total slope in the pre-G2Z period as $\beta_1 + \beta_3$. The endpoint of the pre-G2Z period can be calculated as $\beta_0 + (\beta_1 * 36) + \beta_2 + (\beta_1 + \beta_3) * 30$, where 30 is the number of months in the pre-G2Z period. This endpoint is the level of PrEP patients right before the G2Z interruption (i.e., the intercept of the pre-G2Z regression line at the G2Z interruption).

G2Z segment (period)

At the G2Z period which starts from July 2018 and ends in February 2020. The indicator variables ramp-up and G2Z are both equal to one and COVID-19 is set to zero. Thus, the regression Equation (5.1) becomes:

$$\begin{aligned} Y_t =& \beta_0 + \beta_1 * (\text{months since 2013}) \\ &+ \beta_2 * \text{ramp-up} + \beta_3 * (\text{months since ramp-up}) * \text{ramp-up} \\ &+ \beta_4 * \text{G2Z} + \beta_5 * (\text{months since G2Z}) * \text{G2Z} \end{aligned} \tag{5.4}$$

β_4 is the jump in the intercept or the immediate effect at the G2Z interruption. The level at G2Z after the jump can be calculated as the sum of the endpoint of the

pre-G2Z period and β_4, as $\beta_0 + (\beta_1 * 36) + \beta_2 + (\beta_1 + \beta_3) * 30 + \beta_4$. β_5 gives the change in slope due to G2Z interruption and the total slope for the G2Z period is $\beta_1 + \beta_3 + \beta_5$. The endpoint of the G2Z period can be calculated as $\beta_0 + (\beta_1 * 36) + \beta_2 + (\beta_1 + \beta_3) * 30 + \beta_4 + (\beta_1 + \beta_3 + \beta_5) * 20$, where 20 is the number of months in the G2Z period.

COVID-19 segment (period)

All the indicator variables (i.e., ramp-up, G2Z, and COVID-19) are equal to one. Therefore the regression Equation (5.1) takes the form of:

$$\begin{aligned} Y_t = &\beta_0 + \beta_1 * (\text{months since 2013}) \\ &+ \beta_2 * \text{ramp-up} + \beta_3 * (\text{months since ramp-up}) * \text{ramp-up} \\ &+ \beta_4 * \text{G2Z} + \beta_5 * (\text{months since G2Z}) * \text{G2Z} \\ &+ \beta_6 * \text{COVID-19} + \beta_7 * (\text{months since COVID-19}) * \text{COVID-19} \end{aligned} \tag{5.5}$$

β_6 gives the immediate effect of the COVID-19 interruption. Hence, the absolute intercept after accounting the immediate effect of COVID-19 is given by $\beta_0 + (\beta_1 * 36) + \beta_2 + (\beta_1 + \beta_3) * 30 + \beta_4 + (\beta_1 + \beta_3 + \beta_5) * 20 + \beta_6$, where 20 is the duration of the COVID-19 period. β_7 gives the change in slope due to COVID-19 interruption. Hence we can compute the total slope in the COVID-19 period as $\beta_1 + \beta_3 + \beta_5 + \beta_7$.

5.2.2 Pairwise linear regression

The Pairwise Segmented Linear Regression analysis is conducted by looking at the three interruptions one at a time and building three separate models to compute the effects of the above three interruptions. These three model takes the form of Equation (5.6) for each interruption.

$$Y_t = \beta_0 + \beta_1 * x + \beta_2 * \text{interruption} + \beta_3 * (\text{time since interruption}), \tag{5.6}$$

where, *interruption* is the indicator variable (which is set to zero before and one after the interruption).

For example, at the ramp-up interruption Equation (5.6) takes the form as:

$$\begin{aligned} Y_t = &\beta_0 + \beta_1 * \text{months since 2013} + \beta_3 * \text{ramp-up} \\ &+ \beta_4 * (\text{months since ramp-up}) * \text{ramp-up} \end{aligned} \tag{5.7}$$

5.2.2.1 Computing intercept and slope

From Equation (5.6), β_0 gives the baseline intercept at the beginning of the pre-interruption period. β_1 is the slope of the pre-interruption period. The change in the intercept or the immediate effect of the interruption is given by β_3 and the change in slope after the interruption is given by β_4. Hence, similar to the Single model, we can compute the endpoint of the pre-interruption period as $\beta_1 *$ months in the pre-intervention period, and the total slope in the post-interruption period is $\beta_1 + \beta_3$. The same analyses can be repeated at the three different interruption breakpoints using three different models.

5.3 Results

5.3.1 Segmented linear regression using a single model

The plot from the full single segmented regression model is shown in Figure 5.3 and the coefficients are shown in Table 5.2.

From Table 5.2 we see that two coefficients: "Baseline intercept" and "G2Z," are not statistically significant in the first Single model. We perform backward selection and eliminate these statistically not significant variables one by one. The "Baseline intercept" having the highest p-value of 0.921 is eliminated first, and the resulting model is shown in Equation (5.8).

The plot and coefficients after eliminating the "Baseline intercept" variable are shown in Figure 5.4 and Table 5.3 respectively.

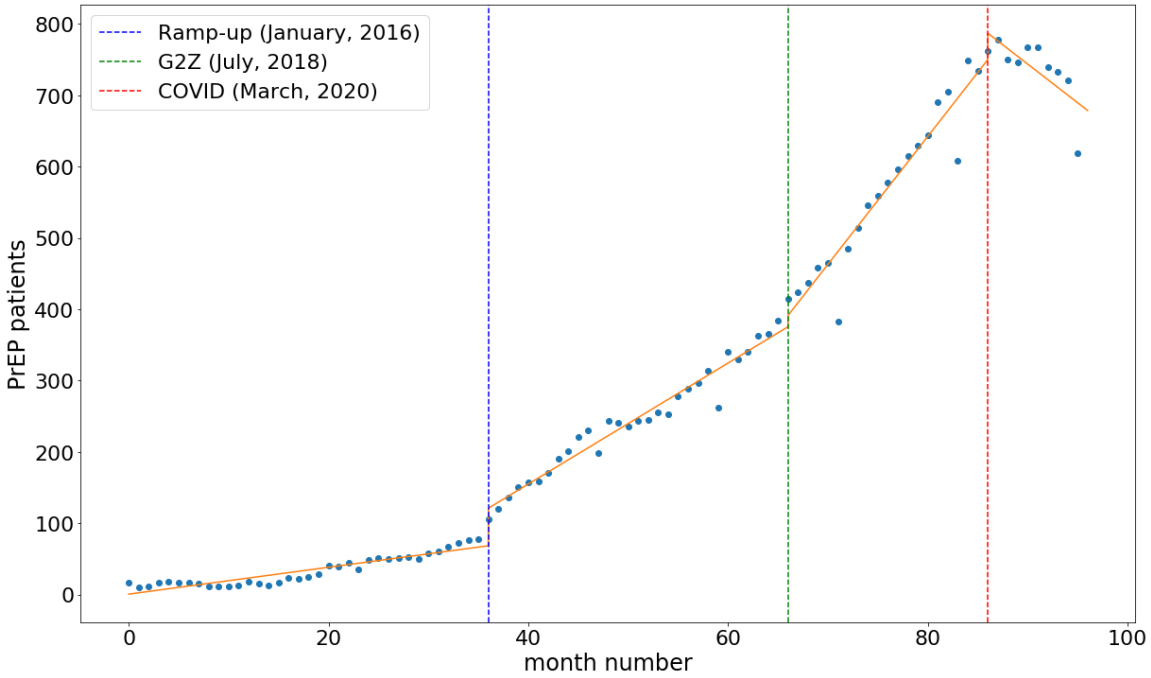


Figure 5.3: Full segmented regression model plot

$$\begin{aligned} Y_t = &\beta_1 * \text{months since 2013} \\ &+ \beta_2 * \text{ramp-up} + \beta_3 * (\text{months since ramp-up}) * \text{ramp-up} \\ &+ \beta_4 * \text{G2Z} + \beta_5 * (\text{months since G2Z}) * \text{G2Z} \\ &+ \beta_6 * \text{COVID-19} + \beta_7 * (\text{months since ramp-up}) * \text{COVID-19}. \end{aligned} \tag{5.8}$$

From Table 5.3, we can see that the coefficient for "G2Z" is not statistically significant with a p-value of 0.193 and is eliminated next. The resulting model after eliminating "G2Z" is shown in Equation (5.9).

