Coverage that Goes the Distance:
Extending Distance Metrics and Clustering Methods to Assess Access to HIV Preventative Care

Samuel Darin Powers
Charlottesville, Virginia

Bachelor of Arts, University of Virginia, 2019

A Thesis presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Master of Arts

Department of Psychology

University of Virginia
May, 2021

Acknowledgements

I would like to express my deepest appreciation to my advisor, Dr. Karen Schmidt, for her personal investment in my growth this year, for the invaluable direction and clarity she provided to this work, and for constantly reminding me to set boundaries and prioritize an achievable scope. I extend my sincere gratitude to Dr. Kathleen McManus in the UVA Division of Infectious Diseases and International Health for her personal mentorship over that past three years as well as for her substantive contributions to the content and interpretation of this work. I also wish to thank Amy Killelea, J.D. — for her policy and subject-matter guidance — and Dr. Steve Boker — both for expanding my statistical imagination over the course of this year and for his readership of this work. Finally, I would like to extend thanks to the Jefferson Scholars Foundation for making my time at the University of Virginia possible through their generosity as well as to my family for their continual emotional support and for always answering the phone.

Abstract

In this work, we extend clustering methodologies to find groups in healthcare plan data, assessing the extent to which specific coverage practices facilitate or restrict effective preventative care for HIV. In doing so, we rely on Gower's underexplored ideas about weighting in distance metrices to create a procedure that handles nested dependencies. Overall, the distance metric chosen effectively translates our priorities based on theory into a form that works well with common clustering algorithms. Running trials with several algorithms, we find convergent structures, settling on a hierarchical approach with three distinct clusters. The clusters exhibit distinct contrasts on how plans cover specific benefits that allow for ordinal interpretations in terms of plan restrictiveness. Broad level interpretation of the output suggests that, across the United States, monthly premiums are not related to plan restrictiveness, prior authorization is less likely among plans where individuals accept higher out of pocket payments for care, state-wide approaches affect what care residents can access, and variability in markets within states, specifically Texas, reflect patterns of discrimination toward individuals at risk of HIV.

Contents

1 Introduction

1.1 Why this Work

Currently, more than 1.2 million Americans live with HIV, a number that grows every year. In 2018, an estimated 36,400 individuals in the US were newly infected with HIV (Center for Disease Control [CDC], 2020). While this increase in people living with HIV (PLWH) is an estimated two-thirds of what the yearly infection rate in the US was during the 1980s at the height of the epidemic, CDC data suggests that the year-to-year infection rate has held constant between 2014 and 2018. During those years, the United States, as a whole, has not made much progress against ending the HIV epidemic (CDC, 2021).

People living with HIV (PLWH) who achieve undetectable viral loads through use of antiretroviral therapy (ART) cannot transmit the virus through sex, a principle known as "Undetectable = Untransmittable" or "U=U" (Eisinger, Dieffenbach & Fauci, 2019). Furthermore, in 2012, the Federal Drug Administration (FDA) approved TDF/FTC, a combination of two drugs traditionally administered as part of antiretroviral therapy (ART) for those diagnosed with HIV, to be used as Pre-Exposure Prophylaxis for HIV (PrEP) among those who are at risk of HIV infection (FDA, 2019). Daily use of TDF/FTC reduces the risk of acquiring HIV through sex by 99% and by at least 74% among people who inject drugs (PWID) (CDC, 2019).

For the last nine years, we have had the tools to prevent new HIV infections. Yet, there are gaps in service provision networks and other structural barriers that have kept powerful preventative measures from reaching those who most need them (CDC, 2021). This is the motivation for the present work — health coverage has a role to play in ensuring those who

need HIV preventative care are both informed and able to acquire it easily. And yet, as an opaque and somewhat disparate system, it can be difficult to audit where health insurance impedes and where it facilitates effective care. The present work explores quantitative methods to accomplish that very task.

1.1.1 Who are Most Affected by New HIV Infections?

Individuals who experience identity-based marginalization on the basis of race, ethnicity, and sexual identity are at a far higher risk for HIV infection. A 2021 report by the CDC demonstrates significant and increasing disparities in the burden of new infections despite the rate of new infections holding constant from 2014. Overall, in 2018, gay and bisexual men who have sex with men (MSM) comprised approximately 67% of new infections. The burden of HIV infections was highest among African American (26%) and Hispanic/Latino (22%) MSM, among whom the rate of infection increased from 2014 to 2018. White MSM comprised the third highest proportion of new infections in 2018 (15.6%), but also saw a 20% decrease in new infections from 2014. After White MSM, African American Heterosexual Women comprised 9.6% of new infections in 2018. Eliminating structural barriers for ending the HIV epidemic is a key tenant of pursuing health equity for US racial and sexual minorities.

Approximately 51% of new HIV cases in 2018 occurred in Southern states (CDC, 2019, September). In 2016 and 2017, over half of new transmissions were identified in just 48 US counties in addition to Washington D.C. and San Juan, Puerto Rico. Key states including Alabama, Arkansas, Kentucky, Missouri, Mississippi, Oklahoma, and South Carolina also reported a

heightened burden of new HIV infections among rural populations, emphasizing the importance of rural care networks to ending the HIV epidemic.

1.1.2 HIV Preventative Care

As mentioned earlier, acquiring HIV is medically preventable due to the innovation of Pre-Exposure Prophylaxis for HIV (PrEP). PrEP is a once daily oral drug capable of preventing 99% of new infections in MSM and 70% of new infections in people who inject drugs (Harawa, et al., 2018). When on PrEP, individuals traditionally at risk of HIV can continue to live their lives with relative freedom. That said, only 18% of the estimated million American who show indications that they could benefit from PrEP are using the medication (CDC, 2021). These gaps are heightened among individuals most at risk for HIV including Black and Hispanic MSM and transgender women (Zarwell et al., 2020; Finlayson et al., 2019).

This issue is also exacerbated in the South. Those who use PrEP in the South comprise only 30% of the overall population of PrEP users nationally despite the South having 51% of new infections each year (Sullivan et al., 2019). As of 2020, PrEP exists in two key formulations, TDF/FTC and TAF/FTC. Both are effective among individuals with male biological sex at birth; however, TAF/FTC has not yet been validated for individuals born biologically female (FDA, 2019, October 3).

While PrEP is a feasible preventative mechanism for many individuals, being on PrEP requires regular monitoring of HIV and potential side effects. The US Preventative Services Taskforce, in accordance with evolving CDC guidelines for standards of care, recommends routine kidney function testing, serologic testing for hepatitis B & C, routine STI testing, quarterly

HIV CD4 counts and viral load screens, and regular pregnancy screenings. It also adds regular behavioral counseling as an important preventative measure to assist those at risk of HIV in exploring behavior-change based methods for lowering their risk of HIV (Owens et al., 2019).

1.1.3 Ending the HIV Epidemic

In early 2019, the US government outlined a new plan to end the HIV epidemic in the US by 2030. *Ending the HIV Epidemic: A Plan for America* (EHE) focuses on diagnosing unknown infections — which is estimated to be about 14% of people living with HIV (PLWH) — treating known infections with anti-retroviral therapy, and preventing over 250,000 new infections in the next 10 years (Fauci et al., 2019). The prevention arm of the EHE plan focuses specifically on expanding access to PrEP.

EHE emphasizes increasing PrEP awareness and demand through community-based outreach and education to increase knowledge and counter PrEP-related stigma. It also emphasizes increasing PrEP accessibility by educating healthcare providers, expanding TelePrEP services, creating clinical guidelines, and offering PrEP at community source locations like STD clinics and school health clinics (CDC, 2021). That said, the *EHE* plan does not directly address the role of health coverage providers in facilitating PrEP access.

1.2 Outline

Section 1 discusses how specific health coverage strategies can be barriers to effective HIV preventative care, introduces issues of comparability between health care plans, reviews past efforts to meaningfully describe trends in coverage, and explicitly states the key questions addressed by this work.

Section 2 introduces the data source informing this study, the HIX Compare; it then justifies the variables selected based on key concerns for HIV prevention and explains how they are operationalized within the data set. The second half of Section 2 introduces the theoretical background for clustering methods. It explains the two-step process of first defining a distance metric before running a clustering algorithm. Section 2.4 elaborates on distance metrics, explains issues posed by the data set, then selects and justifies a distance metric. Section 2.5 concludes by surveying clustering algorithms and laying out the process utilized later to determine an optimal clustering solution.

Section 3 presents results from the cluster analysis. It walks through preliminary fit statistics before reviewing the implications of each model on the data itself. It ends by selecting a final model, the three cluster Ward's solution, and describes the resultant clusters in detail.

Section 4 evaluates both steps of the clustering process — the distance metric and the algorithm — to determine suitability of the methods to the data. It then details the content-based implications of the three cluster solution on health coverage of HIV preventative care and concludes by suggesting research paths for the future.

1.3 Health Coverage as a Barrier to Prevention

The most frequently cited barriers to PrEP use include lack of clinicians with PrEP knowledge, absence of health insurance, stigma, and underestimation of personal HIV risk; however, the role of insurance structures in facilitating (or impeding) access to effective HIV preventative care, in specific, is understudied (Skolnik et al., 2020; Siegler et al., 2018; Seidman et al., 2016; Kay & Pinto, 2020). That said, there is significant literature investigating how insurance companies utilize specific plan factors to guide the choices of those enrolled which we can extend to apply to benefits relevant to HIV prevention. We review those here.

1.3.1 Cost-Sharing

Cost sharing is a healthcare payment structure where patients pay a portion of their health services out of pocket. All insurance benefits have an associated cost sharing structure. Cost sharing is a regulation mechanism used by insurance providers that applies across most variables discussed in this analysis.

From the perspective of health insurers, cost sharing can disincentivize and reduce patient usage of non-effective health services (Remler & Greene, 2009). This strategy of value-based cost sharing, where health insurance companies incentivize lower cost or preventative care and disincentivize care that is seen as less effective, is intended to better align insurer and patient interests leading to better patient health without increasing costs to insurers (Chernew et al., 2010; Thomson, Schang, & Chernew, 2013). Despite this, cost sharing is known to decrease overall drug and health services usage among socioeconomically disadvantaged populations,

increasing disparities in health care (Lexchin & Grootendorst, 2004; Doshi et al; 2016; Chernew et al., 2018).

Within cost-sharing structures, there are two mechanisms for deferring costs to patients: copay models and coinsurance models. In copay models, patients pay a pre-determined fixed amount per service (e.g. $15 for a 30 day supply of a drug). In coinsurance models, patients pay a pre-determined percentage of the total cost for the service (e.g. 30% of the cost for a 30 day supply of a drug). Insurance company usage of coinsurance for cost-sharing as opposed to copay is known to reduce patient usage of specialists and inpatient care (Fronstin & Roebuck, 2020). It has a measurable effect on preventative care including preventative medication and initiation of prescribed specialty drugs (Doshi et al., 2016). A 2010 study by Dor and Encinosa found that when participants were assigned to pay an expected $9 flat copay or a coinsurance percentage with an expected monetary value of $9 for their preventative medication, 34% of patients under copay refilled their medication whereas only 24% refilled under coinsurance. Fronstin and Roebuck (2020) hypothesize this is due to the inherent uncertainty in coinsurance payments.

1.3.2 Plan Tiering

Sometimes plans contain tiered structures. Tiered structures occur when a specific subset of in-network providers are placed on an approved Tier 1 list while a second subset of providers are demarcated as Tier 2. The cost sharing for Tier 1 providers is often lower to encourage utilization through a process known as Value-Based Cost Sharing (Sinaiko, Landrum, & Chernew, 2017). But it is also known to directly limit patient choices of care providers (Frank et al., 2015).

1.3.3 Prior Authorization

According to Cigna, one of the major health care providers in the US,

> The prior authorization (PA) process gives your health insurance company a chance to review how necessary a certain medication may be in treating your medical condition. For example, some brand name medications are very costly. During their review the insurance company may decide a generic or another lower cost alternative may work equally well in treating your medical condition. Other types of medications are dangerous when combined with others you may already be taking, others are very addictive, etc (https://www.cigna.com/).

A review of other major insurance websites finds a few key reasons for Prior Authorization practices, chiefly that (1) some drugs are unsafe when combined with other medications. (2) There may be lower-cost, equally effective alternatives available. (3) Some drugs should only be used for certain health conditions. (4) Some drugs are often misused or abused. and (5) Some drugs often used for cosmetic purposes which insurances do not cover.

However, a survey of medical providers agrees that PA is a key barrier to PrEP access and adherence (Petroll et al., 2017). It adds a degree of, often unnecessary, friction in an already complicated process. In 2017, 35% of PA requests for medication made to Medicare Part D were rejected; however, upon appeal, 73% of denials were overturned (Office of Inspector General, 2019). Patients can easily become frustrated by the process, which adds extra time and can be stigmatizing to patients as insurance companies press into their sexual habits and histories. Changing plans and formularies can also lead to prescriptions requiring additional physician information for release when patients show up arrive at pharmacies to pick up their prescriptions (Resnek, 2020). A 2019 survey suggest that 37% of prescriptions that are rejected at pharmacies due to PA complications are later never picked up (*ePA National Adoption Scorecard*, 2019).

A 5-10 day delay is sufficient time for acquisition of HIV (Kay & Pinto, 2020). For patients already taking PrEP, remembering to request a refill two weeks ahead of time can be difficult. This leads to a lapse in protection while waiting for re-authorization. More broadly, physicians in general tend to agree that PA creates a barrier to efficient and effective care. In a survey run by the American Medical Association in 2018, 91% of physicians reported having observed care delays for patients and 74% reported having observed care abandonment because of PA.

1.3.4 Specialty Tiering

Placing PrEP in a specialty tier, a designation traditionally reserved for drugs that require special administration or are meant to care for rarer diseases, allows justification for higher cost and greater restrictions, whereas placing PrEP on a lower tier indicates a different attitude toward the drug on the part of the insurance provider (Lotvin et al, 2014). Specialty Tiering is also a cost-regulation mechanism as survey data shows that adults are willing to pay higher premium amounts to obtain better specialty drug coverage benefits (Romley et al., 2012).

1.4 Past Work On Comparing Health Plan Offerings

Health care plan data is complex. To give but a brief example, to calculate the price one expects to pay for a year supply of a medication requires first examining the plan formulary to identify how the plan categorizes the drug — as a specialty drug, a preferred drug, a non-preferred drug, or a generic drug. From there, one must find the corresponding benefit in the plan's benefit listings. The benefit may be covered with a coinsurance or a copay. And the corresponding amounts could differ based on when in the year one reached the deductible or if one has hit one's maximum out of pocket amount cap. So, the drug cost really depends on what other medical services one has also paid for that year. Comparing projected drug costs between plans requires one to make assumptions that may only be reasonably asserted by someone who fully understands the system. Because of this, the National Alliance of State and Territorial Aids Directors (NASTAD) created a tool called PrEPcost.org to assist individuals in selecting plans.

Previously research conducted by the McManus Lab at UVA (McManus & Powers et al, 2020) examined this issue of auditing plans to assess equitable offerings across the country by honing in on just one aspect of care — prior authorization for PrEP. Our work demonstrated (1) significant geographic disparities in prior authorization requirements for TDF/FTC with prior authorization at 13 times the rate in the South as in Northern states and (2) characteristics associated with increased cost shifting to patients (economic barriers) were associated with decreased prior authorization (administrative barriers) for TDF/FTC.

That said, if calculating cost for one drug requires a full website dedicated to determining optimal care options and if one dimension — like prior authorization for one drug —can ground an entire inquiry, it is easy to see how comparing plans across a matrix of benefits quickly

becomes a daunting task. And yet, it is important to understand how certain restrictive actions vary together and create multiplicative restrictions. There has been some recent work comparing patient satisfaction and health outcomes across plans from patients of different background to create scoring systems that rate plans on their commitment to serving high quality care to at-risk groups (Agniel et al., 2019; Lyratzopolous et al., 2011). However, those methodologies require in depth data collected from patients across plans as a basis for measurement as opposed to trying to understand the plans themselves.

1.5 Key Questions

In this work, we propose applying clustering methodologies to reduce the dimensions of the data into more easily comparable groups. In this vein, we are interested in both methodological and content-based questions. Our first question of interest is whether clustering methods can be adapted to handle the inherently nested structure of health care benefits data to generate results that are useful to evaluating health equity in care accessibility for HIV prevention. We hypothesize that this algorithmic approach to auditing will reveal useful patterns in healthcare data. This leads into our content-based questions — firstly, what conclusions can we draw about how health insurance companies handle preventative care for HIV and secondly, where in the United States does healthcare plan design propose a barrier to the uptake of preventative services for people at risk of HIV?

2 Data & Methods

2.1 Data Source

This study uses data from the Health Insurance Compare (HIX Compare) database created and maintained by the Robert Wood Johnson Foundation (https://hixcompare.org/). For these analyses, we utilize the 2019 individual state market place files. This data contains plan design and benefit details for all plans offered in the United States through ACA-compliant state-wide marketplaces in 2019. We linked the HIX Compare data with 2019 plan-level formulary data from Vericred (https://vericred.com/) to obtain PrEP coverage details for each plan.

The HIX Compare 2019 dataset included 17,061 unique plans with complete administrative data (Deductible, Maximum out of Pocket, and Premium). Following convention established by the Robert Wood Johnson Foundation in the HIX compare data documentation, a unique plan is defined as a plan (1) with a unique set of benefits, (2) offered in a specific rating area (3) at a specific premium that is (4) not a cost share reduction or child-only derivative of another unique plan.[1] For the purposes of our study, we restrict our consideration of benefits to in-network plan characteristics.

[1] Cost Share Reductions (CSR) are plans that are available to certain individuals if their income subseeds a specific threshold based on the federal poverty line. CSRs are not unique plans because they simply reduce the patient contribution by a specific percentage once the benefit is applied as opposed to altering the benefits themselves. For low income individuals, cost sharing, premiums, and out of pocket maximums will be proportionally reduced.

2.2 Variable Selection

When selecting variables for this analysis, we consider two levels of inclusion criteria: (1) What benefits correspond to the preventative care necessary for individuals at risk for HIV? And (2) what variables quantify the mechanisms insurance companies use to regulate those benefits?

With regards to the first question, there are two key categories (1) HIV-specific Preventative Care Benefits and (2) General Health Care Benefits — like emergency care and primary care — that support continued physical, emotional, and financial health for people at risk of HIV. Within the second, there are three key health insurance mechanisms that regulate access and benefits: (1) Plan Maintenance Costs – the factors associated with having a health insurance plan, (2) Cost Sharing – a structure that regulates how much out of pocket patients pay for their benefits, and (3) Benefit-specific mechanisms such as prior authorization for preventative medications. This section first presents Plan Maintenance Costs and Cost Sharing regulatory factors and then details HIV-specific and general health care benefits, discussing benefit-specific mechanisms alongside the benefits they regulate.

2.2.1 Plan Maintenance Costs

2.2.1.1 Premium

Premiums are the monthly payments individuals make to acquire and maintain a healthcare plan. It is the base level amount one pays per year for health insurance. The HIX Compare dataset records estimated premiums for individuals of specific ages and family arrangements. For the purposes of this project focusing on individuals at risk of acquiring HIV, we include

estimated premiums for single individuals aged 27 given the largest percentage of new infections in 2018 occurred among the 25 to 29 age group (CDC, 2021).