Table 5.2: Full segmented regression model table

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
Baseline intercept	0.6952	6.955	0.1	0.921	-13.126	14.517
months since 2013	1.8825	0.342	5.508	0.000	1.203	2.562
ramp-up	52.6403	10.495	5.016	0.000	31.783	73.497
months since ramp-up	6.5976	0.565	11.687	0.000	5.476	7.719
G2Z	15.8491	12.161	1.303	0.196	-8.319	40.017
months since G2Z	9.456	0.94	10.056	0.000	7.587	11.325
COVID-19	37.012	15.958	2.319	0.023	5.299	68.725
months since COVID-19	-28.7785	2.486	-11.574	0.000	-33.72	-23.837

Table 5.3: Single Model with "Baseline intercept" eliminated

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
months since 2013	1.9119	0.173	11.021	0.000	1.567	2.257
ramp-up	52.2780	9.794	5.337	0.000	32.817	71.739
months since ramp-up	6.5682	0.479	13.704	0.000	5.616	7.521
G2Z	15.8491	12.093	1.311	0.193	-8.180	39.878
months since G2Z	9.456	0.935	10.113	0.000	7.598	11.314
COVID-19	37.012	15.869	2.332	0.022	5.481	68.543
months since COVID-19	-28.7785	2.473	-11.639	0.000	-33.691	-23.866

$$\begin{aligned} Y_t =& \beta_1 * (\text{months since 2013}) \\ &+ \beta_2 * \text{ramp-up} + \beta_3 * (\text{months since ramp-up}) * \text{ramp-up} \\ &+ \beta_5 * (\text{months since G2Z}) * \text{G2Z} \\ &+ \beta_6 * \text{COVID-19} + \beta_7 * (\text{months since COVID-19}) * \text{COVID-19}. \end{aligned} \tag{5.9}$$

The plot and coefficients after eliminating both "Baseline intercept" and "G2Z" are shown in Figure 5.5 and Table 5.4 respectively.

From Table 5.4, we can see that all the coefficients in the above model are statistically significant. The number of monthly PrEP users kept increasing at a rate of $\beta_1 = 1.9119$ PrEP patients per month in the ramp-up period. The duration of the ramp-up period is for 36 months. Hence, we can calculate the endpoint or

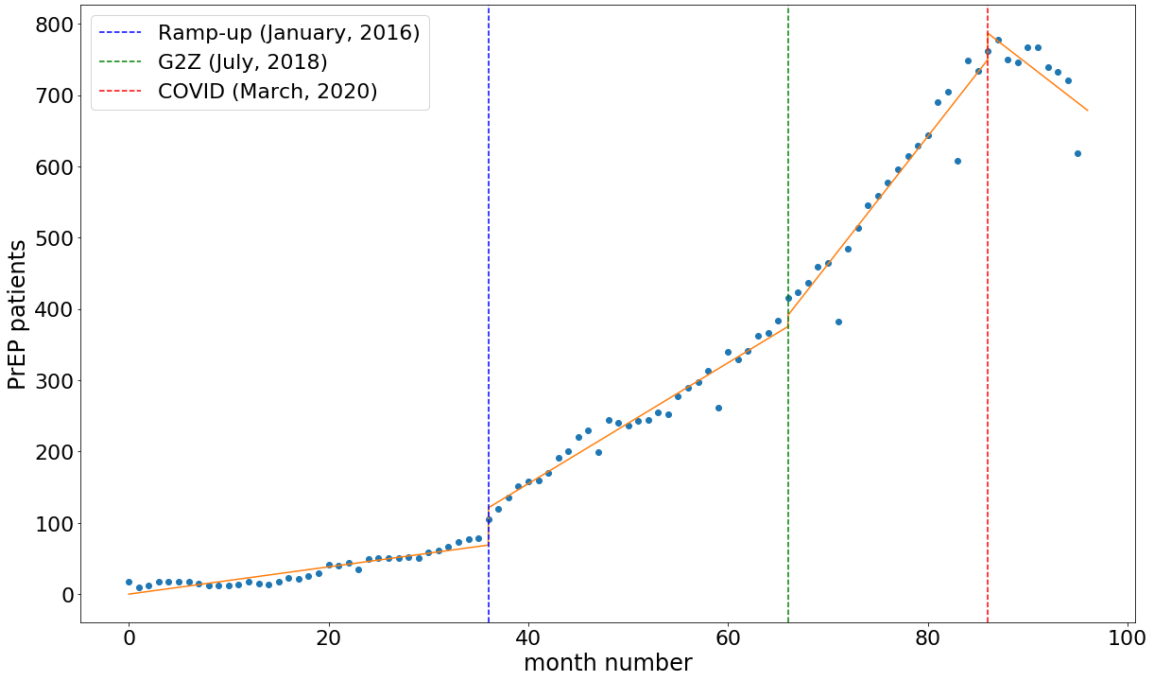


Figure 5.4: Segmented regression model plot with "Baseline intercept" eliminated

the baseline level of the ramp-up period as $1.9119 * 36 = 68.8284$ monthly PrEP patients. The immediate effect or the change in level due to the ramp-up interruption is given by $\beta_2 = 49.0363$ PrEP patients. Therefore, the absolute intercept after accounting the jump caused by the ramp-up interruption is given by; $(1.9119 * 36) + 49.0363 = 117.8647$.

The gradual effect of the ramp-up interruption or the slope after the ramp-up interruption increased by $\beta_3 = 6.9036$ PrEP patients per month. Therefore, the slope in the pre-G2Z period is given by the sum of the slope in the ramp-up period and the change in slope due to the ramp-up interruption, $\beta_1 + \beta_3 = 1.9119 + 6.9036 = 8.8155$.

In the pre-G2Z period, the number of PrEP patients kept increasing at this rate of 8.8155 PrEP patients per month. The duration of the pre-G2Z period is 30 months. Hence, the endpoint of the pre-G2Z period is $117.8647 + (8.8155 * 30) = 382.3297$. The immediate effect or the level change at G2Z interruption is not statistically significant and is eliminated. Hence, the level after the immediate effect of G2Z interruption is

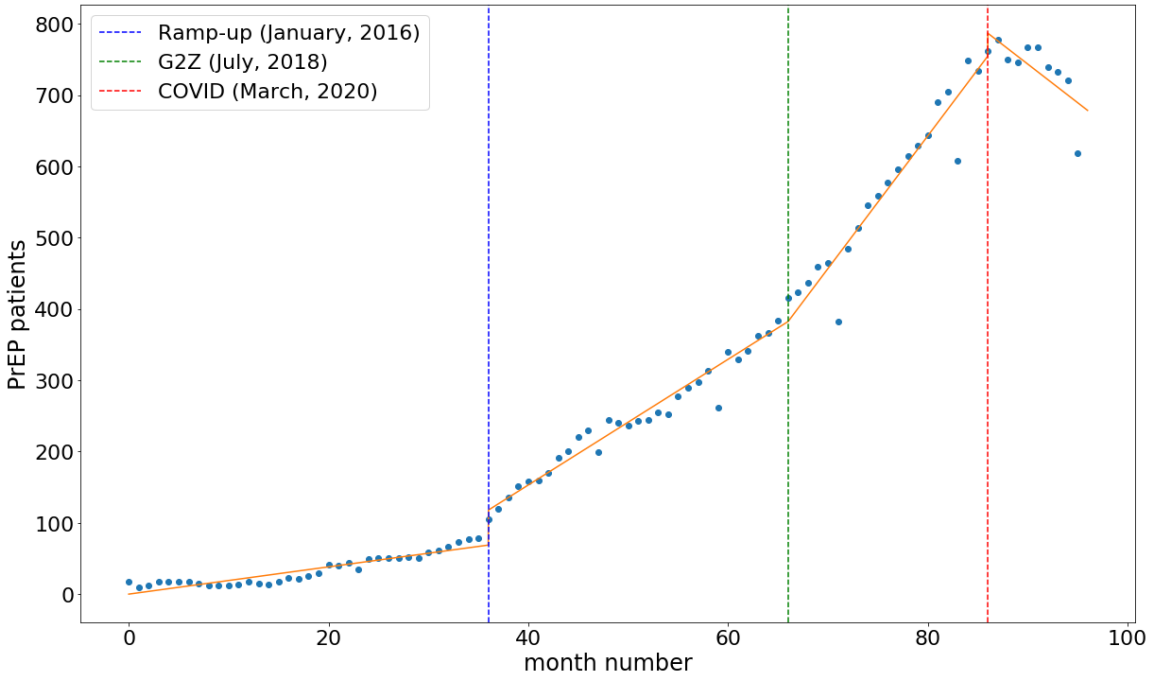


Figure 5.5: Segmented regression plot with "Baseline intercept" and "G2Z" eliminated

382.3297+0 = 382.3297. The change in slope or the gradual effect of G2Z interruption is $\beta_5 = 9.8153$ and the slope in the G2Z period is $\beta_1 + \beta_3 + \beta_5 = 1.9119 + 6.9036 + 9.8153 = 18.6308$. This implies that the number of monthly PrEP patients increased at a rate of 18.6308 for a duration of 20 months until the COVID-19 interruption. Therefore, the endpoint in the G2Z period is $382.3297 + (18.6308 * 20) = 754.9457$. The immediate effect or level change due to the COVID-19 interruption is given by $\beta_6 = 32.1494$ PrEP patients and the level of PrEP patients after the immediate effect of COVID-19 interruption is $754.9457 + 32.1494 = 787.0951$. The slope change or the gradual effect of COVID-19 interruption is -29.4732 PrEP patients per month. Hence the total slope in the COVID-19 period is $18.6308 - 29.4732 = -10.8424$, i.e., the number of PrEP patients kept reducing at a rate of -10.8424 PrEP patients every month after the COVID-19 interruption.