2.2.1.2 Deductible

A deductible is the amount that a plan requires an individual to pay before the insurance plan itself begins to pay; although, oftentimes some services such as yearly check-ups with a primary care provider do not require first meeting the deductible. Some plans have a total deductible for both Drug and Medical services while other plans have separate deductibles for Drug and Medical services. For the purposes of these analyses, we sum Drug and Medical services deductibles to create a total deductible for plans that make such a distinction.

2.2.1.3 Maximum Out of Pocket Cost

The maximum out of pocket cost (MOOP) is a cumulative limit set by an insurance company after which an individual will no longer pay for their health coverage. Payments toward deductibles, copays, and co-insurances contribute toward the MOOP; however, payments toward premiums, out of network care, or services that are not covered by the insurance plan are not included in the total for MOOP. Some plans have a total MOOP for both Drug and Medical services while other plans have separate MOOP for Drug and Medical services. For the purposes of these analyses, we sum Drug and Medical services MOOP to create a total MOOP for plans that make such a distinction.

2.2.2 Cost Sharing

As discussed in Section 1.3.1, cost sharing is an important mechanism that health insurance organizations utilize to guide patient behavior and restrict access to care. Coinsurances psychologically and financially disincentivize care usage whereas copays provide more predictable payment schemes and generally lower financial burdens. Within our dataset, each benefit type (e.g. emergency care or specialty drugs) records the cost sharing structure in four key variables. For each benefit, there are two variables recording the coinsurance structure and two variables recording the copay structure. In the first of each of those variables, the data record an indicator variable describing the structure – whether the cost sharing applies before the deductible, after the deductible, always, never, or if there is just no charge for the benefit. The second column contains the cost information for the benefit – a fixed dollar amount for copays and a percentage for coinsurances. For each benefit either the two variables for copay or the two variables for coinsurance will be filled. No benefits are covered using both coinsurance and copay.

2.2.3 Plan Tiering

For the purposes of our analysis, the presence of tiered benefit structure represents a restrictiveness in the plan's network and is important to account for. However, less than 10% of plans in our dataset are multi-tiered. To account for that restrictiveness while also seeking a parsimonious set of variables, we include an indicator variable to denote when a plan has multi-tiered benefit structures, but do not include the full set of Tier 2 benefit details.

2.2.4 Preventative Care for HIV

A key component of the Ending the Epidemic Plan is to prevent over 250,000 new HIV infections over the next 10 years. This means that facilitating access to preventative care should be a key aim of health care plans available to those at increased risk for HIV. The *United States Preventative Services Task Force* strongly recommends the following for the prevention of HIV (Owens et al., 2019):

- Usage of Pre-Exposure Prophylaxis including:
    - Kidney function testing
    - Serologic testing for Hepatitis B & C
    - Routine STI Testing
    - Quarterly HIV CD4 count and Viral Load screens
    - Regular pregnancy screenings
- Behavioral counseling

The Ryan White HIV/AIDS program (https://hab.hrsa.gov/) adds the following into its care recommendations:

- Quarterly visits with a primary care provider or infectious disease specialist
- Mental health screenings
- Addiction screening and treatment

The following discusses how these components operationalize onto variables within the HIX compare data.

2.2.4.1 PrEP

As discussed in Section 1.1.2, access to PrEP is one of the most important considerations in the accessibility of HIV preventative care. As referenced in Section 1.3, prior authorization, specialty

tiering, and cost-sharing are all relevant to PrEP accessibility. Tiering status, prior authorization and cost sharing information are all available in the Vericred formulary data linked to the HIX compare individual plan data.

2.2.4.1.1 Cost Sharing Considerations for PrEP

There are a number of external funding sources that will cover the expected patient out of pocket costs for PrEP. These include the Gilead Advancing Access program which offers individuals up to $7200 in payments, the Patient Advocate Fund which offers up to $7500 per year, and State PrEP Assistance programs (NASTAD, 2020). Because of these options, individuals may never fully pay for PrEP themselves which mitigates, to some degree, the burden of the dollar value of the cost-sharing. That said, we still include an indicator of coinsurance versus copay to account for the ways in which uncertainty of pricing may affect patient behavior as discussed in Section 1.3.1.

2.2.4.1.2 PrEP-Specific Labs

The below listed tests are all covered under diagnostic testing benefits in the HIX compare dataset.

- Kidney function testing: One potential, although limited, side effect of PrEP is risk of kidney damage. As a precaution, kidney function is regularly monitored among those taking PrEP (Mocroft & Ryom, 2016).
- Serologic testing for hepatitis B and C virus: People at higher risk of HIV are also at elevated risk for contracting Hep B and Hep C. Monitoring these is part of the continuum of care.

- **Testing for other STIs:** Given those who use PrEP are protected from HIV infection, researchers have observed higher incidence of condomless sex among MSM on PrEP which may lead to higher rates of other STIs (Liu et al.l 2016; Barriero, 2018).
- **Pregnancy testing:** Individuals who are pregnant should not continue PrEP treatment.

Because of US Preventative Service Taskforce gradings guidelines, every year, the first of each type of diagnostic test an individual receives may be covered under preventative care benefits and, by regulation, have a $0 Copay (www.uspreventiveservicestaskforce.org/).

2.2.4.2 Routine HIV Labs

HIV preventative care requires continual monitoring of HIV status through regular HIV labs. These are also covered under Diagnostic Testing Benefits.

2.2.4.3 Behavioral Counseling

The US Preventative service and CDC standards of care recommend behavioral counseling for individuals at higher risk of HIV to think through behavioral modifications to limit risk if possible. These services are covered under outpatient mental health benefits in the HIX compare data set

2.2.4.4 Quarterly Visits with Primary Care and Infectious Disease Specialists

Preventative care for HIV requires coordination and consistent check in with individuals who manage patients' holistic health and those with specialized knowledge of HIV. Coverage for these services will be recorded in primary care and specialist benefits. Qualitative work also suggests that full time assistors can increase the likelihood of successful care navigation and lowered

costs for PLWH (McManus et al., 2020). PCP coordination is related to higher PrEP usage, therefore health insurance must make PCPs accessible (Silapaswan, Krakower, & Mayer, 2017

2.2.4.5 Mental Health Screenings

There are noted correlations between depression, anxiety, and elevated risk of HIV given experiences of stigma (Vanable et al., 2006). Mental health screenings should be covered under outpatient mental health benefits.

2.2.4.6 Addiction Screening and Treatment

People who inject drugs (PWID) are one population at heightened risk of HIV. Addiction screening and services are an important set of benefits to help mitigate risks for this population in specific. Coverage for these services should be covered under outpatient substance benefits.

2.2.5 General Benefits

In addition to HIV-prevention specific services, individuals at risk of HIV also require general health benefits. Beyond helping individuals care for their general well-being, this mitigates unforeseen financial risks which could impact their ability to continue with preventative care (Herman, Rissi, & Walsh 2011; Kalousova & Burgard, 2013). In 2012, the US government defined the following additional essential health care benefits that are relevant to our population of interest (Ford & Spicer, 2012):

- Emergency services: These are encoded in the HIX compare data as ambulance & emergency care benefits.

- Hospitalization: Coverage for these services are encoded as inpatient physician care benefits within HIX compare.
- Maternity and newborn care: These services are encoded in Inpatient Birth coverage.
- Prescription drugs: These benefits are encoded in HIX compare benefits covering generic drugs, preferred drug, non-preferred drugs, and specialty drugs.
- Rehabilitative and habilitative services and devices: These are included in our data set through benefits on habilitation services.

2.3 Clustering Methods

Clustering is an unsupervised machine learning technique that finds groups of similar data points within a larger set. These groups are called clusters. Not only do clustering algorithms group similar data points, but they also optimize the distance between dissimilar data points, placing them in different clusters. A good clustering algorithm should generate clusters with small within-cluster variance and large between-clusters variance. Clustering is content-agnostic, meaning that the clustering algorithm itself will only create groups of similar data points based on the data inputted into the algorithm. It is up to the individuals using the clustering algorithm to qualitatively inspect and name the groupings after they have been created. That said, while it is necessary to work with content experts to make sense of clustering output, clustering algorithms can help scientists discover underlying groups within large data sets with many variables and complex variable interactions (Romesburg, 2004).

Clustering is particularly useful for this healthcare study where we have over 17,000 plans, each with its own benefit structure. Historically, clustering has been used in medical research to identify patient risk groups and patterns in patient healthcare utilization (McLachlan, 1992; Liao et al., 2016; Lefèvre et al.). This work shifts the typical paradigm of clustering at the patient level to implement clustering at the plan level. This takes the focus off of patient actions and places scrutiny on insurance company choices. That said, while the unit of analysis is shifted, finding groupings in the types of plans healthcare companies offer to assess equity in their geographic distribution is a natural extension from prior work conducted on individual behavior.

2.3.1 Components of Clustering

There are two steps to conducting a rigorous cluster analysis: (1) calculating pairwise similarity (often distance) between observations and (2) creating clusters based on the similarity matrix. While some programmatic implementation of clustering algorithms group both steps into one procedure — such as the implementation of K-means clustering in R through the kmeans function — this standardizes important decisions which affect the resultant clusters. Recognizing that choice of distance measure is just as influential as choice of clustering algorithm results in a more thoroughly defensible final product (Shirkhorshidi et al., 2015).

2.4 Heterogeneous Distance Functions

Most textbook introductions to clustering analysis rely on the K-Means algorithm which iteratively calculates means of possible clusters until it minimizes the distance between points and cluster centers and maximizes the distance between points and centers of other clusters. The two often unstated but implicitly understood assumptions of K-means, however, are that (1) all of the input data is continuous and (2) distances are calculated using the Euclidian distance function (Romesburg, 2004).

But, the health plan data at hand is not entirely continuous. Each benefit has a coinsurance or copay distinguisher. PrEP can be listed on one of four drug tiers. As such, we cannot calculate a Euclidian distance between adjacent points. Heterogeneous distance functions were designed to bridge this very gap.

A heterogeneous distance function is an algorithm that handles each class of variable — whether it be qualitative, ordinal, or quantitative — with its own function. They frequently are written as piece-wise functions with specific transformations specified for each variable type. Lexically, where quantitative measures of relation tend to be referred to as "distance," heterogeneous measures, which do not translate into a physical coordinate representation of proximity as directly, are termed "similarity" measures.

Heterogenous distance metrics that handle quantitative continuous and qualitative variables are typically comprised of two parts: (1) a numeric distance metric and (2) a qualitative similarity classifier. As Wilson and Martinez (1997) outline in their seminal paper on heterogeneous distance metrics, there are myriad choices of similarity metrics for quantitative variables ranging from the canonical Euclidian distance to more data-reliant methods such as

Mahalanobis' distance. One of the most utilized heterogeneous metrics in instance-based learning literature, the Heterogeneous Euclidian-Overlap Metric (HEOM), uses the Manhattan's distance, calculated as:

$$D(x, y) = \sum_{i=1}^{m} |x_i - y_i|.$$

The Manhattan's distance forces normalization of variables with respect to their range (Aha, Kibler & Albert, 1991; Aha, 1992; Giraud-Carrier & Martinez, 1994). This is in contrast to Euclid's distance which is frequently critiqued for giving preference to larger-scaled variables (Shirkhorshidi et al., 2015).

In regard to qualitative similarity classifiers, the HEOM uses a simple overlap metric where the distance contributed by the qualitative variable equals 0 if the variables are equal and 1 otherwise. However, this can often overweight the qualitative variables in the dataset relative to quantitative variables. As Spencer et al. (2010) explain, unequal qualitative variables automatically contribute a full 1 point dissimilarity whereas unequal quantitative variables must be at the opposites of their range to contribute that significant of a dissimilarity. As such, qualitative variables are frequently over-weighted within the HEOM framework. Wilson and Martinez (1997) extend this within an instance-based learning framework and suggest replacing the overlap metric with the Value Difference Metric (VDM) developed in Stanfill and Waltz (1986). The VDM uses the probability of qualitative similarity within the dataset and cluster of interest to assign a similarity metric between 0 and 1. While this has proven successful within the instance-based learning framework, there is less opportunity to apply the VDM more broadly given it assumes the usage of a training set with pre-defined clusters to establish the cluster probabilities in the subsequent analysis. While this is crucial to research within instance-based

learning, this does not extend to more exploratory analyses where the best option is still the 1/0 overlap function.

2.4.1 Gower's Distance

While there may not be a sufficiently more advantageous measure of qualitative similarity for exploratory analyses, there are other options to re-balance the over-weighting of qualitative variables in heterogeneous distance metrics. John Gower (1971), in his now seminal paper "A General Coefficient of Similarity and Some of its Properties" addressed this vary issue in his formulation of the Gower's distance metric.

Gower's distance, much like HEOM, uses the range-scaled Manhattan's distance for quantitative variables and the overlap function for qualitative variables. However, while HEOM calculates the final distance by taking the square root of the sum of squared individual distance calculations, Gower's distance uses a simple arithmetic mean of the individual distances.

Gower goes on to say that while the arithmetic mean is a simple and straight-forward formulation of his distance metric, it would be entirely reasonable to take a weighted average of individual distances based either on *a priori* theoretic considerations or based on considerations from the data itself. That said, despite the current popularity of Gower's distance, the usage of a non-equal weighting scheme has remained largely unexplored since Gower's article's publication. Writing almost 40 years after Gower first proposed his similarity metric, Pretchey and Gaston (2009) go so far as to say that "so little is known about appropriate or inappropriate trait weightings that further research seems appropriate, rather than outright rejection of any approach at this stage." Chae et al. (2006) do discuss the usage of weighting to balance the

contributions of categorical and numeric data, suggesting that instead of just assigning a value of 1 to dissimilar categorical variables, practitioners should consider weighting categorical variables according to the probability of similarity (i.e. a variable with two levels and a 50% chance of concordance might receive a weighting of .5). Montanari and Mignani (1994) also consider weighting, evaluating its importance in handle missing data. But, as van den Hoven (2015) notes, most articles using Gower's distance set the weighting scheme to one without explaining why, while others do not note the weighting scheme used. A quick review of popular clustering tutorials on the well-referenced practical data science blog "Towards Data Science" makes no reference of weighting in its few postings on clustering with heterogeneous distance metrics (Fillaire, 2018; Shendre, 2020). Weighting Gower's distance is underexplored in current data science practice.

2.4.2 Distance with Hierarchical Structure

One main reason for weighting that Gower explores within his own paper, but that remains unaddressed in further literature, is to account for hierarchical structures within the variables. Referencing Kendrick and Proctor's (1964) work on taxonomy, he suggests that weighting can be used to ensure that similarity or dissimilarity between second-level characteristics never overwhelms similarity or dissimilarity between primary taxonomic distinctions. He provides some numerical guidance on how to calculate these weights; however, he does not provide this guidance within the context of both quantitative and qualitative variables — only within the context of homogenous variable types, presumably qualitative.

That said, this thinking is highly relevant to health care plan data given the hierarchical structure of benefit schemes. Each benefit first notes whether it is covered by copay or by coinsurance and then notes the numeric value of that benefit. The primary distinction between plans is whether or not they both utilize copay or coinsurance given the psychological and behavioral differences in how clients respond to the different payment schemes. While the practical monetary implications noted by the numeric value are important, they do not overwhelm the delineation made by being in the copay versus coinsurance category. Because of this, we choose to maintain weighting values of one for both the benefit structure qualitative and quantitative variables. In the case that two plans are both covered by copay or are both covered by coinsurance, they will have 0 distance between them due to the qualitative variable. Then, the numeric cost variables will be compared. In the case of variables that are close together, $5-10 dollar copay difference or a coinsurance difference of 5-10%, the effect will be minimal. The two variables for the benefit will average out to be close to zero. However, if the numeric cost variables are maximally distant from each other and the distance contributed by the term is equal to 1, then the overall dissimilarity contributed by the benefit would average out to .50. This is sufficient to create distinctions between plans who cover the same benefit with the same copay or coinsurance strategy.

2.4.3 Copay and Coinsurance Hierarchies as Missing Data

Setting the weights to one for both the qualitative and quantitative variables within each benefit is sufficient to create hierarchical distinctions between plans that match on the primary qualitative variable. That said, it does not address the case where the primary qualitative variable

for a specific benefit is different between two plans. Given that copay value and coinsurance value are two different variables with different scales (dollars and percent, respectively), we can think about this case as a type of missing data problem. When a benefit is covered by a copay, the coinsurance value for that benefit will be missing, whereas when a benefit is covered by coinsurance, the copay value for that benefit will be missing. That said, this issue diverges from the concept of missingness in that the missing value should not theoretically exist —plans with copays should not have coinsurance values nor should plans with coinsurances have copay values. As such, methodologies such as imputation to create bounds are inappropriate for this study.

While Gower's distance does not have a mechanism for handling missing data, HEOM does address the issue by suggesting that comparisons with missing values should be set to 1, e.g. assumed to be maximally distant. But, setting both the copay and the coinsurance quantitative variables equal to one would unduly emphasize the differences between plans due to the difference in the primary qualitative variable. In other words, under the HEOM framework, a difference in the qualitative plan structure variable (i.e., one plan being coinsurance and the other being copay) would result in a total of 3 distance points contributed to the numerator and 3 distance points to the denominator, far outweighing potential similarities with other plans.

2.4.4 Selected Distance Metric

Therefore, this thesis will use Gower's proposed weighting variables for hierarchical data to adjust for missingness. In the case of qualitative dissimilarity in the primary variable, the weights

on the quantitative secondary variables will be set to 0 such that the entire benefit contributes only one point of dissimilarity to the numerator and one point to the denominator of the mean distance calculation. In cases where there is a qualitative similarity, the benefit will contribute zero points to the numerator for the qualitative variable and a decimal number of points between 0 and 1 for the quantitative distinction. It will contribute two points overall to the denominator. This approach is supported by approaches to multi-state classification advanced in Romesburg (2004; 158) wherein dissimilarities in variables that could not possibly align between two observations — i.e. coinsurance and copay amounts between one copay using plan and one coinsurance using plan — do not contribute to the denominator of the qualitative similarity coefficient.

We operationalize a Gower's Distance for nested data with a dependency-related missingness pattern as follows:

$$d_{i,j} = \begin{cases} overlap(x_{i,j}, y_{i,j}) & where\ j\ is\ qualitative \\ \frac{|x_{i,j} - y_{i,j}|}{range(j)} & where\ j\ is\ continuous \\ 1 & where\ x_{i,j}\ or\ y_{i,j}\ is\ NA \end{cases}$$

$$w_{i,j} = \begin{cases} 1 & where\ j\ is\ qualitative \\ 1 & where\ j\ is\ continuous \\ 0 & where\ x_{i,j}\ or\ y_{i,j}\ is\ NA \end{cases}$$

$$D_i = \frac{\sum_{j=1}^{u} w_{i,j} d_i, j}{\sum_{j=1}^{n} w_{i,j}}.$$

2.5 Clustering Algorithms

Once we have a pair-wise distance matrix, the next step is to create clusters of plans that are similar in distance to each other but dissimilar to plans in other clusters. There are two main types of clustering algorithms which we will test in our analysis, hierarchical and non-hierarchical. For each, we must consider both how the clusters are created and how we can assess the fit of the clusters, including how to choose an optimal number.