The immediate effect (jump in the intercept at interruption breakpoints) and the gradual effect (change in the slope after interruption breakpoints) for all the three

Table 5.4: Single Model with "Baseline intercept" and "G2Z" eliminated

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
months since 2013	1.9119	0.174	10.977	0.000	1.566	2.258
ramp-up	49.0363	9.515	5.154	0.000	30.134	67.939
months since ramp-up	6.9036	0.407	16.966	0.000	6.095	7.712
months since G2Z	9.8153	0.898	10.936	0.000	8.032	11.598
COVID-19	32.1494	15.491	2.075	0.041	1.375	62.924
months since COVID-19	-29.4732	2.425	-12.156	0.000	-34.290	-24.656

interruption breakpoints are shown in Table 5.5 and Table 5.6, respectively.

Table 5.5: Change in intercept (pre-interruption vs. post-interruption)

Period	Pre-interruption intercept	Jump in intercept	Post-interruption intercept
ramp-up	68.8284	49.0363	117.8647
G2Z	382.3297	0	382.3297
COVID-19	754.9457	32.1494	787.0951

Table 5.6: Change in slope (pre-interruption vs. post-interruption)

Era	Pre-interruption slope	Change in slope	Post-interruption slope
ramp-up	1.9119	6.9036	8.8155
G2Z	8.8155	9.8153	18.6308
COVID-19	18.6308	-29.4732	10.8424

5.3.2 Pairwise linear regression

5.3.2.1 Ramp-up vs. pre-G2Z period

From Table 5.7, we see that the "Intercept" is not statistically significant with a p-value of 0.862 and is eliminated.

Figure 5.7 and Table 5.8 is the plot and summary table after eliminating the "Intercept". The monthly PrEP patients increased at the rate of 1.9119 PrEP patients per month in the ramp-up period of 36 months. Therefore, the endpoint of the ramp-up period is $1.9119 * 36 = 68.8284$. The immediate effect of the ramp-up is 52.2780, PrEP

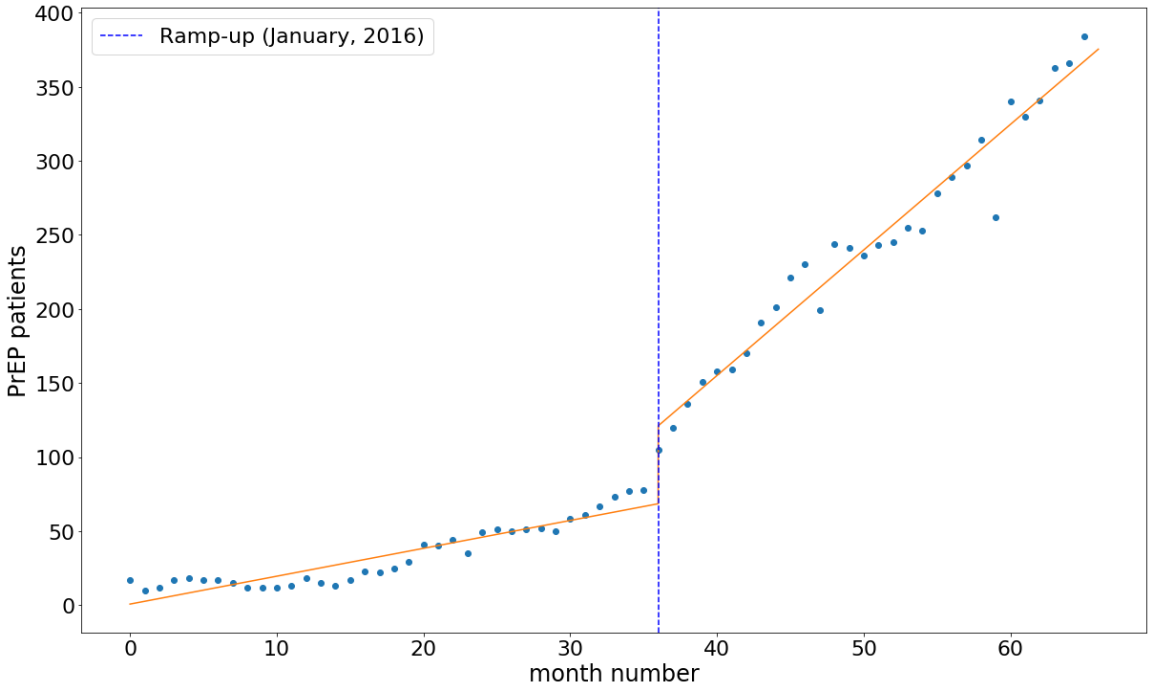


Figure 5.6: Full pairwise regression plot (pre-ramp-up vs. post-ramp-up)

Table 5.7: Full pairwise regression table (ramp-up vs. pre-G2Z)

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
Intercept	0.6952	3.987	0.174	0.862	-7.275	8.666
months since 2013	1.8825	0.196	9.608	0.000	1.491	2.274
ramp-up	52.6403	6.017	8.749	0.000	40.613	64.668
months since ramp-up	6.5976	0.324	20.385	0.000	5.951	7.245

patients. The gradual effect or the change in slope after the ramp-up interruption is 6.5682, and the total slope in the pre-G2Z period is $1.9119 + 6.5682 = 8.4801$.

The coefficients in Table 5.8 are exactly the same as the coefficients from the single model shown in Table 5.4.

5.3.2.2 Pre-G2Z vs. G2Z period

The plot and summary table for the pre-G2Z vs. G2Z period are shown in Figure 5.8 and Table 5.9. We see the coefficient for "G2Z" is not statistically significant with a p-value of 0.274 and is eliminated from the next model.

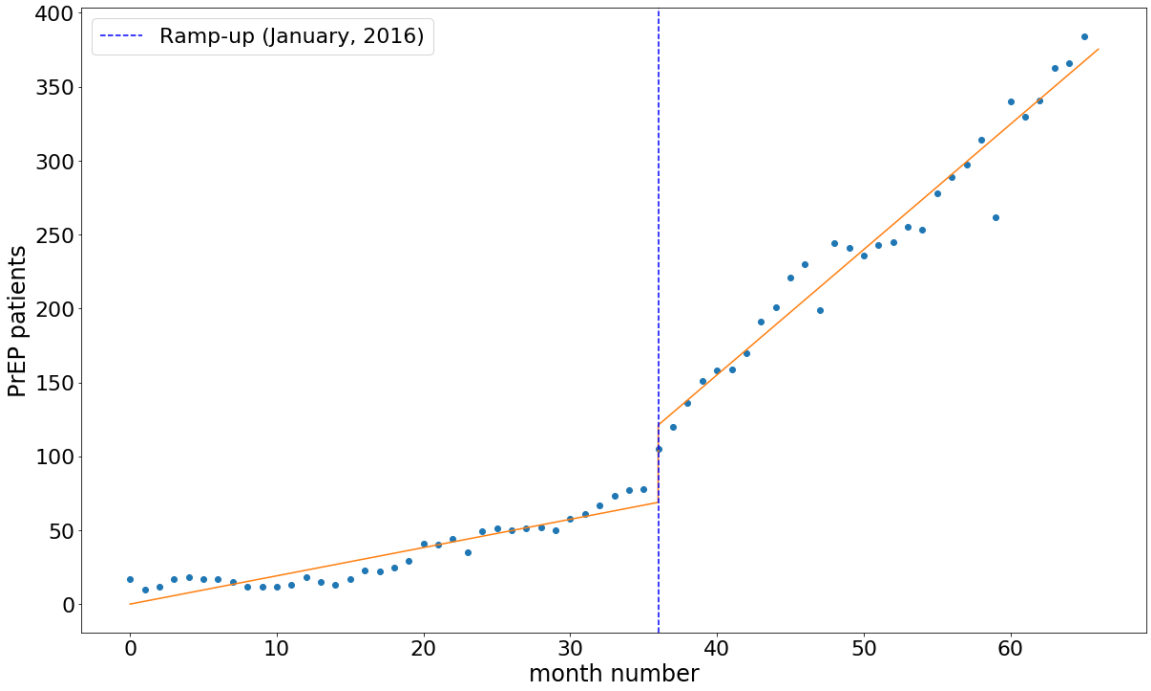


Figure 5.7: Pairwise regression plot (ramp-up vs. pre-G2Z) with "Intercept" eliminated

Table 5.8: Pairwise regression table (ramp-up vs. pre-G2Z) with "Intercept" eliminated

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
months since 2013	1.9119	0.099	19.265	0.000	1.714	2.11
ramp-up	52.2780	5.603	9.33	0.000	41.081	63.475
months since ramp-up	6.5682	0.274	23.954	0.000	6.02	7.116

Figure 5.9 and Table 5.10 show the plots and the summary table for the pre-G2Z vs. G2Z period after eliminating "G2Z". We see all coefficients are significant here. The baseline intercept is 117.8637 PrEP patients which is the endpoint of the ramp-up period in the single model as shown in Table 5.4.