2.5.1 Hierarchical Clustering

Hierarchical clustering methods stem from numerical approaches to taxonomy and are the context within which Gower first developed his distance algorithm (Felsenstein, 2013). Hierarchical methods allow practitioners to view groups within groups and to get a sense practically for where different groups may diverge from each other — they also only rely on pairwise distance metric relationships between observations and do not place other restrictions or requirements on the data (Alpaydin, 2020). Hierarchical methods are useful as well because they provide a degree of "predicted distance" between observations by way of tracing paths on the tree from one observation to another which provides a more concrete basis to evaluate the fit of the cluster model than in other clustering methods (Farris, 1969).

There are two types of hierarchical clustering method: agglomerative and divisive. In agglomerative clustering, smaller groups are joined together based on similarity criterion to form larger groups. In divisive clustering, larger groups are divided in half into possible smaller groups. Each possible split is compared using a chosen metric, be it cluster variance or distinction between clusters, and the split which optimizes that metric is chosen before those groups are

split once again (Roux, 2018). Generally, researchers prefer to use agglomerative clustering for large data sets like the one in question because of the outsized computational load of divisive clustering which traditionally requires fitting many possible splits for each decision point (Sasirekha & Baby, 2013). In this work, we restrict our methods to agglomerative clustering.

When conducting agglomerative clustering, researchers must choose a linkage method to determine which clusters to iteratively merge. These linkage metrics are usually a function of the distances between objects within the clusters proposed to be joined. There are three main distance-based metrics: average linkage, single linkage, and complete linkage. Average linkage calculates the average distance between all objects in the clusters proposed to be merged (Sokal & Michener, 1958), single linkage calculates the distance between the closest two objects in the clusters proposed to be merged (Legendre & Legendre, 1998), and complete linkage calculates the distance between the two furthest objects in the clusters proposed to be merged (Sorensen, 1948). The linkage metric is calculated for all proposed cluster merges and the two clusters with the lowest value of the metric are merged. In practice, complete linkage is a conservative but strong method to detect clusters because it connects clusters based on the maximum distance between them; it is also one of the oldest methods in hierarchical clustering. In contrast, single linkage represents an alternative extreme in that it can pull together two observations that are more distant from each other but that are both close to a third observation, uncovering relationships that are important but perhaps less intuitive. Average linkage is one of the most frequently used methods because it balances the conservative approach of complete linkage with the discovery capacity of single linkage (Romesburg, 2004; Roux, 2018).

There is a second branch of thought within hierarchical clustering — one where all possible cluster joins are performed and the result is chosen based on optimizing a specific function. Ward (1963) suggests joining clusters to minimize within-cluster variance and optimize between-cluster variance. This method, known commonly today as “Ward’s Method” has been validated based on a number of underlying distance metrics (Strauss & von Maltitz, 2017). While it does not directly model the underlying distance matrix, Ward’s method is a viable method to discover well-defined clusters.

2.5.1.1 Assessing Fit

Clustering is an unsupervised machine learning method and, as such, assessing fit of specific clusters can be difficult since its aim is to discover relationships that we did not previously know existed. That said, with hierarchical clustering methods, we can assess the fit of the hierarchical tree to the underlying distance matrix. In doing so, we answer the question, “do the distances between observations when modelled by the tree match the distances between observations in the distance matrix?”

To accomplish this, we need to define the distances between measures predicted by the hierarchical clustering tree. In taxonomy, the most used measure is called the cophenetic distance. It is defined as the height in the clustering tree where the branches connecting two observations connect for the first time (Sneath & Sokal, 1973). Using this, we can create a predicted distance matrix of cophenetic distances.

2.5.1.2 Cophenetic Correlation Coefficient

From the predicted distance matrix, we can calculate a simple correlation between the original distance matrix and the cophenetic distance matrix. This is known as the Cophenetic Correlation Coefficient (CPCC) (Farris, 1969; Sokal & Rohlf, 1962) — a 0 to 1 metric that describes the extent to which the calculated distances between pairwise data points and the predicted distances between pairwise data points on the dendrogram represent similar variation.

2.5.1.3 Index of Agreement

Taking advantage of the cophenetic distance matrix as a prediction for the original distance matrix, we can also calculate Willmot's Index of Agreement (IoA), a coefficient reminiscent of the Average Root Mean Squared Error (RMSEA) traditionally used in supervised machine learning methods. While the CPCC calculates the extent to which the predictions in the model are associated with the underlying data, the IoA, as Willmot (1981) puts it, is *a measure of the degree to which a model's predictions are error free*. We calculate the IoA for a hierarchical model as follows:

$$d = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (|P_i^*| + |O_i^*|)^2},$$

where P_i is the cophenetic distance, O_i is the observed distance, $P^*_i = P_i - \bar{O}$ and $O^*_i = O_i - \bar{O}$.

Using these fit assessment indices, we can directly compare different methods of growing clustering dendrograms based on the same underlying distance matrix. That said, one drawback of the Index of Agreement is its emphasis on exactly matching the underlying distances to the cophenetic tree-modelled distances. In comparing tree-based methods it is

important to review the underlying data to examine the extent to which matching the magnitude of distances is important as compared to simply grouping together plans that display similarities.

2.5.1.4 Choosing Clusters

Once we have grown a dendrogram, we have to decide into how many clusters to cut the structure. As an unsupervised method, we get to make this decision based both on theory and usability of results in addition to quantitative metrics that validate cluster quality and existence. For this analysis, we use a metric called silhouette distance to quantify the quality of the clusters formed. The silhouette distance for observation i, as defined by Rousseeuw (1986) is calculated as follows:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)},$$

where a_i is the average of the distances between observation *i* and all other observations within its own cluster and b_i is the average of the distances between observation *i* and all observations in the closest cluster. Here, we can see that s_i will range between -1 and 1, being maximized when the within cluster distances are small relative to the between-cluster distances and minimized when the between cluster distances are closer to the within cluster differences. A positive *s* indicates that observations have been assigned to the proper clusters whereas a negative *s* indicates that an observation is likely assigned to the wrong cluster. We can take the average across all s_i to get an average silhouette metric which summarizes the overall quality of the clusters. Choosing the number of clusters that maximizes the silhouette distance is a common way to decide how to cut a hierarchical cluster dendrogram, particularly in medical

research (Clifford et al., 2011). However, it is equally important to consider the contrasts between clusters on the underlying data as well.

2.5.2 Non-Hierarchical Clustering and Partitioning Around Medoids

While hierarchical clustering methods assume para– and sub–relationships between and within clusters, non-hierarchical methods model the data as distinct clusters without a hierarchy of relationships. Since we cannot assume too much about the data generating process by which insurance companies generate plans, it is worthwhile to compare our hierarchical models against a model with fewer assumptions about data structure.

There are a number of non-hierarchical partitioning methods, each with their own set of assumptions. For this analysis, we compare our hierarchical clustering against the Partitioning Around Medoids (PAM) algorithm advanced by Theodoridis and Koutroumbas (2006: 635) because it is more robust to noise and to outliers than other methods like the k-means algorithm and because it can accept a pre-created distance matrix to cluster mixed-type data (Romesburg, 2004). PAM works iteratively by randomly selecting a medoid for each cluster, assigning each data point to the closest medoid, selecting another point within the cluster and swapping it with the medoid to see if it improves fit as defined by average distance, and then continuing the process until no more improvements can be made. Unfortunately, PAM is analytically costly and the number of clusters (k) must be specified beforehand. Because of this, there is no way to find the optimal number of clusters without running the full clustering process a number of times across a range of k.

Since non-hierarchical methods do not create a dendrogram structure to model the distances between points, we do not have predicted distances to calculate CPCC and IoA statistics. Because of this, we can only compare clusters created by PAM to hierarchical models using the average silhouette.

2.5.3 Choosing a Final Clustering Model

We follow a two-step process to choose our final clustering model. First, we fit three distance-based hierarchical models, one variance-based model using Ward's method, and one non-hierarchical model using the partitioning around medoids algorithm. We evaluate these models using cluster fit statistics and select a subset for further consideration. Second, we investigate the practical implications of each solution by comparing contrasts between clusters on the actual dataset of interest. We select a final solution on the basis both of cluster quality and on usefulness for practical interpretation. As a reference, the methods section Table 1, on the next page, overviews the clustering algorithms employed in this analysis including their strengths and weaknesses.

Methods Table 1: Clustering Methodologies

Method	Procedure	Strengths	Weaknesses
Hierarchical Clustering	Iteratively joins small clusters to form larger clusters on the basis of a defined metric.	Creates an interpretable hierarchical tree.	An optimal join at a lower level may introduce bias later on.
Single Linkage	Joins two clusters based on the distance between the most similar points between the two clusters.	Can uncover less-intuitive structures by linking two points that are both similar to a third point.	Can result in chaining. Prioritizes similarity of individual points over homogeneity of the cluster.
Complete Linkage	Joins two clusters based on the distance between the least similar points between the two clusters.	Is robust to spurious similarities between clusters and priorities within cluster homogeneity.	May avoid needed linkages due to outliers on cluster edge.
Average Linkage	Joins clusters based on average distance between all points in the two clusters.	More robust to spurious similarities and outliers on the edge.	Regression to the mean creates similarities on average but does not ensure similarity of individual points.
Ward's Method	Joins two clusters based on which join will minimize the within cluster variance of the new cluster.	Prioritizes cluster homogeneity.	Computationally demanding. Does not prioritize matching the cophenetic distances to the magnitudes in the distance matrix.
Non-Hierarchical Clustering	User sets number of clusters. Algorithm creates optimal assignment.	Does not assume hierarchy to groupings. Directly estimates desired number of clusters.	Clustering process is not interpretable.
Partitioning Around Medoids	Generates *n* cluster centers. Sorts points to minimize distance to cluster centers. Swaps cluster center for random point in cluster. Re-sorts points to minimize within cluster variance. Repeats until a minimum in the cost function is identified.	Requires few assumptions and creates cluster homogeneity with respect to the defined distance matrix.	Computationally demanding. Does not create subgroupings.

3 Results

3.1 Descriptive Statistics

The HIX Compare 2019 dataset included 17,061 unique plans with complete administrative data (Deductible, Maximum out of Pocket, and Premium). Monthly premiums ranged from less than $325/month for the lowest quartile to above $501/month for the highest quartile. There was considerable variance in deductible prices which ranged between $2000 and $6500 for the middle 50% of plans. Overall, individuals enrolled in the middle 50% of plans were expected to pay a maximum out of pocket cost of $6500 to $7900 per year for their care in addition to their monthly premium costs and any out-of-network care (see Table 1A). Ten percent of plans guided enrollees to specific care providers using multi-tiered structures.

Table 1A: Plan Characteristics: Administrative Variables

	Q1	Median	Q3
Monetary Variables			
Deductible	2000	4500	6500
Maximum Out of Pocket	6500	7350	7900
Premium/Month (Age 27)	325	411	501
	n	**%**	
Plan Factors			
Contains 2+ Tiers	1788	10.48%	

3.1.1 HIV Prevention Specific Benefits

3.1.1.1 Pre-Exposure Prophylaxis

Almost all plans covered Pre-Exposure Prophylaxis (98.75%). Thirty-two percent did so with coinsurance while 66.2% covered PrEP with a copay. Nineteen percent of plans required

enrollees to obtain prior authorization before they could utilize PrEP and 21.5% restricted access by placing PrEP on a specialty tier.

3.1.1.2 Access to care

In terms of accessing the care services necessary to for HIV preventative, 52% of plans allowed enrollees to pay for Diagnostic Tests with a copay; 54.7% of those copays equaled $0. Seventy-seven percent of plans allowed enrollees to pay for outpatient mental healthcare with copays. For 75% of plans, these payments were less than $40. Eighty-seven percent of plans allowed patients to see a primary care physician using a copay, 75% of which charged $35 or less. Plans were slightly more restrictive about allowing enrollees to see specialists such as infectious disease doctors. Seventy-eight percent provided that coverage through a copay, but 50% of those copays were over $50. Finally, all plans used $0 copays to cover routine preventative services such as routine blood tests. Variables related to routine preventative services were dropped from consideration in all analyses as they did not contribute any variance between plans.

Table 1B: Plan Characteristics — PrEP Variables

	Number of Plans	**Percent**
Coverage		
Covered	16847	98.75
Not Covered	214	1.25
Coverage Strategy		
Coins	5552	32.54
Copay	11295	66.2
Not Covered	214	1.25
Prior Authorization		
No PA	13610	79.77
Requires PA	3237	18.97
Not Covered	214	1.25
Tier		
Non Preferred Brand	791	4.64

Non Preferred Specialty	3674	21.53
Preferred Brand	12382	72.57
Not Covered	214	1.25

3.1.2 General Benefits

As discussed in Section 2.2.5, individuals at higher risk of HIV infection also need general health benefits to ensure financial and general well-being. This facilitates better adherence to preventative treatment regimens. Emergency care and the accompanying costs of unexpected illness have the greatest potential to create unforeseen expense or disability. Comprehensive emergency care includes coverage of Ambulance, Emergency Room, and Inpatient Physician services. Overall, we find that plans offering copays for the aforementioned services (AB: 45.7%, ER: 65.6%, IH: 46.8%) had lower predicted costs of care given that at least 50%, 25%, and 75% of Ambulance, ER, and Inpatient Physician copays, respectively, were free.

Table 1C: Plan Characteristics — Benefits

Benefit	Number of Plans	Percent	Q1	Median	Q3
Ambulance (AB)					
Not Covered	190	1.11			
Coins	9047	53.03	20%	30%	40%
Copay	7824	45.86	$0	$0	$150
Diagnostic Tests (DT)					
Not Covered	3	0.02			
Coins	8178	47.93	20%	30%	40%
Copay	8880	52.05	$0	$0	$50
Emergency Room (ER)					
Not Covered	0	0			
Coins	5868	34.39	20%	30%	50%
Copay	11193	65.61	$0	$250	$350
Generic Drugs (GD)					
Not Covered	0	0			
Coins	1782	10.44	20%	30%	40%
Copay	15279	89.56	$5	$10	$18
Habilitation Services (HA)					
Not Covered	18	0.11			

Coins	7553	44.27	20%	30%	40%
Copay	9490	55.62	$0	$15	$40
Inpatient Birth (IB)					
Not Covered	2	0.01			
Coins	8518	49.93	20%	30%	40%
Copay	8541	50.06	$0	$0	$500
Inpatient Physician (IH)					
Not Covered	283	1.66			
Coins	8798	51.57	20%	30%	40%
Copay	7980	46.77	$0	$0	$0
Outpatient Mental Health (OM)					
Not Covered	2	0.01			
Coins	3824	22.41	20%	30%	40%
Copay	13235	77.57	$0	$25	$40
Primary Care Physician (PC)					
Not Covered	0	0			
Coins	2225	13.04	20%	35%	50%
Copay	14836	86.96	$10	$25	$35
Preventative Care (PV)					
Not Covered	0	0			
Coins	0	0	0%	0%	0%
Copay	17061	100	$0	$0	$0
Specialist (SP)					
Not Covered	0	0			
Coins	3745	21.95	20%	35%	50%
Copay	13316	78.05	$20	$50	$65

3.2 Selecting a Clustering Methodology

Having described our dataset, we evaluate the clusters generated by our distance matrix and clustering algorithm trials and choose a solution that helps us better understand healthcare offerings for people at risk of HIV. To choose the most useful set of clusters, we follow a two-step strategy of (1) comparing model fit statistics to select a subset of clusters that plausibly fit our data and (2) interpreting those optimal clusters in light of the data itself to select the most useful solution.

3.2.1 Preliminary Review of Model Fit Statistics

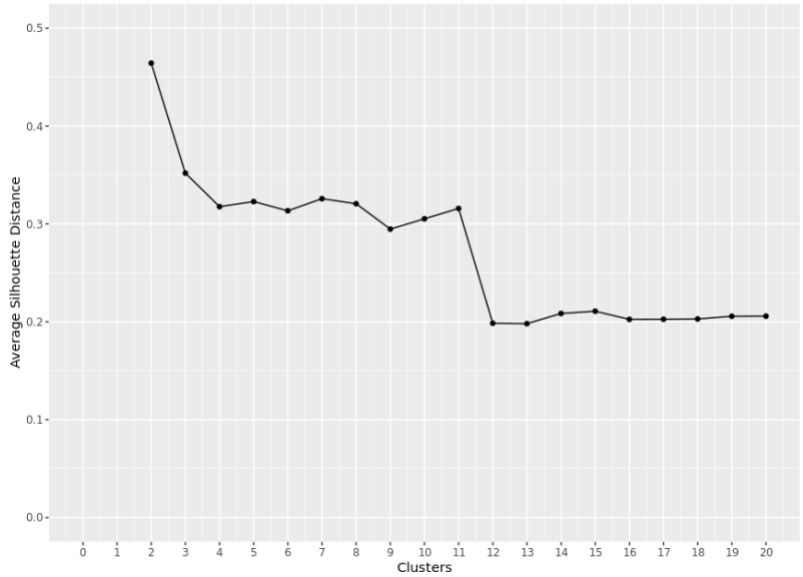
We review model fit for hierarchical models (complete linkage, single linkage, average linkage, and Ward's methods) first before reviewing the partitioning around medoids algorithm.

3.2.1.1 Complete Linkage

The cophenetic distances from our Complete Linkage hierarchical tree correlated with the original distance matrix at 0.734. This means that, on average, the complete linkage model separated plans that were distinct from each other and kept together observations that had lower distance values. The associated Index of Agreement (Willmott, 1982) was approximately .50, which means that the magnitude of the cophenetic distances themselves were well aligned to the original distance matrix. The complete linkage method uncovered two distinct clusters (Silhouette [see section 2.5.1.4] = .464); however, cutting the tree into more clusters did not create groups that held together as well (Silhouettes below .35). Based on these preliminary

statistics, we selected the two cluster complete linkage model for consideration in further analyses.

Figure 1: Average Silhouette Distances from Complete Linkage Method

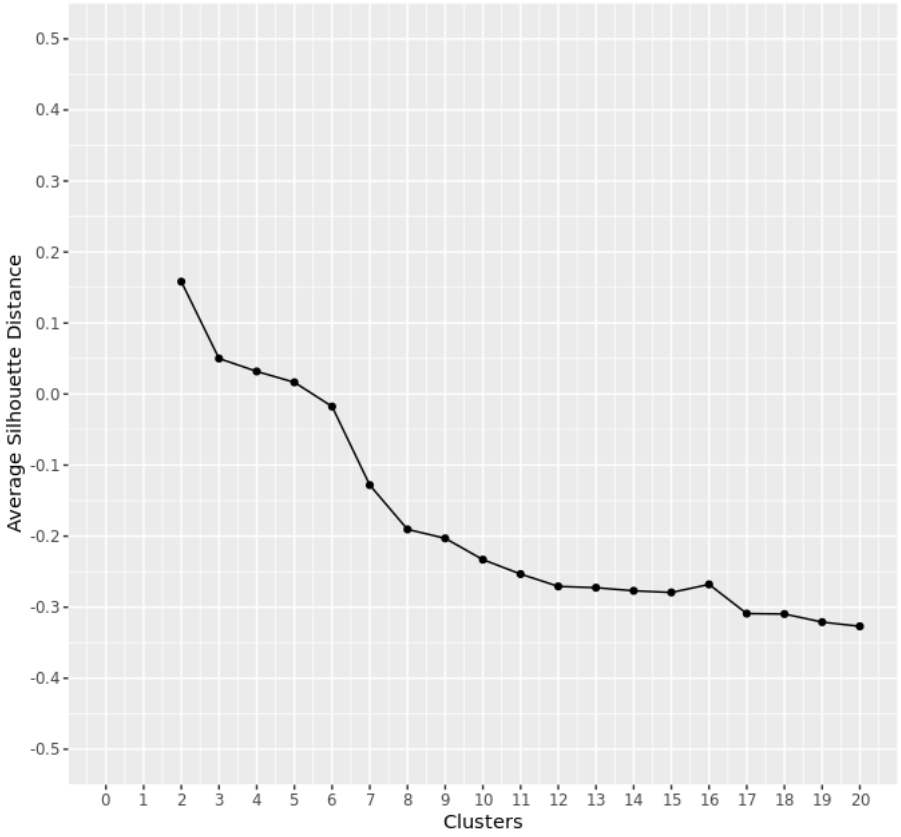


3.2.1.2 Single Linkage

The cophenetic distances from our Single Linkage hierarchical tree correlated with the original distance matrix at 0.333. This means that, in general, plans that were farther apart in the original distance matrix were also farther apart when modeled in the single linkage tree, but this trend is relatively weak. The Index of Agreement of 0.428 suggests a moderate amount of prediction error between the magnitude of the cophenetic distances and the computed distances. Furthermore, the associated plot of average silhouette widths displays weak cluster coherence and even misclassification as the number of clusters increases. Because of this, we drop results from the Single Linkage model from further consideration. Given the liberality with which the single linkage model makes connections — joining clusters by their two closest points — this is not a surprising result. A drawback of Single Linkage is that it can result in chaining as opposed

effective clustering. However, given its flexibility, single linkage has the capacity to bring together less intuitive clustering and so it was still worth exploring.

Figure 2: Average Silhouette Distances from Single Linkage Model

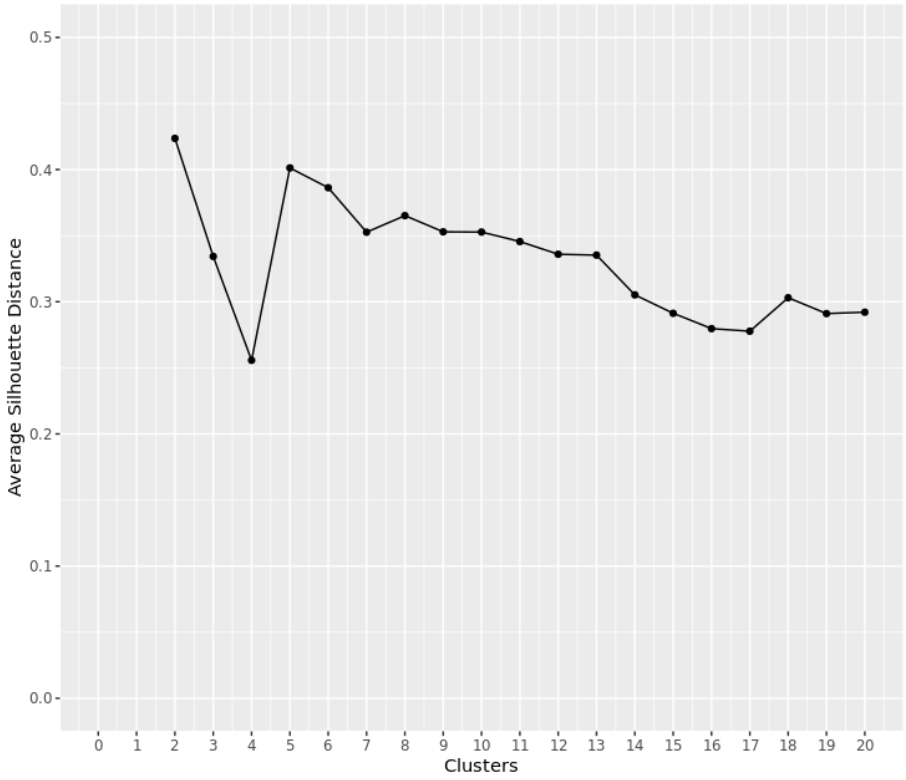


3.2.1.3 Average Linkage

The Average Linkage hierarchical model created cophenetic distances between plans that correlated with the actual distances at 0.781. Generally, plans that were farther apart in the distance matrix were also farther apart in the model. It also modeled the magnitude of those distances with high precision (Index of Agreement = 0.868). This accords with expectation, given that the average linkage method is robust to cluster outliers and optimizes tree concordance with the underlying matrix. However, as shown in the silhouette plot below, the clusters

themselves were not highly coherent. Despite this, the two cluster and the five cluster solutions showed reasonable definition (Silhouettes > .40) and were explored further in the second step.

Figure 3: Average Silhouette Distances from Average Linkage Model

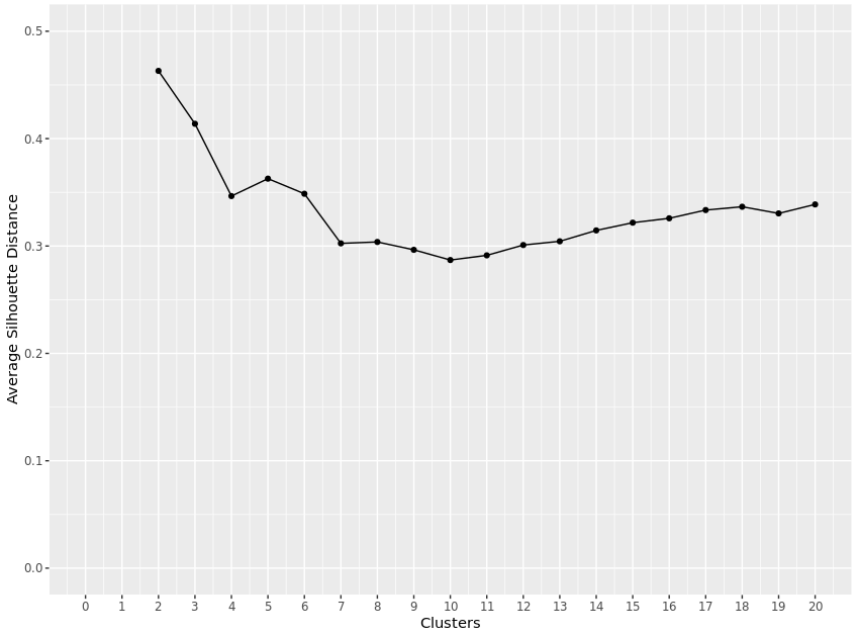


3.2.1.4 Ward's Method

Growing a hierarchical tree using Ward's Method of minimum variance created modelled distances between the points that were well correlated with the original distances (CPCC = 0.683). On average, points that were further apart in the distance matrix ended up further apart in the hierarchical tree. However, Ward's method was not highly successful at predicting the magnitude of those distances (Index of Agreement = 0.014). This result is somewhat unsurprising as Ward's method focuses on creating well-defined clusters with minimum within-cluster variance and not on matching the underlying distance matrix. This means that while it can

uncover a compelling cluster structure, it does not account for distances between clusters in the same way as average linkage. While the resultant clusters may be useful and points that are generally closer to each other may end up closer to each other on the tree, the modelled distances may not approximate the original distances in their magnitude. As shown in the average silhouette plot below, both the two cluster and three cluster solutions from the Ward's Linkage method were defensible solutions (Silhouette = .463 & .414, respectively). We leave it to our content-based investigation in step two to determine if the potential usefulness of these clusters overrides concerns about mismatch between the magnitudes of the pair-wise modelled distances and the originally calculated ones.

Figure 4: Average Silhouette Distances from Ward's Linkage Model

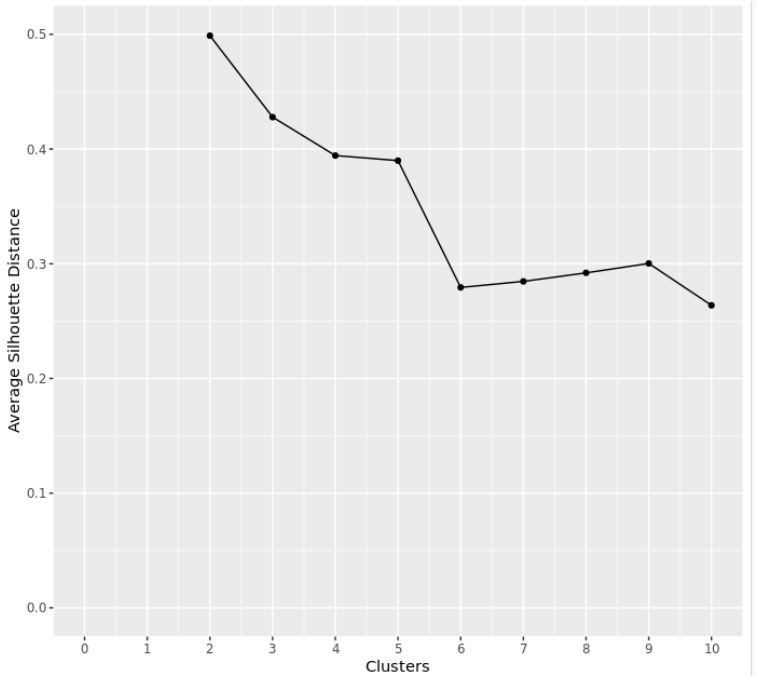


3.2.1.5 Partitioning Around Medoids

As the one non-hierarchical method considered, the Partitioning Around Medoids (PAM) algorithm does not generate a tree with cophenetic distances. Because of this, we can only consider the silhouettes of the resultant clusters as well as their interpretation within the data.

Similar to the clusters resulting from Ward's Method, both the two and three cluster solutions from the PAM presented potentially usable results (Silhouettes = .499 and .427, respectively).

Figure 5: Average Silhouette Distances from PAM Algorithm



3.2.2 Evaluating Model Fit Using Data

After selecting the following possible models: Complete Linkage 2 Clusters, Average Linkage 2 & 5 Clusters, Ward's Linkage 2 & 3 Clusters, PAM 2 & 3 Clusters, we evaluate the models based on their practical usefulness and apparent fit to the data. We first consider cluster size and then consider cluster contrasts on key variables of interest.

3.2.2.1 Cluster Size

Comparing resultant cluster sizes across methodologies, we immediately found that the 5 Cluster Average Linkage Model recovered three core clusters with two additional small clusters that,

collectively, only comprised 0.873% of the data. This suggests that the average silhouette widths reported for the 5 Cluster Average Linkage Model may be arbitrarily high due to over fitting. For this reason, we removed the Average Linkage 5 Cluster model from consideration. Overall, the PAM and the Ward's method models reported similar cluster sizes to each other. The clusters were reasonably balanced in the two cluster formulation and were composed of two equal-sized clusters and a third smaller cluster in the three cluster formulation. Both the Average and Complete Linkage 2 cluster models uncovered one larger cluster and one smaller cluster. These clusters were very unbalanced in the Average case (80% to 20%) and moderately unbalanced in the Complete case (62% to 38%).

Table 2: Cluster Sizes by Method

Method	Clusters	1	2	3	4	5
Average	2	13534	3527			
Average	5	6184	3527	7201	141	8
Complete	2	10562	6499			
Pam	2	8858	8203			
Pam	3	6322	2838	7901		
Ward	2	9894	7167			
Ward	3	6329	3565	7167		

3.2.2.2 Cluster Contrasts on Benefit Characteristics

In defining our distance algorithm, we prioritized distinctions between coinsurance and copay for each set of plan benefits based on considerations from theory. For this reason, we continued our evaluation by considering those distinction in coverage mechanisms. Figure 6 displays the distribution of coinsurance usage for each benefit by clustering method and cluster. Practically, clustering methods that create more distinct and useful clusters will have coinsurance percentages at the extremes — either closer to 100% or 0%. This is because our clusters are only

useful to the extent to which knowing cluster membership of a plan provides useful information about it. Probabilities of coinsurance near 0 or 1 provide more certainty that an individual plan in the cluster takes on a specific characteristic and reduce within-cluster variance, creating more useful separations. The first panel of Figure 6 displays the benefit distribution of all plans which provides a useful comparison for evaluating the relative restrictiveness of plans in each category.

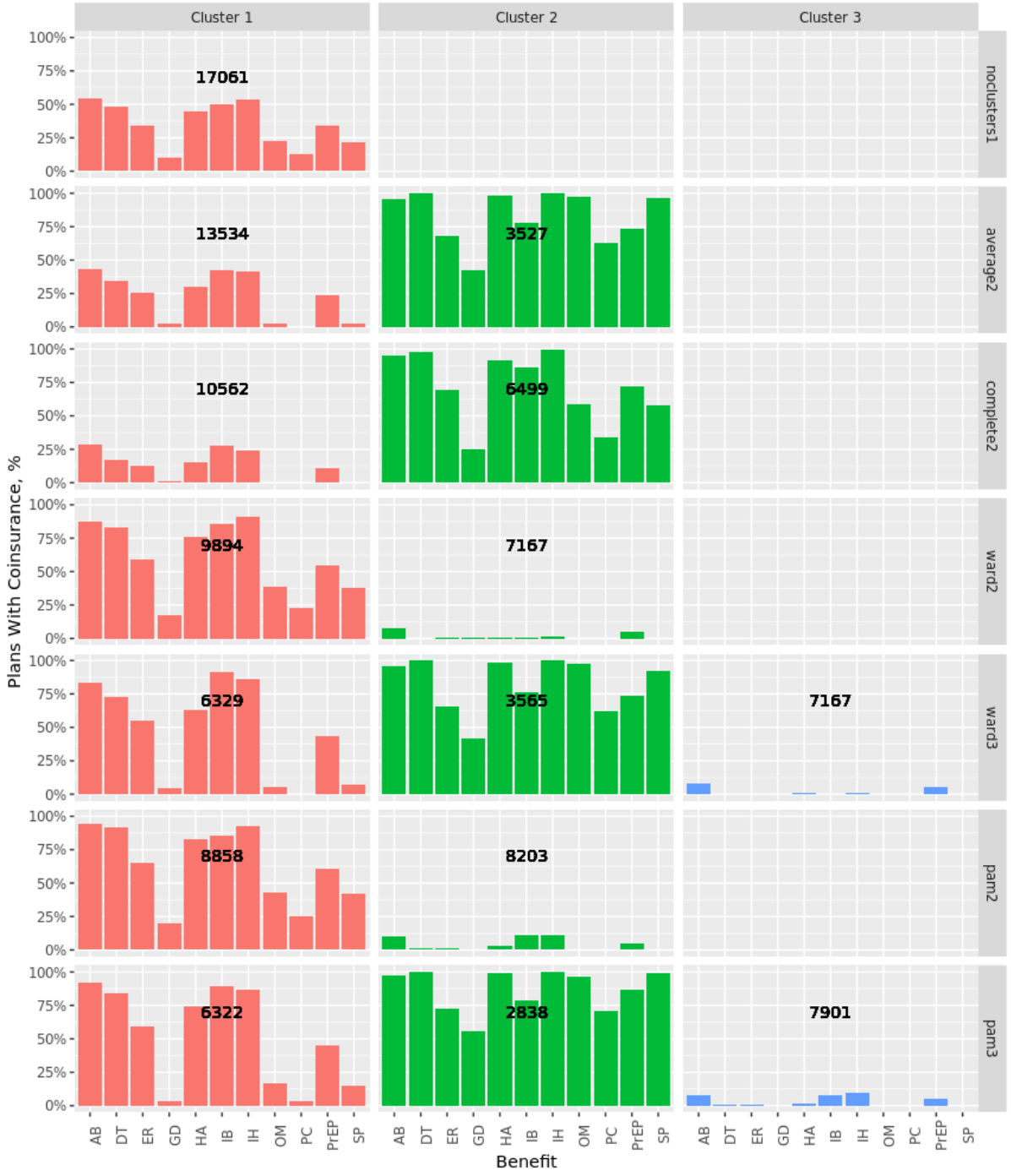
Comparing two-cluster models using Figure 6, we found similarities between the average and complete cluster methods as well as between Ward's method and the PAM algorithm. The average and complete cluster graphs show comparable bar graph shapes with spikes and dips in bars in similar places between them. That said, where the average linkage method placed some plans into Cluster 1 to generate higher coinsurance probabilities among benefits including Specialty Providers and Outpatient Mental Health, the complete linkage method achieved coinsurance probabilities closer to 0 in Cluster 1 by shifting plans into Cluster 2, increasing the variance of some of the coinsurance probabilities in Cluster 2.

The distinctions between the Ward's method and the PAM algorithm follows a similar pattern. While both clustering strategies revealed remarkably similar patterns, the PAM algorithm shifted a few plans over to Cluster 2, raising coinsurance probabilities of benefits in Cluster 1 — including Ambulance Care, Diagnostic Tests, and Habilitation Services — to be closer to 1. In contrast, the Ward's clustering method maintained coinsurance probabilities even closer to 0 among Cluster 2 — particularly for inpatient birth and inpatient physician services — while allowing the coinsurance probabilities among Cluster 1 to dip a bit closer toward .5. That said, the differences between the two clustering methods are subtle.

Overall, the Ward's and PAM two cluster methods created greater contrasts between the clusters than the average and complete methods. The Ward's and PAM methods isolated a cluster of plans that has almost no coinsurance while average and complete methods did not. In addition to lowering the within-cluster variance, functionally, being able to label a cluster as "least restrictive" across all benefits due to its lack of coinsurance is both useful and easily interpretable. Because of these patterns, we found Ward's method and PAM to be the most advantageous moving forward and removed the average and complete linkage methods from consideration.

Distinguishing between the two and three cluster solutions, we see from Figure 6 that the Ward's three cluster solution maintained its least-restrictive cluster (n = 7167 plans) and split the more restrictive cluster into two groups — a larger cluster (n = 6329) with increased access and decreased variance on Generic Drugs, Outpatient Mental Health, Primary Care, and Specialist Physicians and a smaller cluster (*n* = 3565) with coinsurance probabilities very close to 1 for Ambulance, Diagnostic Testing, Habilitation, Inpatient Hospital, Outpatient Mental Health, and Specialist benefits. The shifts between the two and three cluster solutions were very similar for the PAM algorithm. But, with the PAM, the least restrictive cluster did not stay constant between the two —there was a slight shift in probabilities, specifically among inpatient birth and habilitation services, toward 0.

Figure 6: Percent of Plans Requiring Coinsurance in Each Cluster



While we saw from the average silhouette distance plots (Figures 4 and 5) that adding an additional cluster to create a three cluster solution decreased the overall cohesion of the individual clusters themselves, this makes sense in light of the observed data. In both the PAM

and the Ward's three cluster solutions, the probability of coinsurance for a few benefits in some clusters hovered around .5 — such as Emergency Room (Clusters 1 & 2) and Generic Drugs (Cluster 2). For a few other benefits the probabilities remained similar between Clusters 1 and 2— such as Ambulance Care and Inpatient Physician care. This necessarily will increase the within-clusters variance while decreasing the between-clusters variance. That said, while there were some similarities between Clusters 1 and 2 for both PAM and Ward's methods, the three cluster solution still provided useful — and potentially ordinal — contrasts in terms of plan restrictiveness that the two cluster method does not.

Cluster 2, the most restrictive cluster, had the highest rates of coinsurance. While some rates stayed similar between Cluster 2 and Cluster 1, plans in Cluster 1 gained complete access to Copays on four key benefits — Generic Drugs, Outpatient Mental Health, Primary Care, and Specialty Providers. These are all relevant benefits for HIV preventative care. Moving to Cluster 3, almost all plans provided access to all benefits using copay. Because of these distinct tiers in copay percentage, we believe that, despite the lower silhouette coefficients, the three cluster solutions for both PAM and Ward's method was the most practically useful and had clear interpretative value. While the two cluster solutions did provide distinct contrasts, the three-cluster solution demonstrated the existence of a mid-level tier of restrictiveness that, while retaining some similarities with the most restrictive tier, was worth examining as its own distinct cluster.