The endpoint of pre-G2Z period is $117.8637 + (8.8154 * 30) = 382.3257$, which is the same as the endpoint computed in the single model from Table 5.4 $((1.9119 * 36) + 49.0363 + (1.9119 + 6.9036) * 30)$. There is no statistically significant immediate effect of G2Z interruption on monthly PrEP patients. However, the gradual effect or rate

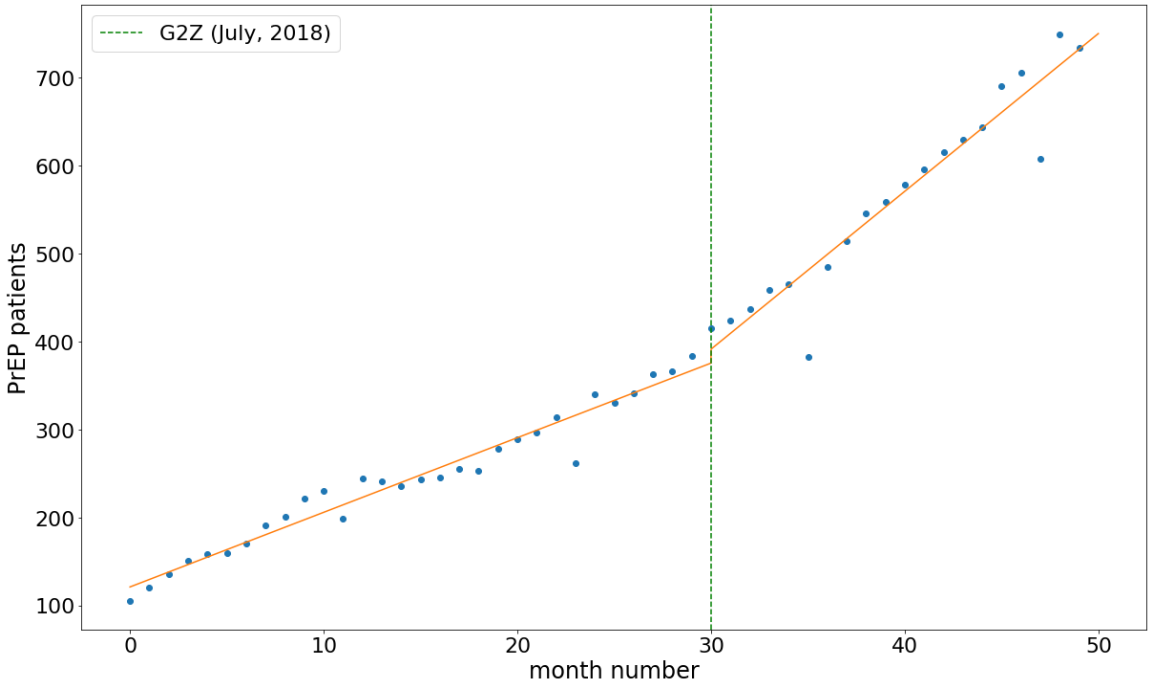


Figure 5.8: Full pairwise regression plot (pre-G2Z vs. G2Z)

Table 5.9: Full pairwise regression table (pre-G2Z vs. G2Z)

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
ramp-up	121.1054	8.926	13.567	0.000	103.138	139.073
months since ramp-up	8.4801	0.529	16.043	0.000	7.416	9.544
G2Z	15.8491	14.307	1.108	0.274	-12.949	44.647
months since G2Z	9.456	1.106	8.548	0.000	7.229	11.683

Full pairwise regression table (pre-G2Z vs. G2Z)

of increase in PrEP patients per month in the G2Z period is 9.8153. Hence, the total slope after G2Z interruption is $8.8154 + 9.8153 = 18.6307$, which is equal to the total slope computed by the Single model shown in Table 5.4 ($1.9119 + 6.9036 + 9.815 = 18.6305$).

5.3.2.3 G2Z vs. COVID-19 period

The plot and summary table for the G2Z vs. COVID-19 period is shown in Figure 5.10 and Table 5.11 respectively. From the Table 5.11, we see that the coefficient for "COVID-19" is not statistically significant (p-value = 0.162) and is eliminated in

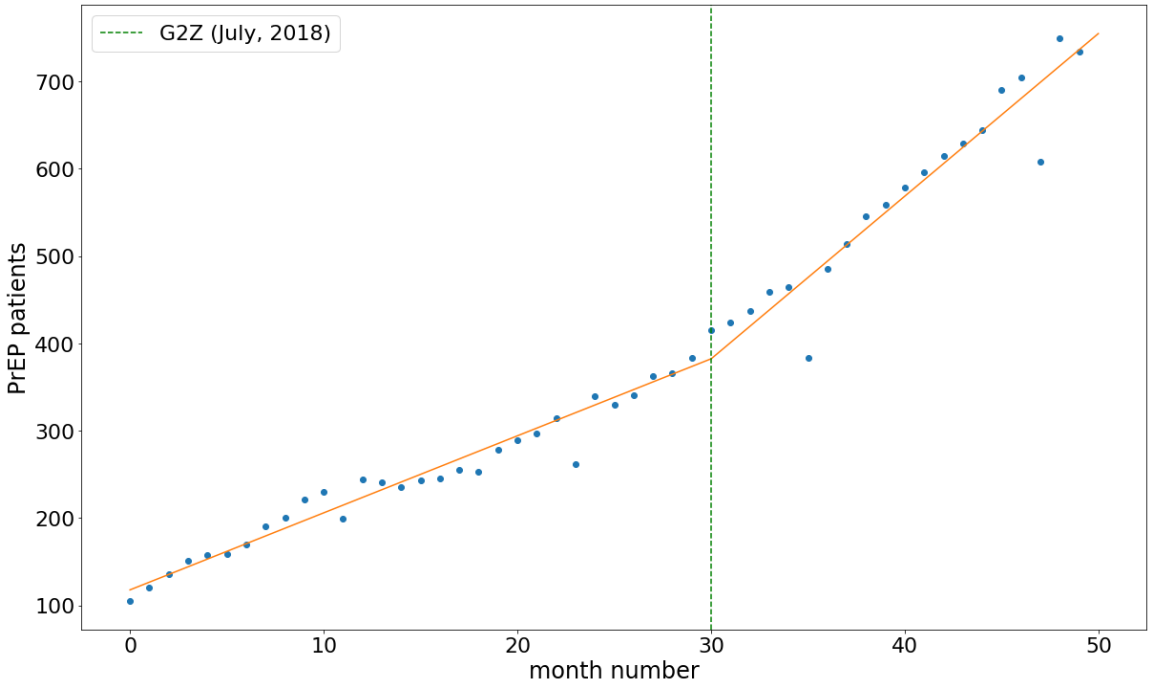


Figure 5.9: Pairwise regression plot (pre-G2Z vs. G2Z) with "G2Z" eliminated

Table 5.10: Pairwise regression table (pre-G2Z vs. post-G2Z) with "G2Z" eliminated

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
ramp-up	117.8637	8.453	13.943	0.000	100.858	134.87
months since ramp-up	8.8154	0.434	20.294	0.000	7.942	9.689
months since G2Z	9.8153	1.060	9.258	0.000	7.683	11.948

the next model.