Having settled on the three cluster model based on practical interpretative value, we had to decide between the PAM and the Ward's cluster method. As previously reported, the 3 cluster PAM and Ward's solutions had average silhouette distances of .427 and .414, respectively. Given

that silhouette metrics prioritize spherical-shaped clusters and may not be as robust to the nested structure and missingness patterns of our data (Lengyal & Botta-Dukát, 2019), this .13 difference is not a meaningful distinction. As such, we once again turned to considerations from the data itself. In the first cluster, Ward's method reported coinsurance likelihoods closer to 0 for Outpatient Mental Health, Primary Care, and Specialist Doctors. In Cluster 3, Ward's method reported coinsurance likelihoods closer to 0 for Inpatient Birth and Inpatient Physician care. Having these lower likelihoods, while not substantially different between the two groups, cements the ordinal nature of low restrictiveness, moderate restrictiveness, and high restrictiveness for Clusters 3, 1, and 2, respectively. For this reason, we believe the 3 Cluster solution using Ward's method has the greatest practical value for understanding healthcare plan groupings for preventative HIV care

3.3 Ward's 3 Cluster Solution

Our chosen solution creates three distinct clusters with three distinct and decreasing levels of restrictiveness of care.[2] Cluster 1, the most restrictive cluster, contains 20.9% (n = 3565) of the considered plans. Cluster 2, the moderately restrictive cluster, contains 37.1% (n = 6329) of the considered plans. Cluster 3, the least restrictive cluster, contains 42% (n = 7167) of the considered plans.

3.3.1 Comparing Benefits

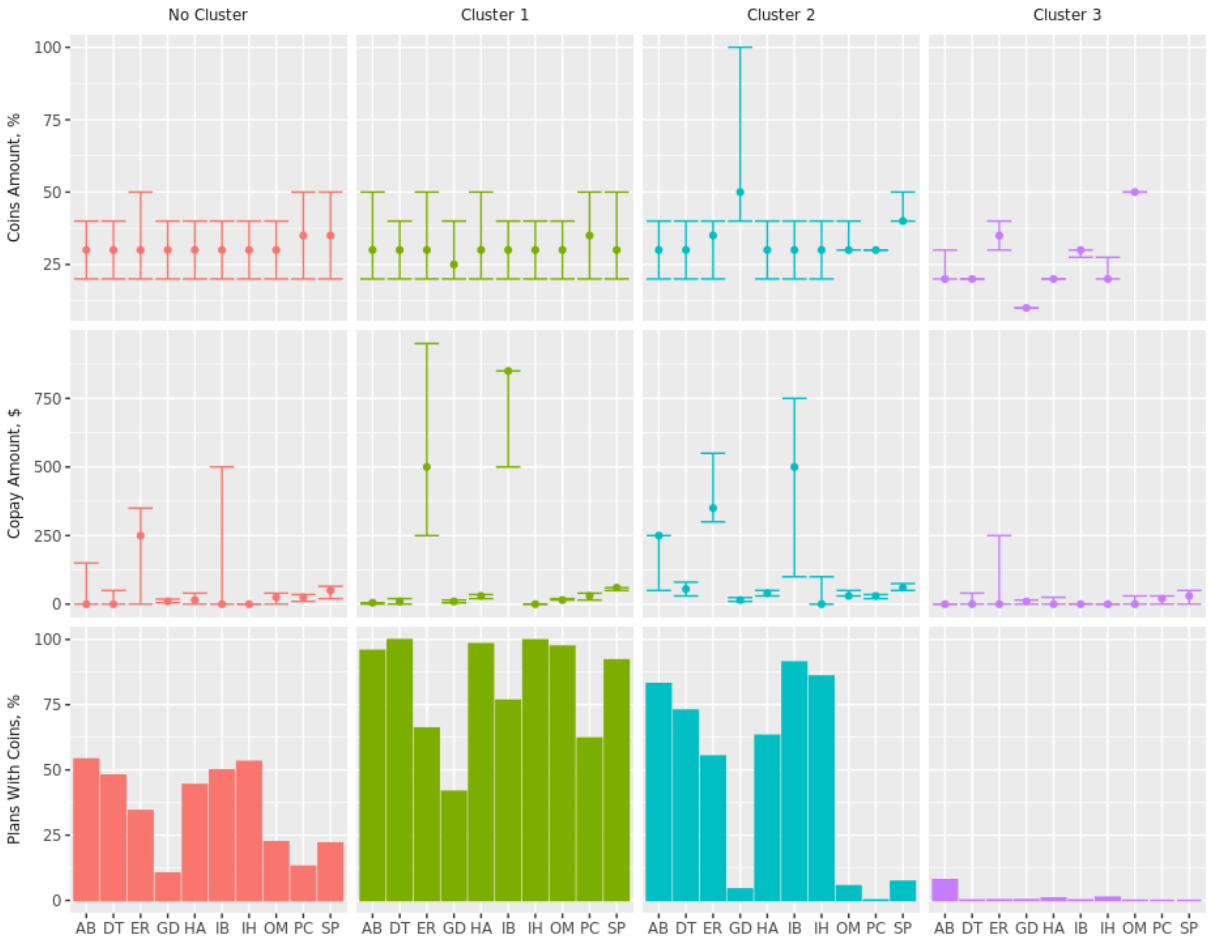
Figure 7 compares percentage of plans using coinsurance, the median coinsurance levels, and the median copay amounts for each benefit in each cluster. It also provides the benefit details for all plans as a reference. We provide a written overview of each cluster below. Detailed tables of cluster characteristics are included for reference in Appendix A.

3.3.1.1 Cluster 1

Plans in Cluster 1, the most restrictive cluster, have coinsurance probabilities that are higher than the coinsurance probabilities for the average plan across all benefits. Benefits that are important to HIV prevention including Diagnostic Tests (DT), Outpatient Mental Health (OM), and Specialist Providers (SP) are restricted through coinsurance usage in more than 90% of plans. While access to Primary Care (PC) is not restricted using coinsurance in all plans in the cluster (62.19%), it does appear in a far greater percentage of plans in this cluster than in the plans

[2] To maintain the ordinality of restrictiveness in our data, we have re-numbered the clusters from those found in the exploratory graphics above. From here forward, Cluster 1 refers to the most restrictive cluster, Cluster 2 refers to moderately restrictive cluster, and Cluster 3 refers to the least restrictive cluster.

Figure 7: Ward's 3 Cluster Solution Cluster Characteristics — Benefits



The top two rows display Median and IQR Coinsurance and Copay amounts for each cluster. The bottom row displays percentage of plans with coinsurance within each cluster.

when considered as a whole (13.04% as reported in Table 1C). This demonstrates an important correlation between restricted access to three dimensions important to HIV care (DT, OM, & SP) and higher likelihood of restrictiveness for PC.

Plans in Cluster 1 also defer significant financial risk and cost uncertainty to patients in cases of medical emergency. 94.7% cover ambulance care with coinsurance and 98.65% ask patients to pay a portion of their inpatient physician hospital bills. While there is more variance in coinsurance usage for emergency room visits with 66% of plans taking that strategy, the

copays among the remaining 34% of plans require a median payment of $500, which is double the median payment among the entire set of plans ($250).

3.3.1.2 Cluster 2

Plans in Cluster 2, the moderately restrictive cluster, relax key restrictions around preventative care for HIV and add certainty into payment amounts. But, they still do not facilitate optimal financial security and access to the range of services necessary for HIV prevention. Plans in Cluster 2 have coinsurance probabilities for Outpatient Mental Health, Primary Care, and Specialist Care that are close to 0. That said, the copays used for Specialist Care are generally higher than those used for Specialist Care across all plans (Cluster 2 Median [IQR] = $60 [25]; All Plans Median [IQR] = $50 [45]).

In terms of facilitating general patient health and financial stability, plans in Cluster 2 facilitate increased access to generic drugs with low coinsurance probabilities (.044) and low copays (Median [IQR] = $15 [14]). While the probability of coinsurance for Emergency Care decreased relative to Cluster 1 (.66 to .55), the copays among the remaining plans were still high (Median [IQR] = $350 [250]). Overall, Cluster 2 provides greater access to HIV preventative care than Cluster 1 without significantly expanding access to the care necessary for general health as well.

3.3.1.3 Cluster 3

Plans in Cluster 3, the least restrictive cluster, almost completely eliminate coinsurance usage across all benefits. They also significantly reduce copay amounts, with median copays of $0 for

benefits such as Ambulance care, Diagnostic tests, Emergency Room Care, Habilitation Services, Inpatient Birth, Inpatient Physician care, and Outpatient Mental Health Care.

3.3.2 Accessibility of PrEP

Figure 8: Ward's 3 Cluster Solution — PrEP

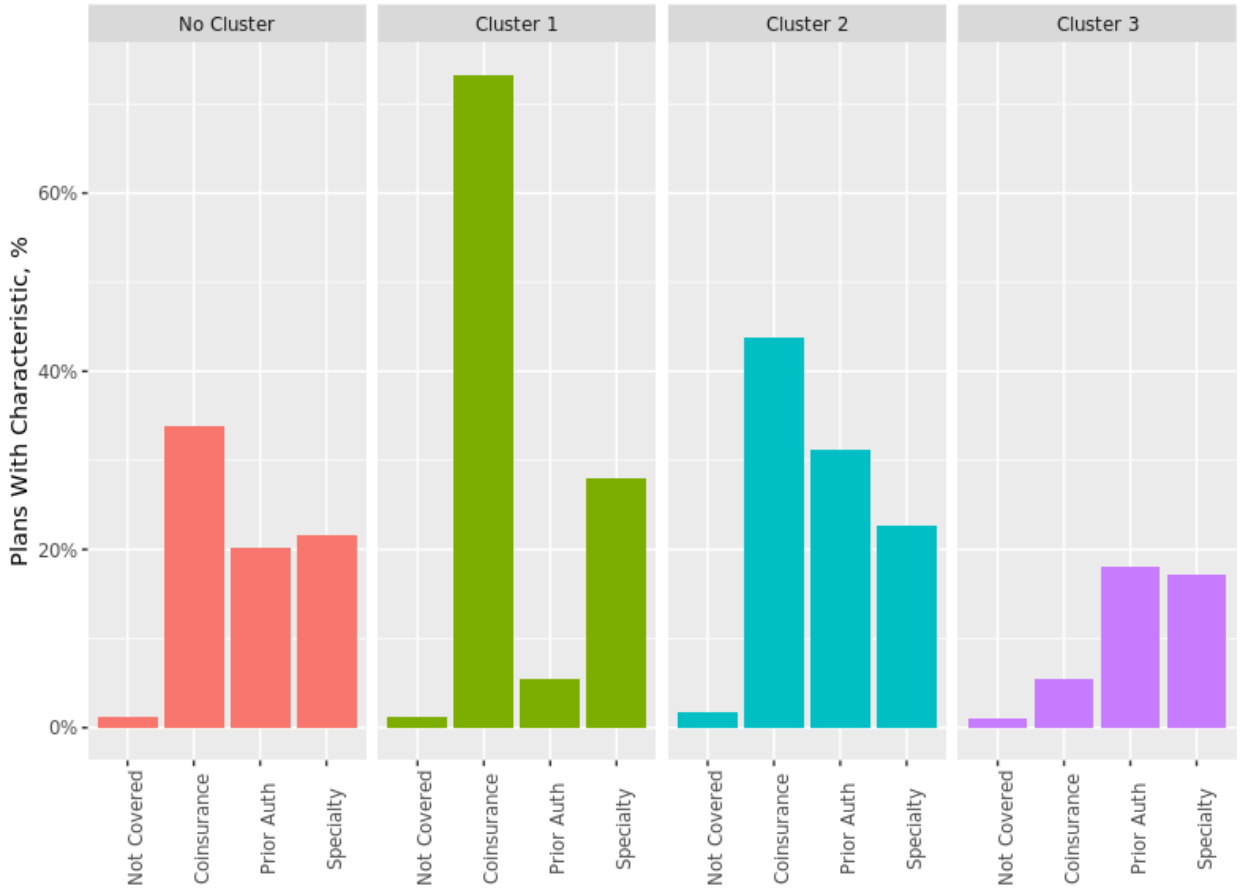


Figure 8 displays the percentage of plans in each cluster that place specific restrictions on Pre-Exposure Prophylaxis. Coverage of PrEP was not an influential variable in the clustering process given 98.75% of all plans covered PrEP in some way. Because of this, coverage of PrEP does not vary significantly between clusters. As expected, given patterns of coinsurance prevalence across other benefits, 72.15% of Cluster 1 plans, 42% of Cluster 2 plans, and only 4.45% of Cluster 3

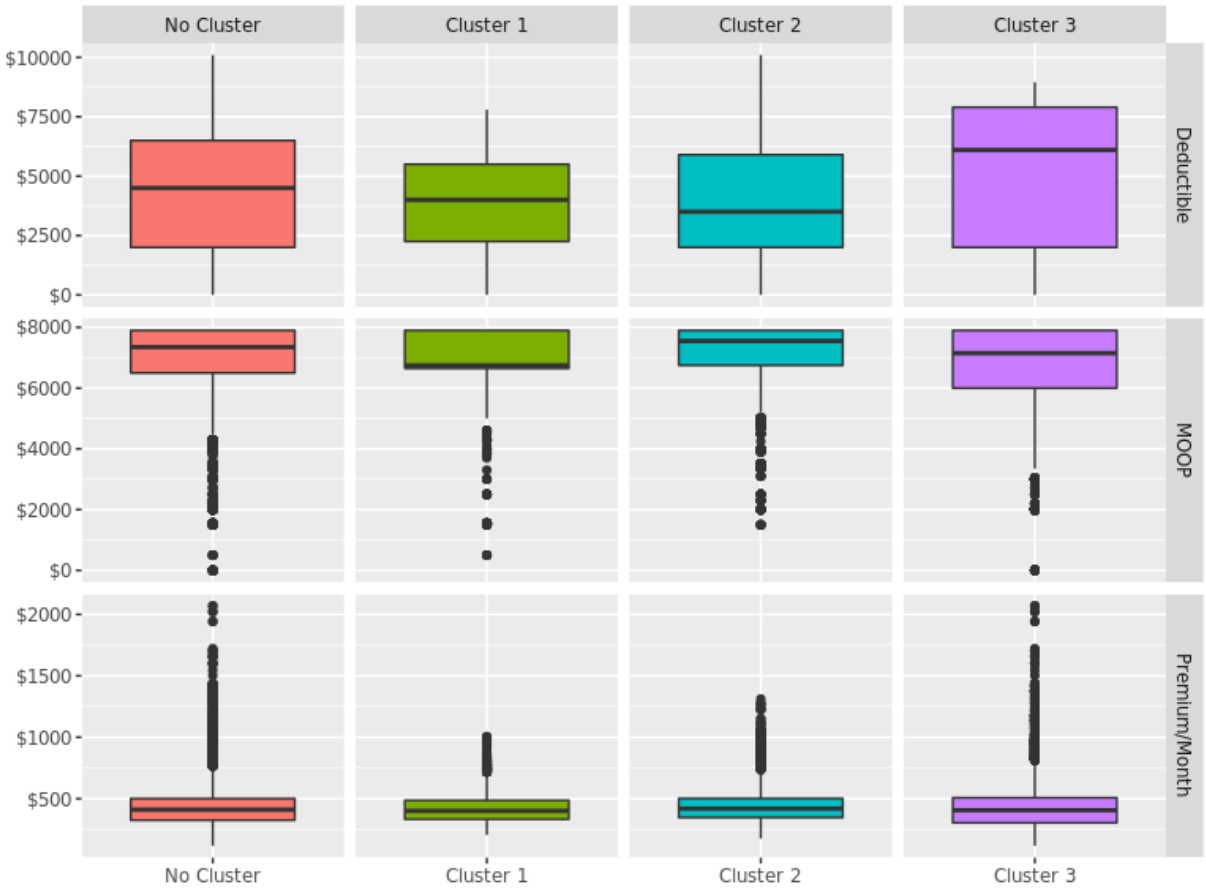
plans used coinsurance to cover PrEP. This represents a significant contrast between Cluster 3 and the other two clusters. Use of specialty tiering for PrEP mirrored the restrictiveness pattern and was most prevalent among plans in Cluster 1 (28.0%), but the contrasts were not as stark between clusters (Cluster 2: 22.7%; Cluster 3: 17.3%). This suggests that usage of specialty tiering for PrEP, while still positively correlated with other variables of restrictiveness, is only weakly so. Finally, use of Prior Authorization for PrEP exhibits a completely different pattern entirely, spiking in Cluster 2 at 29.4% and appearing in very few plans in Cluster 1 (4.24%). This suggests that prior authorization is less likely to be imposed when cost restrictions are higher and that "restrictiveness" as captured by coinsurance usage does not correlate positively with "restrictiveness" as defined in terms of prior authorization.

3.3.3 Administrative Variables

Figure 9 displays distributions of administrative cost variables by cluster. Overall, plans in Cluster 3 had the highest median deductible at $6100 as compared to $4000 for Cluster 1 and $3500 for Cluster 2. Despite this, Cluster 2 had the highest median maximum out of pocket costs at $7550 as compared to $7150 for Cluster 3 and $6750 for Cluster 1. This suggests that while Cluster 1 may be the most restrictive, enrollees paying for high-cost care — such as individuals taking preventative medications for PrEP — may actually pay more for their care overall in Cluster 2. Finally, monthly premiums were relatively equivalent between the three groups at $401/month, $420/month and $406/month for Clusters 1, 2, and 3, respectively. This suggests that, in addition to having fewer barriers to care access once enrolled, individuals who are covered by plans in Cluster 3 do not pay more than individuals in Cluster 1 or 2 to maintain their enrollment.

Finally, as reported in Appendix A, we found that 33% of Cluster 1 plans used a multi-tiered structure as compared to 2.64% in Cluster 2 and 5.97% in Cluster 3.

Figure 9: Ward's 3 Cluster Solution — Administrative Variables



3.4 Geographic Distribution of Clusters

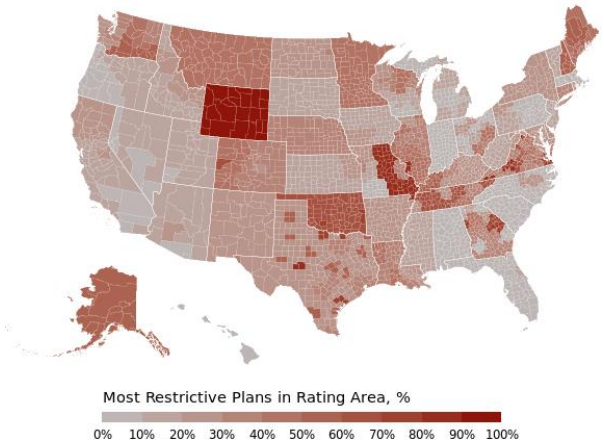
Figure 10 displays the proportion of plans in each Rating Area that fall into each cluster Overall, we find the highest rates of the most restrictive plans in Wyoming, Virginia, Illinois, Missouri, Texas, Georgia, and Oklahoma. Each of these states have rating areas where at least 75% of plans offered are in the most restrictive cluster.

There are no rating areas where more than 75% of the plans offered fall into the moderately restrictive cluster. However, the moderately restrictive plans are most prevalent in Michigan, Utah, Georgia, Rhode Island, Hawaii, and North Carolina. In each of these locations, moderately restrictive plans comprise at least two- thirds of the plans offered.

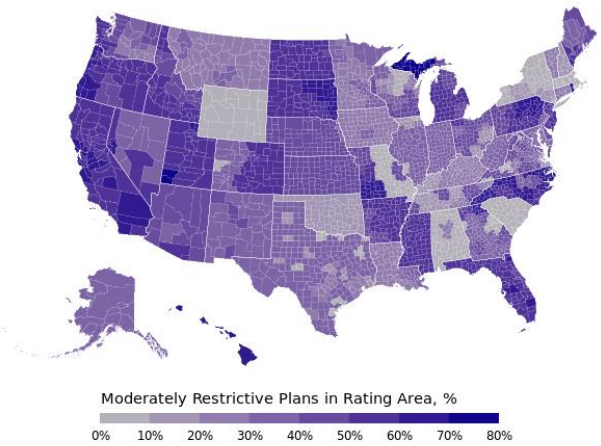
In Alabama, there are several rating areas where 100% of plans offered are categorized as least restrictive. Other states with rating areas with high prevalence of least restrictive plans include Massachusetts and South Carolina. Two rating areas in Arkansas do not offer any of the least restrictive plans. This is also the case in three rating areas in Wyoming. One rating area in Washington, four in Maine, and one in Tennessee all have markets in which least restrictive plans comprise less than 10% of the plans offered.

Figure 10: Percent of Plans in Rating Area by Cluster

Most Restrictive



Moderately Restrictive



Least Restrictive

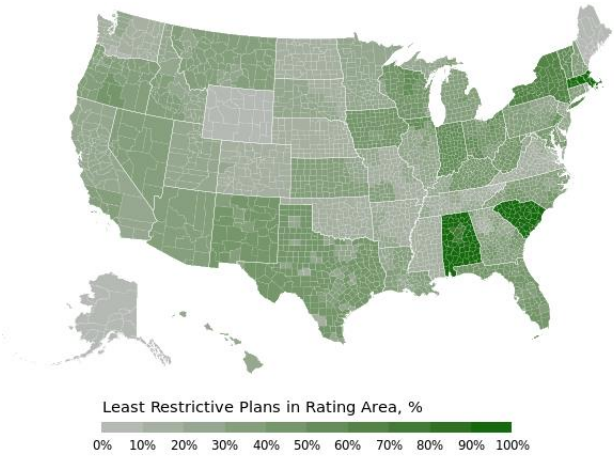
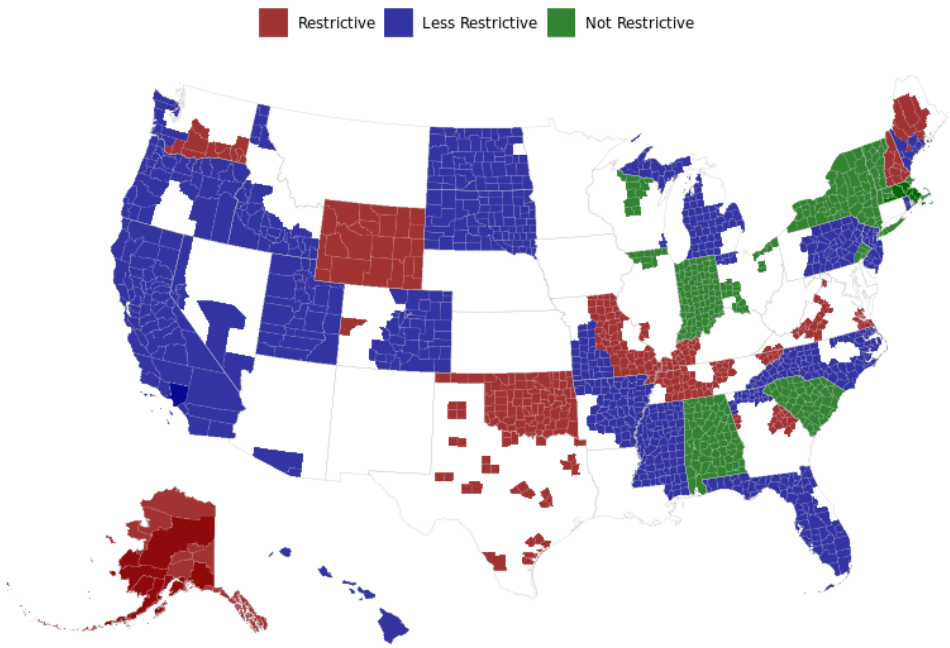


Figure 11 colors rating areas to show places where a specific cluster comprises a majority of plans, displaying the degree to which local markets are competitive or dominated by specific types of plans. States including Oregon, Nevada, Arizona, Louisiana, Montana, Nebraska, Kansas, Minnesota, Iowa, West Virginia, Maryland, and Georgia, have markets with higher degrees of heterogeneity. Plans from one cluster do not comprise more than 50% of the market share of almost all rating areas in those states. This allows individuals greater agency when tailoring their benefits to their needs. In contrast, states like Missouri, Tennessee, Virginia, and Texas have high

inter-rating area variability in plan offerings. While less restrictive plans are readily available in one rating area, moving to another rating area within the state restricts an individual's choices.

Figure 11: Cluster with Majority Share of Plans



Areas that remain white do not contain one cluster that exceeds 50% of plans offered

4 Discussion

4.1 Methodological Performance

At the outset of this investigation, we asked if clustering methodologies could be reasonably extended to handle the inherently nested structure of health plan data. We then, contingent on the usability of the methodology, asked content-based questions — what conclusions can we draw about how insurance companies cover preventative care for HIV? Is there a geographic pattern to that coverage? For this reason, we begin our discussion assessing the performance of our methodology and the usability of our results before moving into substantive considerations of what our work suggests for the field of health coverage and HIV prevention.

4.1.1 Distance Metric

The primary methodological innovation of our work was extending Gower's conception of distance and hierarchy to account for the nesting and patterned missingness inherent in health plan data (1971). In setting weights for our distance metric, we chose to prioritize categorical distinctions between copays and coinsurances due to our belief from theory in the behavioral importance of coinsurances with regard to consumer choice. In setting up our metric, we defined plans to be maximally distant from each other if one used coinsurance and the other used copay for a given benefit. If they used the same coverage strategy, we added an extra term to quantify the degree of that distinction. In this way, plans that cover the same benefit with copays of different magnitudes could still be distinct from each other, but never maximally

distinct. Functionally, using our weighting scheme, we only calculated distances between two plans based on the factors they shared in common.

It is certainly possible to contest this assertion from a theoretical standpoint — should a plan with a $400 copay for specialist care share some similarity with a plan that asks for a $5 copay for specialist care but not with a plan with a 90% specialist care coinsurance just because it uses a copay and not a coinsurance? We believe the answer is yes given the empirical work reviewed in Section 1.3.1 that suggests individuals behave differently when covered by coinsurances as opposed to copays because there is inherent uncertainty in the final cost of care even when those costs are equal. However, distinctions in patient behavior between plans with low coinsurance and plans with high copay are understudied and could influence theoretical considerations for weighting in the future. That said, challenging this assertion is a content-based argument not necessarily a methodological one.

Rather, the contrasts between clusters suggest that the weighting strategy utilized effectively operationalized the theoretical distinctions we hoped to encode. Foremost, the distinctions between plans in Cluster 3 and those in Clusters 1 and 2 are stark — plans in Cluster 3 have almost no coinsurance liability whereas those in Clusters 1 and 2 have high percentages. In creating a distance metric to prioritize copay and coinsurance distinctions, we recovered clusters that do the same.

Additionally, for benefits with lower rates of coinsurance across the board such as Emergency Services, we also recovered contrasts on the copay amounts between clusters. For example, where Cluster 2 has a 50% coinsurance rate for Emergency Services, it also has a high median copay amount relative to Cluster 3. This suggests that the secondary concern — allowing

for contrasts between plans that take similar coverage strategies but have divergent costs for those strategies — was also recovered, although with lower priority. But, those contrasts were not recovered as significantly across all benefits. We see that for some benefits, like ambulance services, that have similar rates of coinsurance between Clusters 1 and 2, the third quartile of coinsurances is higher among plans in Cluster 1. This is not true across the board. For example, there is little contrast in both the rate and the amount of coinsurance for Inpatient Physician services between Clusters 1 and 2. This is reasonable given the clustering methodology represents a multivariate analysis. If two benefit characteristics do not covary across plans, it may not be possible to create meaningful contrasts on them both at the same time (Romesburg, 2004).

This reflection of theoretical considerations within the clusters is encouraging, particularly given the amount of data included. As Ronan et al. (2016) discuss, issues of high dimensionality in clustering methodologies can easily lead to spurious results since the probability of two observations contrasting on a few variables increases with the number of variables considered. This reduces the range of the distance metric and creates random associations as opposed to recovering meaningful clusters. However, this does not seem to be the case within our result. The extent to which the relationships between variables aligns with content-based considerations as later discussed in Section 4.2 suggests we are not simply uncovering a structure of noise.

In contrast to Gower's suggestion of setting weightings based on theory, work from van de Hoven (2015) has suggested iteratively constructing the distance matrix weights in conjunction with feedback from growing a hierarchical tree, selecting the set of weights that

optimizes the cophenetic correlation coefficient. Tuning the weights this way yields a potentially interpretable weight matrix that reflects the relative influence of each variable when seeking the best possible defined clusters. However, while this process serves the end of discovering clusters, it does not necessarily prioritize discovering interpretable clusters that have theoretical value. While van de Hoven's method has inherent value as an exploratory process — i.e. discovering unforeseen clusters and the variables that drive their formation — in this work we sought to define clusters that comment on the relationships between benefits across plans. As such, optimizing the distance matrix to best fit into a hierarchical tree would not have provided resultant clusters that account for equal contribution between benefits and would not enable as easily the exploration of the multivariate relationships discussed in Section 4.2.

While algorithmically setting the weights may not have provided the interpretable results desired from this analysis, noting the relative lack of contrast between groups on the quantitative administrative variables (Deductible, MOOP, and Premium) suggests that a future direction for this type of work may be to group variables by type —administrative variables, general benefits, and HIV preventative benefits, for example — and set weights to balance contributions to the distance metric. Chae, Kim, & Yang (2006) provide such an example of balancing the contributions of continuous variables and categorical variables while the precedent for grouped weighting extends back to Gower himself (1971).

Overall, the contrasts on both prevalence and values of copay and coinsurance between clusters suggest that our weighting choice accurately encoded key theoretical considerations and translated them through the distance matrix into the output. While the theoretical considerations themselves can be challenged, the weighting strategy used to translate the

nested structure of the data into a usable distance matrix was effective in light of the resultant cluster contrasts.

4.1.2 Clustering Performance

In Section 1.4, we discussed the difficulties of comparing health plan offerings. As the outcome of complex decision-making processes within companies, the high dimensionality of and dependencies within the data restrict the interpretability of descriptive comparisons. We proposed clustering as a possible solution to reduce dimensions and enable more usable comparisons. Having carried out the investigation, we must decide if those comparisons are viable through our solution.

Clustering high dimensional data brings with it a number of pitfalls. It is not always robust to small perturbations in methodology and has a growing likelihood of detecting spurious results as the number of clusters increases (Ronan et al., 2016). This was likely the case in the five clustered solution to the average linkage hierarchical tree. Because of this, Ronan, Qi, & Naegle (2016) suggest running multiple clustering strategies. Finding convergent structures between methodologies provides evidence that clusters uncovered are substantive as opposed to spurious. Since Ward's method and Partitioning Around Medoids (PAM) converged in both the two and three cluster solutions, we take this as support that our final solution represents a repeatable phenomenon as opposed to an idiosyncrasy of an individual algorithm.

As previously discussed in the results section, from a mathematical perspective, the three-cluster solution chosen is defensible. Albeit, from an absolute perspective, the average silhouette distance of .414, while confirmatory of the existence of true clusters, is not indicative

of the most well-defined cluster solutions in the field (Lengyel & Botta-Dukát, 2019). However, this does not invalidate the usefulness of the results. The average silhouette width, while widely used, was developed as a graphical aid to help researchers visualize how closely together the points in their clusters clung (Rousseeuw, 1987). Because of this, there are no explicit rules of thumb or cut points — just a general consensus that values near zero suggest clusters may not exist within the data and that negative values represent a complete misfit between the data and the model (Lengyel & Botta-Dukát, 2019).

Interpreting the average silhouette in light of the number of dimensions we incorporated into the cluster analysis, it is not necessarily concerning that it is not higher. With so many variables, there would need to be significant covariance across all variables between plans to create more distinct contrasts. This is why Ronan (2016) also suggests drawing heavily on ideas of practical significance. Perhaps not all variables between clusters contrast, but if there are key contrasts that provide useful information, then the solution is of value.

The ordinality of the clustered results — three clusters with generally decreasing restrictiveness as discussed in Section 3.3 — provides the most compelling evidence of the usefulness of the solution and of the practical distinctness of the clusters even if some variables are not well contrasted. The graphical evidence shown earlier in Figure 7 is the most intuitive demonstration of this value.

The ordinality and cluster distinctiveness are not without qualification. Chiefly, benefits where coinsurance probabilities are close to .5 — such as Emergency Room in Clusters 1 & 2, Habilitation Services in Cluster 2, and Primary Care in Cluster 1 — do not represent as useful categorizations. The clusters are useful to the extent that knowing cluster membership says

something meaningful about an individual plan. Yes, being able to say that a plan in Cluster 3 almost assuredly does not have coinsurance for Emergency Room services but that a plan in Cluster 2 has about a 55% chance of it is useful. However, we really do not know with any certainty better than chance whether a random plan in Cluster 2 will or will not require coinsurance for an ER visit. But with so many prevalences close to either 1 or 0 across a number of benefits in each cluster, these results support the conclusion that clustering is a useful tool to illustrate macro level trends in health insurance and enable plan comparisons.

4.2 Insights from Cluster Relationships

Having concluded that the results are methodologically sound enough to facilitate interpretation, we move to examine what the macro-level patterns between the clusters reveal about how insurance companies construct healthcare plans. Viewing clustering as a type of multivariate analysis, examining which benefits co-occur in plans that are grouped together provides insight into how plans are constructed.

4.2.1 Coinsurance and Copay Patterns

Interpreting the pattern of coinsurances and copays across the three clusters provides key exploratory takeaways about how insurance companies construct plans. First, the pattern of copays within Cluster 3 suggests that, in general, plans that have less copays on some benefits tend to have less restrictive copays on other benefits. There is a whole subset of plans created by insurance companies that requires lower cost fixed payments for services that would be more easily navigable to a person on a preventative regimen for HIV. This contrasts starkly with the conclusions supported by Cluster 1, specifically that insurance companies also create plans that have high coinsurance-driven restrictiveness. The existence of these two extreme clustering patterns suggests a lot about how insurance companies set benefits, offering either full access or more complete restriction.

Adding in considerations from Cluster 2 further draws the tiered systems of restrictiveness into focus. The way in which higher copays for Diagnostic Tests and Emergency Care are used for plans that do not impose coinsurance suggests some degree of positive covariance between more restrictive copays and higher restrictiveness on other benefits. Given

how we set up our distance metric, the only way plans with coinsurance and plans with copay for Emergency and Diagnostic Test benefits would end up in the same cluster is if they contained similarities on other benefits. What the pattern shows, then, is that, in some plans, coinsurances are swapped for high copays while maintaining the same patterns of coinsurance-based restrictiveness elsewhere.

However, the patterns within Cluster 2 support deeper insights as well — specifically that when insurance companies choose to lower some restrictions, they prioritize access to Primary Care, Specialist Care, Generic Drugs, and Outpatient Mental Health. These benefits are all important for individuals who are on preventative regimens to guard against HIV and maintain general well-being (Owens et al., 2019). That said, benefits that provide a financial cushion in more catastrophic events — Ambulances, Emergency benefits, and Inpatient Hospital care, remain highly restricted.

These patterns of restrictiveness comport generally with the Bronze, Silver, Gold and Platinum plan stratifications created by the ACA which intend to balance out-of-pocket costs with premium costs (healthcare.gov). We will further explore cost patterns in Section 4.2.3 to see if the monetary portion of that balance holds true.

4.2.2 Prior Authorization Patterns

The pattern of lowest prior authorization for PrEP among the most restrictive group suggests enrollee willingness to accept higher out of pocket costs is associated with higher agency in initiating PrEP. Patients that were willing to accept greater restrictions and financial responsibility elsewhere were less likely to need approval before initiating PrEP drugs. Framing

this from an insurance provider perspective as opposed to a consumer agency perspective, this finding also suggests that insurance companies feel less need to restrict access using prior authorization when there are already cost-based restrictions controlling care. Yet, in plans where cost-based restrictions are relaxed, prior authorization is instituted as a control. This runs contrary to messaging around prior authorization advanced by insurance companies that prior authorization is for patient protection (See Section 1.3.3). In theory, individuals should not be able to buy the right to opt out of a regulation designed for their safety, nor should insurance companies utilize prior authorization as a restrictive counterbalance to accessible pricing.

This comports with previous work from McManus et al. (2020) which found that being enrolled in a plan that covers PrEP with coinsurance as opposed to copay lowers the odds of PA by a factor of 0.51 when other plan factors re held constant. Lower PA was also associated with other measures of PrEP restrictiveness such as specialty tiering. The findings from this cluster analysis augment those findings to suggest that reduced PA requirements for PrEP are not just related to increased restrictions on PrEP but also increased restrictions on all other plan factors and that relaxing restrictive plan factors raises likelihood of an insurer instituting PA.

4.2.3 Cost Patterns

The relative equality of premium prices per month between clusters suggests that contrasts on other variables were more useful in determining clusters than contrasts on premiums. It also suggests that restrictiveness of benefits across plans may vary more independently from price than posited by healthcare.gov and the metal level rating structure referenced in Section 4.2.1. This does not mean that premium and restrictiveness are necessarily independent among plans

offered by the same insurance company. But it does mean that, across the United States, plans with lesser restrictions do not necessarily cost individuals more per month to remain enrolled. The implications of this finding for people living with HIV are significant — it is possible for insurance companies to provide care that has lowered restrictiveness without raising premiums. To End the HIV Epidemic, these are the types of plans that insurance companies should be incentivized to offer in key regions of the US.

That said, plans in Cluster 1 — which required larger and less predictable out-of-pocket costs for services — did have lower median yearly maximum out of pocket costs. This enables increased certainty when budgeting for maximum medical costs but individuals must have the liquidity to pay more unpredictable contributions over the course of the year. While individuals enrolled in Cluster 3 plans would have a higher MOOP, they would also have relatively low copays and would be less likely to reach the MOOP. The implications of this metric are difficult to evaluate without simulating costs for an individual for an entire year. But from a total spending perspective, there are benefits and drawbacks to being in either Cluster 1 or 3. Cluster 2, with the highest median MOOP and high prevalence of copays among specific benefits while enabling easy access to specific types of routine care could quickly become the most burdensome in a catastrophic or emergent event.