The baseline intercept in the G2Z period is 384.7625. The rate of PrEP patients increases at 18.9773 patients per month in the G2Z period. There was no significant level change due to COVID-19 interruption. But, there was a gradual negative impact with the change in the trend of -26.227 PrEP patients every month. The level at COVID-19 interruption can be calculated as $384.7625 + (18.9773 * 20) = 764.3085$. The total slope in the COVID-19 period can be calculated as $18.9773 + (-26.2227) = -7.2454$. Since this pairwise analysis was not started from time 0 (January 2013), the intercepts and slopes do not match with the Single model, where intercept at the G2Z

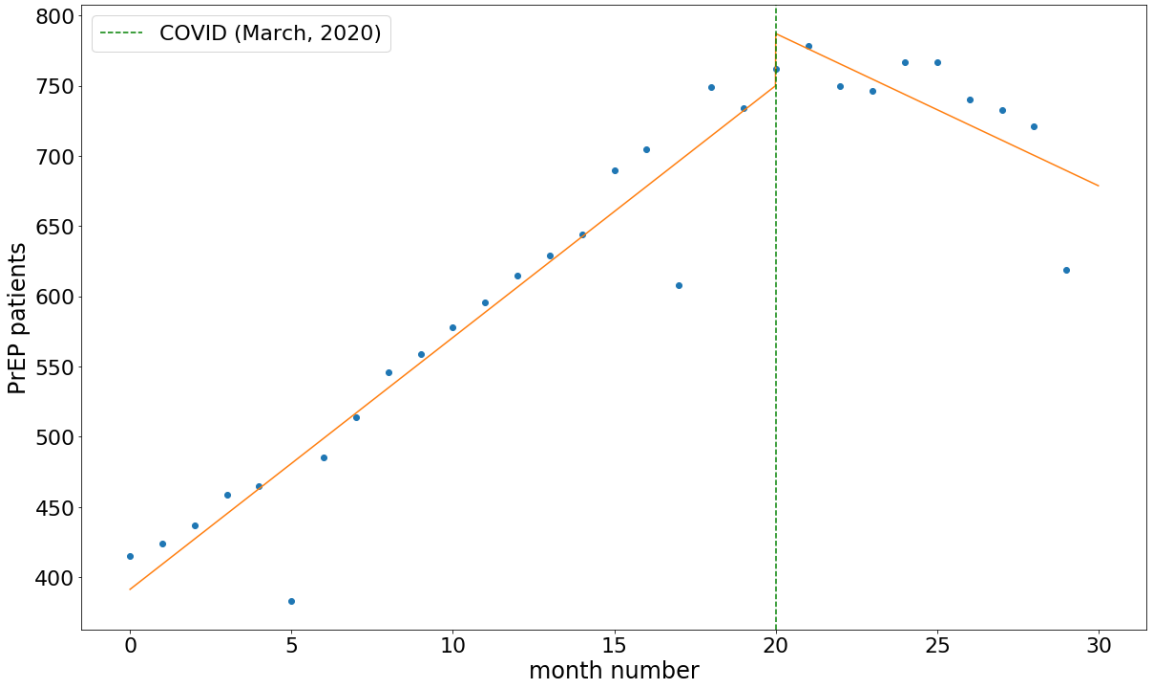


Figure 5.10: Full pairwise regression plot (G2Z vs. COVID-19 period)

Table 5.11: Full pairwise regression table (G2Z vs. COVID-19 period)

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
G2Z	391.3571	14.804	26.436	0.000	360.927	421.787
months since G2Z	17.9361	1.332	13.464	0.000	15.198	20.674
COVID-19	37.012	25.736	1.438	0.162	-15.888	89.912
months since COVID-19	-28.7785	4.01	-7.177	0.000	-37.021	-20.536

interruption breakpoint is offset by 382.3257 patients. (The endpoint of the pre-G2Z period and the starting point of the G2Z period coincide as the jump at the G2Z interruption breakpoint is not significant). Removing the non-significant variable "COVID-19" influences the slopes in the COVID-19 period. Since the intercept at "COVID-19" was significant in the single model and due to the offset in the baseline intercept, we see differences in the significance of "COVID-19" from the single model. However, these differences in the coefficients are not statistically significant from the ones computed in the Single model.

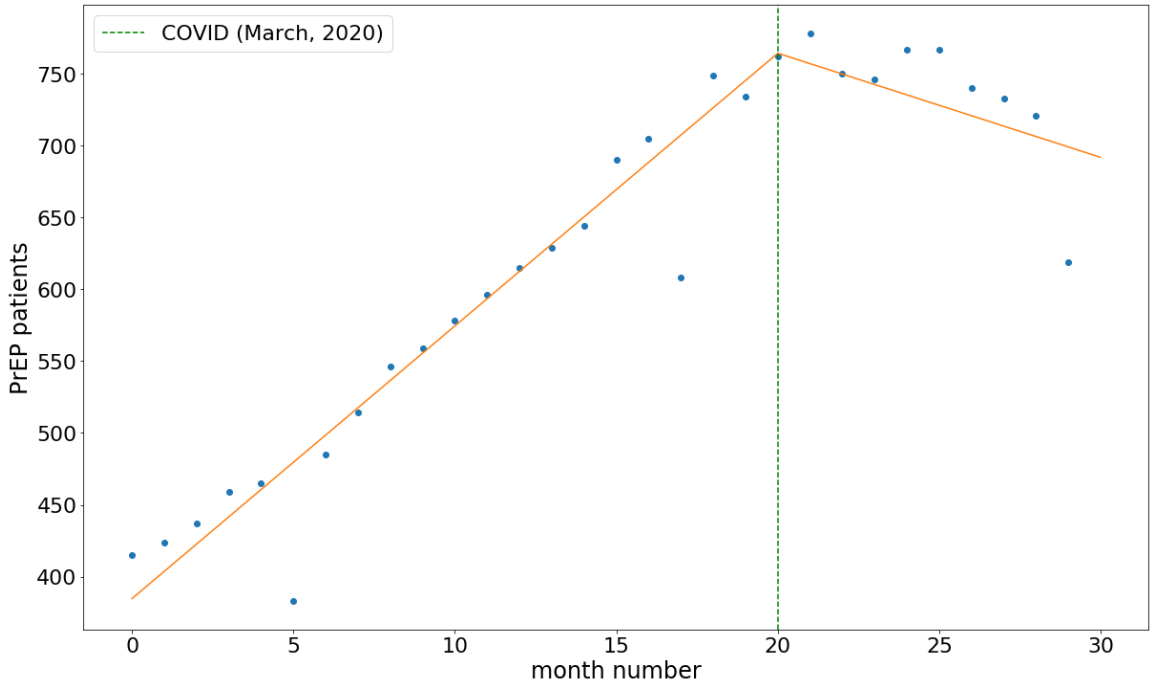


Figure 5.11: Pairwise regression plot (pre-COVID-19 vs. post-COVID-19) with "COVID-19" eliminated

Table 5.12: Pairwise regression table (G2Z vs. COVID-19 period) with "COVID-19" eliminated

	coefficient	standard error	t-statistics	p-value	[0.025	0.975]
G2Z	384.7625	14.352	26.809	0.000	355.315	414.210
months since G2Z	18.9773	1.140	16.645	0.000	16.638	21.317
months since COVID-19	-26.2227	3.665	-7.155	0.000	-33.743	-18.703

5.4 Conclusion

Segmented regression analysis of interrupted time series data can quantify the immediate and gradual effects of the interruptions on the outcome of interest [61]. We show that the segmented regression using a single model and pairwise models empirically yields similar results. We also show that the single-segmented analysis provides better insight into the entire time series than pairwise models, where the information from the entire time series is not considered at all times.

We show a slow rise in the patients using PrEP in the ramp-up era from January

2013 to December 2015. This can be attributed to the slow ramp-up in patients using PrEP in Mecklenburg County in the early days of its introduction. We also show an immediate jump in the monthly PrEP patients at the ramp-up interruption and a significant gradual effect or increasing trend in the pre-G2Z period. We show that the G2Z-MC intervention had no significant immediate effect on the number of PrEP patients. G2Z-MC did have a significant gradual effect on the number of patients using PrEP in a month in the G2Z period. The G2Z-MC intervention increased the number of monthly PrEP patients by 8.8153 since its inception. We show that the COVID-19 pandemic interruption in heath care delivery had an immediate effect and reversed the upward trend in PrEP patients. However this trend was gradual with a slope in patient loss less severe than patient gain in the COVID-19 period.

The monthly HIV incidences in the county displayed considerable variability and did not follow a linear trend from 2013 to 2019. We do not have access to actual monthly population in the county and since any interpolated value for monthly population follows the interpolation function and is redundant to use as a denominator in a regression analysis. Hence, we conducted segmented and pairwise regression analysis only on the actual monthly PrEP patient counts and not on population-adjusted PrEP patients or PrEP-to-need ratio (PnR). This method establishes a framework to evaluate any other new Public Health Intervention (PHI) efforts carried out by MCPH to quantify its effects on PrEP use in the county.