4.2.4 Policy Implications: Cost Sharing for PrEP

In 2019, the US Preventative Service Taskforce (USPSTF) gave PrEP a grade A rating for preventative care (Owens et al., 2019). Under ACA guidelines, private insurance plans must cover services that receive a grade A rating without any cost sharing at all. For PrEP, this shift was set

to occur at the start of 2021. Going forward, individuals on all plans should have access to PrEP without coinsurance or copay. In line with our discussion of the behavioral implications of coinsurance and copay, this shift may help patients be more confident about their ability to continue their treatment. However, from a financial perspective, it is not as significant a shift as it seems at face value. In the past, PrEP financial assistance programs administered by Gilead — the pharmaceutical company that produces PrEP — and by state agencies covered out of pocket PrEP medication costs for insured individuals (NASTAD, 2020). Removing cost sharing from PrEP across all health plans will allow those financial assistance programs to concentrate their focus on the uninsured. But for those who were previously receiving reimbursement for PrEP, it does not mitigate some of the most important barriers. The USPSTF recommendation does not make it clear what, if any, ancillary services plans have to cover in order to provide the PrEP intervention. Firsthand reports from patients who had their PrEP drug costs covered previously note that while they could obtain the medication itself, the provider visits and labs required to maintain a PrEP regimen presented a financial barrier and, in some cases, led to discontinuation (Andrews, 2019 & 2021).

This is concerning given the patterns of health coverage observed for wraparound services. Eliminating cost sharing for PrEP will mainly assist individuals enrolled in plans from Clusters 1 and 2. And yet, usage of coinsurance and elevated copays remain high for Diagnostic Tests and Primary and Specialist care in Cluster 1, in specific. Even with lesser restrictiveness in vital services among Cluster 2 plans, costs associated with Diagnostic Tests — one of the most critical requirements for maintaining PrEP treatment — remain restrictive. For the US Preventative

Service's regulation to have its intended effect of increasing PrEP access, it must also address restrictive cost sharing for associated services. (Kay & Pinto, 2021).

Finally, with increased competitiveness in the PrEP market as generics, different chemical formulations, and long-acting injectable PrEP make their way through the FDA pipeline, more policy analysis will be needed to examine how insurance companies react to these innovations and continue to interpret the USPSTF mandate (Coelho et al., 2019).

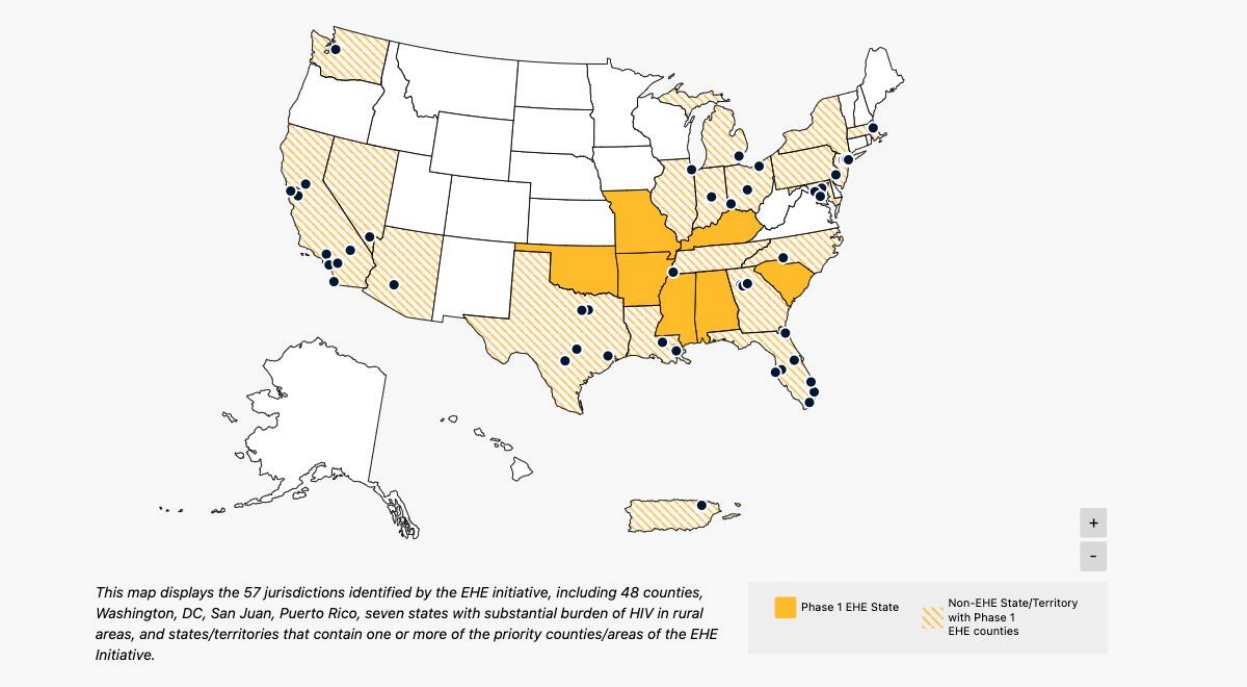
4.3 Geographic Insights

In 2019, the US government identified 57 jurisdictions with increased HIV transmission for prioritization in the Ending the HIV Epidemic Plan (EHE). These jurisdictions, shown in Figure 12, include areas where individuals are at higher risk of acquiring HIV and emphasize places in which there is a higher burden of rural transmission. It is important that these jurisdictions offer plans that reduce barriers to accessing PrEP and other preventative services. In general, EHE jurisdictions vary in what types of plans are available. Rating areas in Oklahoma, for example, have high incidence of plans from the most restrictive cluster while rating areas in South Carolina have high incidence of plans from the least restrictive cluster.

We also observe significant stratification in plan types by state with plans in the moderately restrictive cluster occupying the greatest market share overall, specifically in the West, and the least restrictive plans concentrating primarily in the Northeast, Indiana, South Carolina, and Alabama. This emphasizes the importance of state politics in setting health care priorities. For EHE, federal policymakers will have to work with states with differing policy environments to harmonize a collective strategy.

EHE must also focus on reducing prior authorization requirements to make accessing PrEP on Cluster 2 and 3 plans both financially and logistically feasible. Finding higher prevalences of Cluster 2 and 3 plans in Southern US states, specifically in Florida and Mississippi, compared with other regions comports with observations from McManus et al. (2020) that note increased prior authorization restrictions in those areas.

Figure 12: Ending the HIV Epidemic Priority Jurisdictions



Source: Ahead.hiv.gov

Our finding of within-state variability, particularly in Texas, also gives reason for pause. In most rating areas in Texas, individuals have the option to choose between all three types of plans. But, those healthcare options solidify in major city areas and guide individuals toward the most restrictive options. A 2018 report by the Texas Department of State Health Services notes that over 75% of PLWH in Texas, who are overwhelmingly Black and Hispanic MSM, reside in 5 major urban areas. This means individuals in urban areas are at the highest risk of acquiring HIV. This pattern, where care is restricted in places people need it most, is a clear demonstration of how healthcare systems discriminate against individuals at risk of HIV and contribute to a lasting system of institutionalized racism in healthcare.

4.4 Limitations & Future Directions

This body of work represents a first foray into using clustering methodologies to audit healthcare offerings across the US. The methodology generated useful results but is not without limitations. Firstly, while the clustering algorithm accounts for as many plan factors as possible, this analysis is still limited to in network benefits and does not account for the accessibility of in network providers. An individual may be enrolled in a Cluster 3 plan with access to low copays for vital services for in network providers, but it is possible the providers in their network are over booked or geographically distant. In that case, benefits covering out of network providers would provide more accurate information about an individual's lived experience of finding care.

Secondly, while there were contrasts between clusters on benefit characteristics, the methodology itself does not provide the interpretative value. We drew conclusions about restrictiveness and access based on theory about how patients act in response to specific plan factors. While our method is expedient in that it requires only administrative data on plan designs, it would be strengthened by reported experiences of patients navigating preventative care for HIV while enrolled in archetypal plans from each cluster.

Thirdly, as an exploratory methodology, clustering is difficult to validate. We spent much of this work justifying the fit of our solution based on its practical interpretation. And yet, not all aspects of it — the marginal contrasts on specialty tiering for PrEP, for example — are of practical use. To some degree, this reflects the issue of clustering data with many dimensions (Ronan, Qui, Naegle, 2016). But at the same time, it is a reminder that not every dimension falls into a neat three category solution of meaningfully decreasing restrictiveness. As an aggregate level descriptor of a complicated system, cluster membership does not strictly imply that

randomly choosing one plan from each cluster and comparing them will always align with the pattern that arises from the whole group.

Keeping the above in mind, if this type of method is to be used in auditing plan offerings in the future, more work is needed to strengthen the theory linking specific plan characteristics to the experiences of patients. This could involve analyzing patient satisfaction data across various plan types or, with granular longitudinal data, could include investigating linkages between plan offerings in rating areas and the uptake of preventative services such as PrEP. This type of content-driven work would improve the interpretative value of the methodology. In terms of the methodology itself, with increased linkage between plan characteristics and patient outcomes, it would be worthwhile to experiment with other weighting matrices that create equal contributions between benefits or groups of benefits based on theoretical considerations.

Overall, clustering health plan data to examine trends in preventative care for HIV is a viable method to describe a very complex system. While the results from one clustering may not provide the basis for a complete overhaul of policy, obtaining a macro-level view of the entire system can help researchers and policymakers identify specific content areas for further research and investigation.

References

Agniel, D., Martino, S. C., Burkhart, Q., Hambarsoomian, K., Orr, N., Beckett, M. K., ... & Elliott, M. N. (2019). Incentivizing excellent care to at-risk groups with a health equity summary score. *Journal of general internal medicine*, 1-11.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, *6*(1), 37-66.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press. CDC.

Andrews, M. (2019, July 15). Even When HIV Prevention Drug is Covered, Other Costs Block Treatment. *Kaiser Health News.*

Andrews, M. (2021, January 5). Many Health Plans Now Must Cover Full Cost of Expensive HIV Prevention Drugs. *Kaiser Health News.*

Barreiro P. (2018). Sexually Transmitted Infections on the Rise in PrEP Users. *AIDS reviews*, *20*(1), 71.

CDC. Monitoring Selected National HIV Prevention and Care Objectives by Using HIV Surveillance Data—United States and 6 Dependent Areas, 2018. HIV Surveillance Supplemental Report 2020;25(No. 2).

Center for Disease Control (2019, November 12). *Effectiveness of Prevention Strategies to Reduce the Risk of Acquiring or Transmitting HIV.* US Department of Health and Human Services.

Center for Disease Control. (2019, September). *HIV in the Southern United States*. US Department of Health and Human Services.

Center for Disease Control. (2020, May 5). *Preventative Services Coverage.* US Department of Health and Human Services. https://www.cdc.gov/nchhstp/highqualitycare/preventiveservices

Center for Disease Control. (2021). *HIV Prevention in the United States: Mobilizing to End the Epidemic.* US Department of Health and Human Services. https://www.cdc.gov/hiv/pdf/policies/cdc-hiv-prevention-bluebook.pdf

Chae, S. S., Kim, J. M., & Yang, W. Y. (2006). Cluster analysis with balancing weight on mixed-type data. *Communications for Statistical Applications and Methods*, *13*(3), 719-732.

Chernew M , Juster I , Shah M , Wegh A , Rosenberg S , Rosen A . Evidence that value-based insurance can be effective . *Health Aff (Millwood)* . 2010 ; *29* (3): 530 – 6 .

Chernew, M., Gibson, T. B., Yu-Isenberg, K., Sokol, M. C., Rosen, A. B., & Fendrick, A. M. (2008). Effects of increased patient cost sharing on socioeconomic disparities in health care. *Journal of general internal medicine*, *23*(8), 1131-1136.

Clifford, H., Wessely, F., Pendurthi, S., & Emes, R. D. (2011). Comparison of clustering methods for investigation of genome-wide methylation array data. *Frontiers in genetics*, *2*, 88.

Coelho, L. E., Torres, T. S., Veloso, V. G., Landovitz, R. J., & Grinsztejn, B. (2019). Pre-exposure prophylaxis 2.0: new drugs and technologies in the pipeline. *The Lancet HIV*, *6*(11), e788 e799.

CoverMyMeds (2019) *ePA national adoption scorecard*. https://www.covermymeds.com/main/insights/scorecard/impact/.

Cuadras, C. M., Fortiana, J. & Arenas, C., 1998. Some computational aspects of a Distance-Based model for Prediction. Communication in Statistics- Simulation and Computation, 25(3).

Center for Disease Control [CDC]. Diagnoses of HIV Infection in the United States and Dependent Areas, 2018 (Updated). HIV Surveillance Report 2020; 31.

Dor, A., & Encinosa, W. (2010). How does cost-sharing affect drug purchases? Insurance regimes in the private market for prescription drugs. *Journal of Economics & Management Strategy*, *19*(3), 545-574.

Doshi, J. A., Li, P., Ladage, V. P., Pettit, A. R., & Taylor, E. A. (2016). Impact of cost sharing on specialty drug utilization and outcomes: a review of the evidence and future directions. *Am J Manag Care*, *22*(3), 188-197.

Eisinger, R. W., Dieffenbach, C. W., & Fauci, A. S. (2019). HIV viral load and transmissibility of HIV infection: undetectable equals untransmittable. *Jama*, *321*(5), 451-452.

Enflo, K. (2020). Measures of Similarity. *Theoria*, *86*(1), 73-99.

Farris, J. S. (1969). On the cophenetic correlation coefficient. *Systematic Zoology*, *18*(3), 279-285.

Fauci, A. S., Redfield, R. R., Sigounas, G., Weahkee, M. D., & Giroir, B. P. (2019). Ending the HIV epidemic: a plan for the United States. *Jama*, *321*(9), 844-845.

Felsenstein, J. (Ed.). (2013). *Numerical taxonomy* (Vol. 1). Springer Science & Business Media.

Filaire (2018). *Clustering on Mixed Data Types.* Towards Data Science.

Finlayson, T., Cha, S., Xia, M., Trujillo, L., Denson, D., Prejean, J., ... & National HIV Behavioral Surveillance Study Group. (2019). Changes in HIV preexposure prophylaxis awareness and

use among men who have sex with men—20 urban areas, 2014 and 2017. *Morbidity and Mortality Weekly Report*, *68*(27), 597.

Ford, M. A., & Spicer, C. M. (2012). Implications of Health Care Reform for People with HIV in the United States. In *Monitoring HIV Care in the United States: A Strategy for Generating National Estimates of HIV Care and Coverage*. National Academies Press (US).

Frank, M. B., Hsu, J., Landrum, M. B., & Chernew, M. E. (2015). The impact of a tiered network on hospital choice. *Health services research*, *50*(5), 1628-1648.

Fronstin & Roebuck, 2020. Managing Use of Health Care Services After People Satisfy Their Deductible: What Do Copayments and Coinsurance Do? EBRI Issue Brief, no. 519.

Gavin, D. G., Oswald, W. W., Wahl, E. R. & Williams, J. W., 2003. A statistical approach to evaluating distance metrics. Quaternary Research, Volume 60.

Gonçalves, L. et al., 2008. Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions. GMR, 7(4).

Gower, J. C. (1970). A note on Burnaby's character-weighted similarity coefficient. *Journal of the International Association for Mathematical Geology*, *2*(1), 39-45.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857 871.

Herman, P. M., Rissi, J. J., & Walsh, M. E. (2011). Health insurance status, medical debt, and their impact on access to care in Arizona. *American journal of public health*, *101*(8), 1437-1443.

Kalousova, L., & Burgard, S. A. (2013). Debt and foregone medical care. *Journal of health and social behavior*, *54*(2), 204-220.

Kay, E. S., & Pinto, R. M. (2020). Is insurance a barrier to HIV preexposure prophylaxis? Clarifying the issue. *American journal of public health*, *110*(1), 61-64.

Lefèvre, T., Rondet, C., Parizot, I., & Chauvin, P. (2014). Applying multivariate clustering techniques to health data: the 4 types of healthcare utilization in the Paris metropolitan area. *PloS one*, *9*(12), e115064.

Legendre, P. & Legendre, L., 1998. Numerical Ecology. 2nd ed. s.l.:Elsevier Science.

Lengyel, A., & Botta-Dukát, Z. (2019). Silhouette width using generalized mean—A flexible method for assessing clustering efficiency. *Ecology and evolution*, *9*(23), 13231-13243.

Lexchin, J., & Grootendorst, P. (2004). Effects of prescription drug user fees on drug and health services use and on health status in vulnerable populations: a systematic review of the evidence. *International Journal of Health Services*, *34*(1), 101-1t22.

Liao, M., Li, Y., Kianifard, F., Obi, E., & Arcona, S. (2016). Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC nephrology*, *17*(1), 1-14.

Liu, A. Y., Cohen, S. E., Vittinghoff, E., Anderson, P. L., Doblecki-Lewis, S., Bacon, O., Chege, W., Postle, B. S., Matheson, T., Amico, K. R., Liegler, T., Rawlings, M. K., Trainor, N., Blue, R. W., Estrada, Y., Coleman, M. E., Cardenas, G., Feaster, D. J., Grant, R., Philip, S. S., … Kolber, M. A. (2016). Preexposure Prophylaxis for HIV Infection Integrated With Municipal- and Community-Based Sexual Health Services. *JAMA internal medicine*, *176*(1), 75–84. https://doi.org/10.1001/jamainternmed.2015.4683

Lotvin, A. M., Shrank, W. H., Singh, S. C., Falit, B. P., & Brennan, T. A. (2014). Specialty medications: traditional and novel tools can address rising spending on these costly drugs. *Health Affairs*, *33*(10), 1736-1744.

Lyratzopoulos, G., Elliott, M. N., Barbiere, J. M., Staetsky, L., Paddison, C. A., Campbell, J., & Roland, M. (2011). How can health care organizations be reliably compared? Lessons from a national survey of patient experience. *Medical care*, 724-733.

McLachlan, G. J. (1992). Cluster analysis and related techniques in medical research. *Statistical Methods in Medical Research*, *1*(1), 27-48.

McManus, K. A., Killelea, A., Honeycutt, E., An, Z., & Keim-Malpass, J. (2020). Assisters Succeed in Insurance Navigation for People Living with HIV and People at Increased Risk of HIV in a Complex Coverage Landscape. *AIDS Research and Human Retroviruses*, *36*(10), 842-851.

McManus, K. A., Powers, S., Killelea, A., Tello-Trillo, S., & McQuade, E. R. (2020). Regional Disparities in Qualified Health Plans' Prior Authorization Requirements for HIV Pre exposure Prophylaxis in the United States. *JAMA Network Open*, *3*(6), e207445-e207445.

Mocroft, A., & Ryom, L. (2016). The benefits and risks of PrEP and kidney function. *The Lancet HIV*, *3*(11), e501-e502.

Montanari, A. & Mignani, S., 1994. Notes on the bias of dissimilarity indices for incomplete data sets: the case of archaeological classification. Qüestiió, 18(1). M

Mouchet, M. et al., 2008. Towards a consensus for calculating dendrogram-based functional diversity indices. Oikos, Volume 117.

Murtagh, F., & Contreras, P. (2011). Methods of hierarchical clustering. *arXiv preprint arXiv:1105.0121*.

National Alliance of State and territorial AIDS Directors (NASTAD). (2020). *PrEP Assistance Programs.* https://www.nastad.org/prepcost-resources/prep-assistance-programs

Office of Inspector General. (2019). *Some Medicare Part D beneficiaries face avoidable extra steps that can delay or prevent access to prescribed drugs.* US Department of Health and Human Services. https://oig.hhs.gov/oei/reports/oei-09-16-00411.pdf

Owens, D. K., Davidson, K. W., Krist, A. H., Barry, M. J., Cabana, M., Caughey, A. B., ... & Wong, J. B. (2019). Preexposure prophylaxis for the prevention of HIV infection: US Preventive Services Task Force recommendation statement. *Jama*, *321*(22), 2203-2213.