CHAPTER 6: DISCUSSIONS

6.1 Implications of results

Living with HIV comes with a huge economic and social burden on patients and society. The "Ending the HIV epidemic" started in 2019 calls to end the HIV epidemic in the US by 2030, with major emphasis on PrEP usage. The major objective of G2Z-MC drafted by MCPH as a part of the EHE plan is to increase PrEP use within Mecklenburg County, NC. Lack of previous studies monitoring PrEP use within the county poses a challenge for MCPH to make informed decisions to allocate resources in the county. Our study quantifies and monitors PrEP use within the county from 2013 to 2019 in various demographic and geographic groups, providing MCPH with the needed insights to plan future intervention efforts to increase PrEP coverage. We showed the annual trends in population-adjusted PrEP users, PrEP pills dispensed, and PnR bounds annually from 2013 to 2019. The effectiveness of G2Z-MC, both immediate and gradual effects, was quantified, which provides actionable insights on the intervention's impact in increasing PrEP usage in the county.

6.2 Impact of the study

Monitoring PrEP coverage using traditional metrics like population-adjusted PrEP patients, PrEP pills, PnR ratio, along with the novel metric daPnR in the study, can be used to confidently identify PrEP coverage corresponding to its need in various other similar counties. This study provides a framework to monitor PrEP use in geographically smaller subgroups, which was impossible due to the low number of PrEP users and risk of patient privacy. Geographically granular insights can warn county public health officials at an early stage about an outbreak or unusual patterns

in PrEP usage before it spreads to the entire county, decreasing the response time by the county public health department.

The study highlights the differences in PrEP use in different demographic and geographic groups, which can address disparities in underserved subgroups. This would mean that more PrEP is given to where there is a disparity, which reduces new infection as people with the most risk are getting PrEP. The study can be used to continuously monitor PrEP use and intervention effects in the future within a county acting as an alternative framework for informed policy-making and evaluation.

The framework established in this study can be used to monitor and evaluate prevention-to-need in other infectious diseases. This enables to devise demographically and geographically targeted interventions like vaccine boosters in groups with the highest need. The study can also evaluate the impact of other public health interventions like mandated shelter in place and face coverings in terms of new incidences reduced, viral load or trends in vaccine use.

6.3 Limitations

We used pharmacy claims data to quantify PrEP use which is not a direct but an indirect indicator of PrEP use. Since we did not use diagnostic codes (to use a more comprehensive dataset) but used a rule-based algorithm to identify PrEP episodes, this may have misidentified some episodes as for PrEP. On the other hand, the elimination of the requirement of medical coding for cross-validation avoided the exclusion of patients from analyses because of lack of available medical diagnosis coding. Also, our rule-based algorithm was developed based on clinically rational PrEP and ARV use scenarios in thousands of patient and drug regimens.

We use ZIP code of the pharmacy's location instead of patient's residence to derive geographic insights. Even though this allowed us to use a more comprehensive dataset (around 92% coverage), further research calls for a sensitivity analysis using the patient's residence ZIP code.

6.4 Future work

The study can be extended to continuously monitor PrEP use in the future within the county. Focusing on providing PrEP to demographically and geographically underserved groups needs to be emphasised to realize the goals of EHE and G2Z-MC programs. More research is needed to study the influence of socio-economic factors on PrEP use in underserved demographic and geographic groups. Further investigation is needed to identify and evaluate any other external factors that may have influenced PrEP use in the county other than G2Z-MC intervention such as national advertising campaigns, federal government aid, and other national initiatives to increase PrEP use.

Further research is needed to understand the PrEP prescription patterns in patients. Conducting survival analyses to quantify time to PrEP discontinuation or HIV diagnosis in patients can provide deeper insights in PrEP prescription and risk patterns. For example, quantifying time to discontinue PrEP in different demograhic, geographic, or socio-economic groups can enable us to understand which groups tend to stay on PrEP as opposed to others. Similarly quantifying time to HIV diagnosis in PrEP patients can quantify risk in not having a strict adherence to PrEP in at-risk patients. Conducting similar analyses between various subgroups can also quantify the effects of external factors like household income, insurance type, etc on adherence to PrEP and risk of HIV acquisition.

REFERENCES

[1] Centers for Disease Control and Prevention, "HIV Surveillance Report, 2019; vol.32." https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html, Published May, 2021. Accessed July 15, 2020.

[2] Office of Infectious Disease and HIV/AIDS Policy, HHS, "Overview – HIV.gov." https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/overview, Published May, 2020. Last accessed on August 10, 2020.

[3] N. S. Harris, A. S. Johnson, Y.-L. A. Huang, D. Kern, P. Fulton, D. K. Smith, L. A. Valleroy, and H. I. Hall, "Vital Signs: Status of Human Immunodeficiency Virus Testing, Viral Suppression, and HIV Preexposure Prophylaxis — United States, 2013–2018," *MMWR Morbidity and Mortality Weekly Report 2019*, vol. 68, no. 48, pp. 1117–1123, 2019. doi: http://dx.doi.org/10.15585/mmwr.mm6848e1.

[4] Centers for Disease Control and Prevention, "CDC's HIV Work Saves Lives and Money Infographics | CDC." https://www.cdc.gov/nchhstp/budget/infographics/hiv.html, Published February, 2019. Last accessed on October 15, 2019.

[5] B. R. Schackman, J. A. Fleishman, A. E. Su, B. K. Berkowitz, R. D. Moore, R. P. Walensky, J. E. Becker, C. Voss, A. D. Paltiel, M. C. Weinstein, K. A. Freedberg, K. A. Gebo, and E. Losina, "The lifetime medical cost savings from preventing HIV in the United States," *Medical Care*, vol. 53, pp. 293–301, mar 2015. doi: https://doi.org/10.1097/MLR.0000000000000308.

[6] A. S. Fauci, R. R. Redfield, G. Sigounas, M. D. Weahkee, and B. P. Giroir, "Ending the HIV Epidemic: A Plan for the United States," *Journal of the American Medical Association*, vol. 321, pp. 844–845, 03 2019. doi:10.1001/jama.2019.1343.

[7] Y. L. A. Huang, W. Zhu, D. K. Smith, N. Harris, and K. W. Hoover, "HIV Preexposure prophylaxis, by race and ethnicity — United States, 2014–2016," *MMWR Morbidity and Mortality Weekly Report 2018*, vol. 67, no. 41, pp. 1147–1150. doi: http://dx.doi.org/10.15585/mmwr.mm6741a3.

[8] R. M. Grant, J. R. Lama, P. L. Anderson, V. McMahan, A. Y. Liu, L. Vargas, P. Goicochea, M. Casapía, J. V. Guanira-Carranza, M. E. Ramirez-Cardich, O. Montoya-Herrera, T. Fernández, V. G. Veloso, S. P. Buchbinder, S. Chariyalertsak, M. Schechter, L.-G. Bekker, K. H. Mayer, E. G. Kallás, K. R. Amico, K. Mulligan, L. R. Bushman, R. J. Hance, C. Ganoza, P. Defechereux, B. Postle, F. Wang, J. J. McConnell, J.-H. Zheng, J. Lee, J. F. Rooney, H. S. Jaffe, A. I. Martinez, D. N. Burns, and D. V. Glidden, "Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men," *New England Journal of Medicine 2010*, vol. 363, no. 27, pp. 2587–2599. doi: https://doi.org/10.1056/NEJMoa1011205.

[9] J. M. Baeten, D. Donnell, P. Ndase, N. R. Mugo, J. D. Campbell, J. Wangisi, J. W. Tappero, E. A. Bukusi, C. R. Cohen, E. Katabira, A. Ronald, E. Tumwesigye, E. Were, K. H. Fife, J. Kiarie, C. Farquhar, G. John-Stewart, A. Kakia, J. Odoyo, A. Mucunguzi, E. Nakku-Joloba, R. Twesigye, K. Ngure, C. Apaka, H. Tamooh, F. Gabona, A. Mujugira, D. Panteleeff, K. K. Thomas, L. Kidoguchi, M. Krows, J. Revall, S. Morrison, H. Haugen, M. Emmanuel-Ogier, L. Ondrejcek, R. W. Coombs, L. Frenkel, C. Hendrix, N. N. Bumpus, D. Bangsberg, J. E. Haberer, W. S. Stevens, J. R. Lingappa, and C. Celum, "Antiretroviral Prophylaxis for HIV Prevention in Heterosexual Men and Women," *New England Journal of Medicine 2012*, vol. 367, no. 5, pp. 399–410. doi: https://doi.org/10.1056/NEJMoa1108524.

[10] M. C. Thigpen, P. M. Kebaabetswe, L. A. Paxton, D. K. Smith, C. E. Rose, T. M. Segolodi, F. L. Henderson, S. R. Pathak, F. A. Soud, K. L. Chillag, R. Mutanhaurwa, L. I. Chirwa, M. Kasonde, D. Abebe, E. Buliva, R. J. Gvetadze, S. Johnson, T. Sukalac, V. T. Thomas, C. Hart, J. A. Johnson, C. K. Malotte, C. W. Hendrix, and J. T. Brooks, "Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana," *New England Journal of Medicine 2012*, vol. 367, no. 5, pp. 423–434. doi: https://doi.org/10.1056/NEJMoa1110711.

[11] K. Choopanya, M. Martin, P. Suntharasamai, U. Sangkum, P. A. Mock, M. Leethochawalit, S. Chiamwongpaet, P. Kitisin, P. Natrujirote, S. Kittimunkong, R. Chuachoowong, R. J. Gvetadze, J. M. McNicholl, L. A. Paxton, M. E. Curlin, C. W. Hendrix, and S. Vanichseni, "Antiretroviral prophylaxis for HIV infection in injecting drug users in Bangkok, Thailand (the Bangkok Tenofovir Study): A randomised, double-blind, placebo-controlled phase 3 trial," *The Lancet 2013*, vol. 381, no. 9883, pp. 2083–2090. doi: https://doi.org/10.1016/S0140-6736(13)61127-7.

[12] Food and Drug Administration (FDA), "FDA approves first drug for reducing the risk of sexually acquired HIV infection." https://aidsinfo.nih.gov/news/1254/fda-approves-first-drug-for-reducing-the-risk-of-sexually-acquired-hiv-infection, Published July 16, 2012. Last accessed on May 5, 2020.

[13] "TRUVADA for PrEP® (Pre-Exposure Prophylaxis)." Available online: https://www.truvada.com/, Last accessed on August 18, 2021.

[14] Food and Drug Administration (FDA), "FDA approves second drug to prevent HIV infection as part of ongoing efforts to end the HIV epidemic." https://www.fda.gov/news-events/press-announcements/fda-approves-second-drug-prevent-hiv-infection-part-ongoing-efforts-end-hiv-epidemic, Published October 03, 2019. Last accessed on July 7, 2020.

[15] "What is DESCOVY® for PrEP." Available online: https://www.descovy.com/, Last accessed on August 18, 2021.

[16] Centers for Disease Control and Prevention, "Monitoring selected national HIV prevention and careobjectives by using HIV surveillance data—United States and 6 dependent areas, 2016." https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-23-4.pdf, Published June, 2018. Accessed October 15, 2019.

[17] Li, Zihao and Purcell, David W. and Sansom, Stephanie L. and Hayes, Demorah and Hall, H. Irene, "Vital signs: HIV transmission along the continuum of care — United States, 2016," *Morbidity and Mortality Weekly Report*, no. 11, pp. 267–272. doi: http://dx.doi.org/10.15585/mmwr.mm6811e1.

[18] K. L. Hess, X. Hu, A. Lansky, J. Mermin, and H. I. Hall, "Lifetime risk of a diagnosis of HIV infection in the United States," *Annals of Epidemiology 2017*, vol. 27, no. 4, pp. 238–243. doi: https://doi.org/10.1016/j.annepidem.2017.02.003.

[19] Centers for Disease Control and Prevention, "US public health service: Preexposure prophylaxis for the prevention of hiv infection in the United States—2017 update: a clinical practice guideline.." https://www.cdc.gov/hiv/pdf/risk/prep/cdc-hiv-prep-guidelines-2017.pdf.Published, Published 2018. Accessed September 16, 2020.

[20] P. Dawn K. Smith, MD; Michelle Van Handel, MPH; Richard J. Wolitski, PhD; Jo Ellen Stryker, PhD; H. Irene Hall, PhD; Joseph Prejean, PhD; Linda J. Koenig, PhD; Linda A. Valleroy, "Vital Signs: Estimated Percentages and Numbers of Adults with Indications for Preexposure Prophylaxis to Prevent HIV Acquisition — United States, 2015," pp. 1291–1295, 2015. Available at: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6446a4.htm.

[21] V. Supervie, M. Barrett, J. S. Kahn, G. Musuka, T. L. Moeti, L. Busang, and S. Blower, "Modeling dynamic interactions between pre-exposure prophylaxis interventions treatment programs: Predicting HIV transmission resistance," *Scientific Reports*, vol. 1, 2011. doi: https://doi.org/10.1038/srep00185.

[22] J. E. Volk, J. L. Marcus, T. Phengrasamy, D. Blechinger, D. P. Nguyen, S. Follansbee, and C. B. Hare, "No New HIV Infections with Increasing Use of HIV Preexposure Prophylaxis in a Clinical Practice Setting," *Clinical Infectious Diseases*, vol. 61, pp. 1601–1603, nov 2015. doi: https://doi.org/10.1093/cid/civ778.

[23] S. McCormack, D. T. Dunn, M. Desai, D. I. Dolling, M. Gafos, R. Gilson, A. K. Sullivan, A. Clarke, I. Reeves, G. Schembri, N. Mackie, C. Bowman, C. J. Lacey, V. Apea, M. Brady, J. Fox, S. Taylor, S. Antonucci, S. H. Khoo, J. Rooney, A. Nardone, M. Fisher, A. McOwan, A. N. Phillips, A. M. Johnson, B. Gazzard, and O. N. Gill, "Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): Effectiveness results from the pilot phase of a pragmatic open-label randomised trial," *The Lancet*, vol. 387, pp. 53–60, Jan 2016. doi: https://doi.org/10.1016/S0140-6736(15)00056-2.

[24] J. L. Marcus, L. B. Hurley, C. B. Hare, D. P. Nguyen, T. Phengrasamy, M. J. Silverberg, J. E. Stoltey, and J. E. Volk, "Preexposure prophylaxis for HIV prevention in a large integrated health care system: adherence, renal safety, and discontinuation," vol. 73, pp. 540–546, Dec 2016. doi: https://doi.org/10.1097/QAI.0000000000001129.

[25] A. E. Grulich, R. Guy, J. Amin, F. Jin, C. Selvey, J. Holden, H. M. A. Schmidt, I. Zablotska, K. Price, B. Whittaker, K. Chant, C. Cooper, S. McGill, B. Telfer, B. Yeung, G. Levitt, E. E. Ogilvie, N. J. Dharan, M. A. Hammoud, S. Vaccher, L. Watchirs-Smith, A. McNulty, D. J. Smith, D. M. Allen, D. Baker, M. Bloch, R. I. Bopage, K. Brown, A. Carr, C. J. Carmody, K. L. Collins, R. Finlayson, R. Foster, E. Y. Jackson, D. A. Lewis, J. Lusk, C. C. O'Connor, N. Ryder, E. Vlahakis, P. Read, D. A. Cooper, D. Smith, C. Ooi, S. Martin, T. M. Soo, D. Templeton, D. Townson, B. Forssman, N. Doong, G. Han Thai, B. Anderson, H. MacLeod, and N. Parkhill, "Population-level effectiveness of rapid, targeted, high-coverage roll-out of HIV pre-exposure prophylaxis in men who have sex with men: the EPIC-NSW prospective cohort study," *The Lancet HIV*, vol. 5, pp. e629–e637, Nov 2018. doi: https://doi.org/10.1016/S2352-3018(18)30215-7.

[26] A. J. Siegler, F. Mouhanna, R. M. Giler, K. Weiss, E. Pembleton, J. Guest, J. Jones, A. Castel, H. Yeung, M. Kramer, S. McCallister, and P. S. Sullivan, "The prevalence of pre-exposure prophylaxis use and the pre-exposure prophylaxis-to-need ratio in the fourth quarter of 2017, United States," *Annals of Epidemiology*, vol. 28, pp. 841–849, Dec 2018. doi: 10.1016/j.annepidem.2018.06.005.

[27] P. S. Sullivan, R. M. Giler, F. Mouhanna, E. S. Pembleton, J. L. Guest, J. Jones, A. D. Castel, H. Yeung, M. Kramer, S. McCallister, and A. J. Siegler, "Trends in the use of oral emtricitabine/tenofovir disoproxil fumarate for pre-exposure prophylaxis against HIV infection, United States, 2012-2017," *Annals of Epidemiology*, vol. 28, pp. 833–840, Dec 2018. doi: 10.1016/j.annepidem.2018.06.009.

[28] S. Robertino Mera, David Magnuson, Trevor Hawkins, Staci Bush, Keith Rawlings and McCallister, "Changes in Truvada® for HIV Pre-exposure Prophylaxis Utilization in the USA: 2012-2016 - Disparities." Available online: https://www.natap.org/2017/IAS/IAS_72.htm.

[29] P. S. Sullivan, F. Mouhanna, R. Mera, E. Pembleton, A. D. Castel, C. Jaggi, J. Jones, M. R. Kramer, P. McGuinness, S. McCallister, and A. J. Siegler, "Methods for county-level estimation of pre-exposure prophylaxis coverage and application to the U.S. Ending the HIV Epidemic jurisdictions," *Annals of Epidemiology*, vol. 44, pp. 16–30, Apr 2020. doi: 10.1016/j.annepidem.2020.01.004.

[30] N. W. Furukawa, D. K. Smith, C. J. Gonzalez, Y. L. A. Huang, D. B. Hanna, U. R. Felsen, W. Zhu, J. H. Arnsten, and V. V. Patel, "Evaluation of Algorithms Used for PrEP Surveillance Using a Reference Population From New York City,

July 2016–June 2018," *Public Health Reports*, vol. 135, pp. 202–210, Mar 2020. doi: 10.1177/0033354920904085.

[31] AIDSVu, "Charlotte (Mecklenburg County)." https://aidsvu.org/state/north-carolina/charlotte/, Published February 18, 2019. Last accessed on June 20, 2020.

[32] Mecklenburg County Public Health, "Getting to Zero Mecklenburg." Available online: https://www.mecknc.gov/HealthDepartment/GettingToZero, Accessed on: 2021-03-03.

[33] H. Leyla, "Personal communication via presentation: Symphony Health, a PRA Health Sciences Company," May 19, 2020.

[34] "HIV Surveillance | Reports| Resource Library | HIV/AIDS | CDC." Available online: https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html, Accessed on: 2021-06-18.

[35] National Research Council, *Sharing Research Data*. Washington, DC: National Academies Press, 1985. doi: https://doi.org/10.17226/2033.

[36] D. J. Hutchon, "Publishing raw data and real time statistical analysis on e-journals.," *BMJ (Clinical Research Edition)*, vol. 322, p. 530, mar 2001. doi: https://doi.org/10.1136/bmj.322.7285.530.

[37] A. J. Vickers, "Whose data set is it anyway? Sharing raw data from randomized trials," *Trials*, vol. 7, no. 1, p. 15, 2006. doi: https://doi.org/10.1186/1745-6215-7-15.

[38] D. G. Altman and C. Cates, "Authors should make their data available.," nov 2001. Available online: https://pubmed.ncbi.nlm.nih.gov/11691772/.

[39] G. D. Smith, "Increasing the accessibility of data," *BMJ*, vol. 308, no. 6943, pp. 1519–1520, 1994. doi: https://doi.org/10.1136/bmj.308.6943.1519.

[40] C. J. Hogue, "Ethical issues in sharing epidemiologic data.," *Journal of Clinical Epidemiology*, vol. 44 Suppl 1, pp. 103S–107S, 1991. doi: https://doi.org/10.1016/0895-4356(91)90183-a.

[41] J. Kirwan, "Making original data from clinical studies available for alternative analysis," *The Journal of Rheumatology*, vol. 24, pp. 822–825, jun 1997. Available online: https://pubmed.ncbi.nlm.nih.gov/9150066/.

[42] U.S. Department of Health and Human Services, "Methods for De-identification of PHI | HHS.gov." Available online: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.htmlprotected, Accessed on: 2021-01-18.

[43] T. Dalenius, "Finding a needle in a haystack- or identifying anonymous census record," *Institute for Social and Economic Research*, 1986. Available online: https://www.proquest.com/docview/1266806751.

[44] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," *SRI International*, 1998. doi: https://doi.org/10.1184/R1/6625469.v1.

[45] K. El Emam and F. K. Dankar, "Protecting Privacy Using k-Anonymity," *Journal of the American Medical Informatics Association*, vol. 15, pp. 627–637, Sep 2008. doi: 10.1197/jamia.M2716.

[46] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-Anonymization Using Clustering Techniques BT - Advances in Databases: Concepts, Systems and Applications," in *Springer Berlin Heidelberg* (R. Kotagiri, P. R. Krishna, M. Mohania, and E. Nantajeewarawat, eds.), (Berlin, Heidelberg), pp. 188–200, 2007. doi: https://doi.org/10.1007/978-3-540-71703-4_18.

[47] C. C. Aggarwal and P. S. Yu, "A Condensation Approach to Privacy Preserving Data Mining BT - Advances in Database Technology - EDBT 2004," in *Springer Berlin Heidelberg* (E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, and E. Ferrari, eds.), (Berlin, Heidelberg), pp. 183–199, 2004. Available online: https://www.semanticscholar.org/paper/A-Condensation-Approach-to-Privacy-Preserving-Data-Aggarwal-Yu/c90ce2aa1fc2360ddb2dd75e1efddaca16dfe688.

[48] J. J. Kim, J. J. Kim, W. E. Winkler, and B. O. T. Census, "Masking Microdata Files," *Proceedings of the Survey Research Methods Section, American Statistical Association*, vol. 14, pp. 114–119, 1995. Available online: http://www.asasrms.org/Proceedings/papers/1995_017.pdf.

[49] C. C. Aggarwal and P. S. Yu, "Privacy-Preserving Data Mining," in *Models and Algorithm*, pp. 1–9, Springer, 2008. doi: https://doi.org/10.1007/978-1-84996-338-1_6.

[50] H. Leyla, "Personal communication: Symphony Health, a PRA Health Sciences Company," May 19, 2020.

[51] G. Brummelen Van, *Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry - Glen Van Brummelen*. Princeton University Press, 2012. Available online: https://books.google.com/books?id=0BCCz8Sx5wkCpg=PR7v=onepageqf=false, Accessed on: 2021-02-11.

[52] S. Kettle, "Distance on a Sphere: The Haversine Formula - GeoNet, The Esri Community." Available online: https://community.esri.com/t5/coordinate-reference-systems-blog/distance-on-a-sphere-the-haversine-formula/bap/902128, Accessed on: 2021-02-11.

[53] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. Available online: https://docs.python.org/3/.

[54] Gurobi Optimization LLC, "Gurobi Optimizer Reference Manual." Available online: http://www.gurobi.com, Accessed on: 2021-01-18, 2021.

[55] Gurobi Optimization LLC, "Gurobi Optimizer Reference Manual - MIPFocus." Available online: https://www.gurobi.com/documentation/9.1/refman/mipfocus.html, Accessed on: 2021-01-18, 2021.

[56] E. Kontopantelis, T. Doran, D. A. Springate, I. Buchan, and D. Reeves, "Regression based quasi-experimental approach when randomisation is not an option: Interrupted time series analysis," *BMJ (Online)*, vol. 350, jun 2015. doi: 10.1136/bmj.h2750.

[57] B. Guthrie, K. Payne, P. Alderson, M. E. McMurdo, and S. W. Mercer, "Adapting clinical guidelines to take account of multimorbidity," oct 2012. doi: https://doi.org/10.1136/bmj.e6341.

[58] C. Saunders, C. D. Byrne, B. Guthrie, R. S. Lindsay, J. A. Mcknight, S. Philip, N. Sattar, J. J. Walker, S. H. Wild, I. Brady, J. Chalmers, H. Colhoun, S. Cunningham, R. Elder, A. Emslie-Smith, C. Fischbacher, L. Govan, N. Jones, G. Lese, D. Levine, S. Livingstone, H. Looker, R. McAlpine, A. Morris, D. Pearson, and J. Petrie, "External validity of randomized controlled trials of glycaemic control and vascular disease: How representative are participants?," *Diabetic Medicine*, vol. 30, pp. 300–308, mar 2013. doi: https://doi.org/10.1111/dme.12047.

[59] Institute for Work and Health, "Observational vs. experimental studies," 2016. Available online: https://www.iwh.on.ca/what-researchers-mean-by/observational-vs-experimental-studies, Last accessed on August 15, 2021.

[60] A. Abadie, A. Diamond, and J. Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, vol. 105, pp. 493–505, jun 2012. doi: https://doi.org/10.1198/jasa.2009.ap08746.

[61] A. K. Wagner, S. B. Soumerai, F. Zhang, and D. Ross-Degnan, "Segmented regression analysis of interrupted time series studies in medication use research," *Journal of Clinical Pharmacy and Therapeutics*, vol. 27, no. 4, pp. 299–309, 2002. doi: https://doi.org/10.1046/j.1365-2710.2002.00430.x.

[62] R. M. Pinto and S. Park, "COVID-19 Pandemic Disrupts HIV Continuum of Care and Prevention: Implications for Research and Practice Concerning Community-Based Organizations and Frontline Providers," *AIDS and Behavior*, vol. 24, p. 1, sep 2020. doi: https://doi.org/10.1007/s10461-020-02893-3.

[63] S. Brawley, J. Dinger, C. Nguyen, and J. Anderson, "Impact of COVID-19 related shelter-in-place orders on PrEP access, usage and HIV risk behaviors in the United States," *Journal of the International Aids Society*, pp. 178–178, 2020. Available online: https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/en/covidwho-704620.