Petchey, O. L. & Gaston, K. J., 2009. Dendrograms and measures of functional diversity: A second instalment. Oikos, 118(7).

Ramos, H. et al., 2012. Multivariate analysis to determine the genetic distance among backcross papaya (Carica papaya) progenies. GMR, 11(2).

Remler, D. K., & Greene, J. (2009). Cost-sharing: a blunt instrument. *Annual review of public health*, *30*, 293-311.

Resneck, J. S. (2020). Refocusing medication prior authorization on its intended purpose. *Jama*, *323*(8), 703-704.

Romesburg, C. (2004). *Cluster analysis for researchers*.

Romley, J. A., Sanchez, Y., Penrod, J. R., & Goldman, D. P. (2012). Survey results show that adults are willing to pay higher insurance premiums for generous coverage of specialty drugs. *Health Affairs*, *31*(4), 683-690.

Ronan, T., Qi, Z., & Naegle, K. M. (2016). Avoiding common pitfalls when clustering biological data. *Science signaling*, *9*(432), re6-re6.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Roux, M. (2018). A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, *35*(2), 345-366.

Sasirekha, K., & Baby, P. (2013). Agglomerative hierarchical clustering algorithm-a. *International Journal of Scientific and Research Publications*, *83*, 83.

Sergios Theodoridis & Konstantinos Koutroumbas (2006). Pattern Recognition 3rd ed. p. 635.

Shendre (2020). *Clustering datasets having both numerical and categorical variables.* Towards Data Science.

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, *10*(12), e0144059.

Siegler, A. J., Mouhanna, F., Giler, R. M., Weiss, K., Pembleton, E., Guest, J., ... & Sullivan, P. S. (2018). 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*, *28*(12), 841-849.

Silapaswan, A., Krakower, D., & Mayer, K. H. (2017). Pre-exposure prophylaxis: a narrative review of provider behavior and interventions to increase PrEP implementation in primary care. *Journal of general internal medicine*, *32*(2), 192-198.

Sinaiko, A. D., Landrum, M. B., & Chernew, M. E. (2017). Enrollment in a health plan with a tiered provider network decreased medical spending by 5 percent. *Health Affairs*, *36*(5), 870 875.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification*.

Sokal, R. & Michener, C., 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, Issue 38, pp. 1409-1438.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*(2), 33-40.

Sorensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter, Issue 5, p. 1–34.

Spencer, M. S., Prins, S. C. B., & Beckom, M. S. (2010). Heterogeneous distance measures and nearest-neighbor classification in an ecological setting. *Missouri Journal of Mathematical Sciences*, *22*(2), 108-123.

Strauss T, von Maltitz MJ (2017) Generalising Ward's Method for Use with Manhattan Distances. PLoS ONE 12(1): e0168288.

Texas Department of State Health Services. (2018). *Texas HIV Epidemiologic Profile.* Texas Health and Human Services. https://www.dshs.state.tx.us/hivstd/reports/epi_profile/files/EpiProfile.pdf

Thompson, K. et al., 2009. Little evidence for limiting similarity in a long-term study of a roadside plant community. Journal of Ecology, 98(2).

Thomson, S., Schang, L., & Chernew, M. E. (2013). Value-based cost sharing in the United States and elsewhere can increase patients' use of high-value goods and services. *Health Affairs*, *32*(4), 704-712.

U.S. Food & Drug Administration (FDA). (2019, October 3). *FDA approves second drug to prevent HIV infection as part of ongoing efforts to end the HIV epidemic*.

U.S. Food & Drug Administration (FDA). (2019). *FDA In Brief: FDA continues to encourage ongoing education about the benefits and risks associated with PrEP, including additional steps to help reduce the risk of getting HIV.*

van den Hoven, J. (2015). Clustering with optimised weights for Gower's metric. *Netherlands: University of Amsterdam*.

Vanable, P. A., Carey, M. P., Blair, D. C., & Littlewood, R. A. (2006). Impact of HIV-related stigma on health behaviors and psychological adjustment among HIV-positive men and women. *AIDS and Behavior*, *10*(5), 473-482.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236-244

Willmott, C. J. (1981). On the validation of models. *Physical geography*, *2*(2), 184-194.

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of artificial intelligence research*, *6*, 1-34.

Zarwell, M., John, S. A., Westmoreland, D., Mirzayi, C., Pantalone, D. W., Golub, S., ... & Grov, C. (2020). PrEP Uptake and Discontinuation Among a US National Sample of Transgender Men and Women. *AIDS and Behavior*, 1-9.

Appendix A: Ward's 3 Cluster Solution Cluster Attribute Tables

Cluster 1 — Most Restrictive

Ward's Cluster 1 Plan Characteristics — Admin Variables

	Q1	Median	Q3
Monetary Variables			
Deductible	2250	4000	5500
Maximum Out of Pocket	6650	6750	7900
Premium/Month (Age 27)	332.87	401	486.61
	n	**%**	
Plan Factors			
Contains 2+ Tiers	1193	33.46%	

Ward's Cluster 1 Plan Characteristics — PrEP

Characteristic	Number of Plans	Percent
Coverage		
Covered	3524	98.85
Not Covered	41	1.15
Coverage Strategy		
Coins	2572	72.15
Copay	952	26.7
Not Covered	41	1.15
Prior Authorization		
No PA	3373	94.61
Requires PA	151	4.24
Not Covered	41	1.15
Tier		
Non Preferred Brand	91	2.55
Non Preferred Specialty	997	27.97
Preferred Brand	2436	68.33
Not Covered		1.15

Ward's Cluster 1 Plan Characteristics — Benefits

Benefit	Number of Plans	Percent	Q1	Median	Q3
Ambulance (AB)					
Not Covered	172	1.01			
Coins	3378	94.75	20%	30%	50%
Copay	151	4.24	$0	$5	$5
Diagnostic Tests (DT)					
Not Covered	5	0.03			
Coins	3560	99.86	20%	30%	40%
Copay	4	0.11	$0	$10	$20
Emergency Room (ER)					
Not Covered	0	0			
Coins	2353	66	20%	30%	50%
Copay	1212	34	$250	$500	$950
Generic Drugs (GD)					
Not Covered	0	0			
Coins	1489	41.77	20%	25%	40%
Copay	2076	58.23	$5	$10	$15
Habilitation Services (HA)					
Not Covered	14	0.08			
Coins	3502	98.23	20%	30%	50%
Copay	60	1.68	$20	$30	$35
Inpatient Birth (IB)					
Not Covered	0	0			
Coins	2732	76.63	20%	30%	40%
Copay	833	23.37	$500	$850	$850
Inpatient Physician (IH)					
Not Covered	191	1.12			
Coins	3517	98.65	20%	30%	40%
Copay	8	0.22	$0	$0	$0
Outpatient Mental Health (OM)					
Not Covered	0	0			
Coins	3472	97.39	20%	30%	40%
Copay	93	2.61	$15	$15	$20
Primary Care Physician (PC)					
Not Covered	0	0			
Coins	2217	62.5	20%	35%	50%
Copay	1348	37.81	$15	$30	$40
Specialist (SP)					
Not Covered	0	0			
Coins	3284	92.12	20%	30%	50%
Copay	281	7.88	$50	$60	$60

Cluster 2 — Moderately Restrictive:

Ward's Cluster 2 Plan Characteristics — Admin Variables

	Q1	Median	Q3
Monetary Variables			
Deductible	2000	3500	5900
Maximum Out of Pocket	6750	7550	7900
Premium/Month (Age 27)	346.95	419.81	502.04
	n	**%**	
Plan Factors			
Contains 2+ Tiers	167	2.64%	

Ward's Cluster 2 Plan Characteristics — PrEP

	Number of Plans	Percent
Coverage		
Covered	6224	98.34
Not Covered	105	1.66
Coverage Strategy		
Coins	2661	42.04
Copay	3563	56.3
Not Covered	105	1.66
Prior Authorization		
No PA	4359	68.87
Requires PA	1865	29.47
Not Covered	105	1.66
Tier		
Non Preferred Brand	308	4.87
Non Preferred Specialty	1440	22.75
Preferred Brand	4476	70.72
Not Covered	105	1.66

Ward's Cluster 2 Plan Characteristics — Benefits

Benefit	Number of Plans	Percent	Q1	Median	Q3
Ambulance (AB)					
Not Covered	248	1.45			
Coins	5164	81.59	20%	30%	40%
Copay	1073	16.95	$50	$250	$250
Diagnostic Tests (DT)					
Not Covered	0	0			
Coins	4612	72.87	20%	30%	40%
Copay	1717	27.13	$30	$55	$80
Emergency Room (ER)					
Not Covered	0	0			
Coins	3499	55.29	20%	35%	40%
Copay	2830	44.71	$300	$InCl	$550
Generic Drugs (GD)					
Not Covered	0	0			
Coins	275	4.35	40%	50%	100%
Copay	6054	95.65	$10	$15	$24
Habilitation Services (HA)					
Not Covered	16	0.09			
Coins	3995	63.12	20%	30%	40%
Copay	2328	36.78	$30	$40	$50
Inpatient Birth (IB)					
Not Covered	0	0			
Coins	5778	91.29	20%	30%	40%
Copay	551	8.71	$100	$500	$750
Inpatient Physician (IH)					
Not Covered	442	2.59			
Coins	5275	83.35	20%	30%	40%
Copay	890	14.06	$0	$0	$100
Outpatient Mental Health (OM)					
Not Covered	0	0			
Coins	351	5.55	30%	30%	40%
Copay	5978	94.45	$30	$30	$50
Primary Care Physician (PC)					
Not Covered	0	0			
Coins	8	0.13	30%	30%	30%
Copay	6321	99.87	$20	$30	$35
Specialist (SP)					
Not Covered	0	0			
Coins	461	7.28	40%	40%	50%
Copay	5868	92.72	$50	$60	$75

Cluster 3 — Least Restrictive

Ward's Cluster 3 Plan Characteristics — Admin Variables

	Q1	Median	Q3
Monetary Variables			
Deductible	2000	6100	7900
Maximum Out of Pocket	6000	7150	7900
Premium/Month (Age 27)	303.905	406.11	507.96
	n	**%**	
Plan Factors			
Contains 2+ Tiers	428	5.97%	

Ward's Cluster 3 Plan Characteristics — PrEP

	Number of Plans	Percent
Coverage		
Covered	7099	99.05
Not Covered	68	0.95
Coverage Strategy		
Coins	319	4.45
Copay	6780	94.6
Not Covered	68	0.95
Prior Authorization		
No PA	5878	82.01
Requires PA	1221	17.04
Not Covered	68	0.95
Tier		
Non Preferred Brand	392	5.47
Non Preferred Specialty	1237	17.26
Preferred Brand	5470	76.32
Not Covered	68	0.95

Ward's Cluster 3 Plan Characteristics — Benefits

Benefit	Number of Plans	Percent	Q1	Median	Q3
Ambulance (AB)					
Not Covered	148	0.87			
Coins	505	7.05	20%	20%	30%
Copay	6600	92.09	$0	$0	$0
Diagnostic Tests (DT)					
Not Covered	5	0.03			
Coins	6	0.08	20%	20%	20%
Copay	7159	99.89	$0	$0	$40
Emergency Room (ER)					
Not Covered	0	0			
Coins	16	0.22	30%	35%	40%
Copay	7151	99.78	$0	$0	$250
Generic Drugs (GD)					
Not Covered	0	0			
Coins	18	0.25	10%	10%	10%
Copay	7149	99.75	$0	$10	$15
Habilitation Services (HA)					
Not Covered	21	0.13			
Coins	56	0.78	20%	20%	20%
Copay	7102	99.09	$0	$0	$25
Inpatient Birth (IB)					
Not Covered	5	0.03			
Coins	8	0.11	27.50%	30%	30%
Copay	7157	99.86	$0	$0	$1
Inpatient Physician (IH)					
Not Covered	188	1.1			
Coins	6	0.08	20%	20%	27.50%
Copay	7082	98.81	$0	$0	$0
Outpatient Mental Health (OM)					
Not Covered	5	0.03			
Coins	1	0.01	50%	50%	50%
Copay	7164	99.96	$0	$0	$30
Primary Care Physician (PC)					
Not Covered	0	0			
Coinsurance	0	0			
Copay	7167	100	$0	$20	$30
Specialist (SP)					
Not Covered	0	0			
Coinsurance	0	0			
Copay	7167	100	$0	$30	$50

Appendix B: R Code

Variable Selection:

```
## ---------------------------
## Script name: 1 - Variable Selection.R
##
## Author:Sam Powers
## Date Created: 2021-02-17
##
## ---------------------------
## Purpose of script: To create the dataset that will be used in analysis for my thesis on natural groupings in healthcare plan
coverage for HIV
##  This is script #1 in the analysis
##
## ---------------------------
## set working directory

setwd("/Volumes/GoogleDrive/My Drive/School/MA/ThesisThinking/Analysis")

## ---------------------------
## load up the packages we will need:

library(tidyverse)
options(scipen = 6, digits = 4) # I prefer to view outputs in non-scientific notation

## ---------------------------
## read in data:

plans2019 <- read_csv("Data/plans_2019_raw.csv")


# Data Assumptions ------------------------------------------------------------
# We are only talking about individuals

# Benefits that we want to consider:
# HIV- Specific Set
# - SP: Specialist (Infectious Diseases Dr.)
# - DT: Diagnostic testing (HIV Labs); STIs, Kidney Function, Pregancy testing
# - PV: Preventative Care (PV)
# - OM: Outpatient Mental Health (For Behavioral Change Counseling)
# - OS: Outpatient Substance
# - PrEP coverage

## Maybe group on the above and then group on the whole?

# Generally Necessary for Monetary & Peace of Mind Purposes
# - PC: Primary Care Physician
# - AB: Ambulance
# - ER: Emergency Room
# - IH: Inpatient Physician
# - GD: Generic Drugs
# - HA: Habilitation Services
# - OH: Outpatient Physician Care
# - IB: Inpatient Birth


names(plans2019)

# What do we do with the tiering?
# Do we want tier 1, tier 2 & out of network considerations? We could just assume that everyone goes in network

plans2019 %>%
  filter(CSR == 0, CHILDONLY == 0) %>%
  pull(MULTITIERED) %>% mean()

# Only 10% of the plans are multi-tiered. I don't think that justifies keeping Tier 2 in here. Maybe just keeping an indicator
for the Multitiered variable
# In order to account for restrictiveness.

# Maybe we keep in network and out of network?
reduced_plans <-
plans2019 %>%
  filter(CSR == 0, CHILDONLY == 0) %>%
  select(YEAR,
         hios_id = PLANID,
         ST,
         AREA,
         CARRIER,
         PLANNAME,
         METAL,
         PLANTYPE,
         MULTITIERED,
         PREMI27,
         PREMI50,
         NETWORKID,
         contains(c("SP_", "DT_", "PV_", "OM_", "Ded", "MOOP",
                    "PC_", "AB_", "ER_", "IH", "GD", "HA", "OM", "IB" # The second set of usefulness
                    )
                  )
```

```
        ) %>%
  select(
    YEAR,
    hios_id,
    ST,
    AREA,
    CARRIER,
    PLANNAME,
    METAL,
    PLANTYPE,
    MULTITIERED,
    PREMI27,
    PREMI50,
    NETWORKID,
    contains("InnTier1"),
    # contains("InnTier2"),
    contains("OutofNet"),
        -contains("Complex"),
        -contains("Family")
        )

mean(plans2019$MULTITIERED)

# How do I handle the Deductibles? In the data they can be split into drug and medical.

# Data things to consider
# Restrictiveness - is there some degree of usefulness in keeping in network, out of network, and


# Bring in Truvada Formulary info ----------------------------------------

truv2019 <- read_csv("./Data/dpf2019_subset.csv") %>%
  filter(ndc_package_code %in% c("61958-0701-01"))

formulary_plans_2019 <- read_csv("Data/formulary_plans_2019.csv")

formulary_plans_2019_plus_benefits <-
plans2019 %>%
  filter(CSR == 0, CHILDONLY == 0) %>%
  select(hios_id = PLANID, contains(c("ND_", "SD_", "PD_", "GD_", "OS_")) , -contains(c("Tier2", "OutofNet", "LIMITED", "TIERS",
"Complex") ) ) %>%
  left_join(formulary_plans_2019)


formulary_plans_2019_plus_benefits_plus_truv <-
  formulary_plans_2019_plus_benefits %>%
  left_join(truv2019)

prep_coverage <-
formulary_plans_2019_plus_benefits_plus_truv %>%
  mutate(

tier = case_when(
  tier == "not_listed" ~ NA_character_,
  TRUE ~ tier
  ),

PrEP_Coverage_Type = case_when(
  tier == "non_preferred_brand" & ND_CopayInnTier1 %in% c(0,99) & ND_CoinsInnTier1 %in% c(0,99) ~ NA_character_,
  tier == "non_preferred_brand" & ND_CopayInnTier1 == 0 & ND_CoinsInnTier1 > 0 ~ "Coinsurance",
  tier == "non_preferred_brand" & ND_CopayInnTier1 >= 0 & ND_CoinsInnTier1 == 0 ~ "Copay",

  tier == "non_preferred_specialty" & SD_CopayInnTier1 %in% c(0,99) & SD_CoinsInnTier1 %in% c(0,99) ~ NA_character_,
  tier == "non_preferred_specialty" & SD_CopayInnTier1 == 0 & SD_CoinsInnTier1 > 0 ~ "Coinsurance",
  tier == "non_preferred_specialty" & SD_CopayInnTier1 >= 0 & SD_CoinsInnTier1 == 0 ~ "Copay",

  tier == "preferred_brand" & PD_CopayInnTier1 %in% c(0,99) & PD_CoinsInnTier1 %in% c(0,99) ~ NA_character_,
  tier == "preferred_brand" & PD_CopayInnTier1 == 0 & PD_CoinsInnTier1 > 0 ~ "Coinsurance",
  tier == "preferred_brand" & PD_CopayInnTier1 >= 0 & PD_CoinsInnTier1 == 0 ~ "Copay",

  tier == "generic_brand" & GD_CopayInnTier1 %in% c(0,99) & GD_CoinsInnTier1 %in% c(0,99) ~ NA_character_,
  tier == "generic_brand" & GD_CopayInnTier1 == 0 & GD_CoinsInnTier1 > 0 ~ "Coinsurance",
  tier == "generic_brand" & GD_CopayInnTier1 >= 0 & GD_CoinsInnTier1 == 0 ~ "Copay",

  TRUE ~ NA_character_
 ),

PrEP_Covered = case_when(
  tier %in% c("non_preferred_brand", "non_preferred_specialty", "preferred_brand", "generic_brand") ~ "Covered",
  TRUE ~ "Not Covered"
),

prior_authorization = case_when(
  PrEP_Covered == "Not Covered" ~ NA_character_,
  TRUE ~ as.character(prior_authorization)
)

) %>%
  select(hios_id, PrEP_Covered, PrEP_Coverage_Type, PrEP_Tier = tier, PrEP_PA = prior_authorization ) %>%
  unique()


# Final HIV Risk Factors Data ----------------------------------------------
hiv_prevention_benefits <-
reduced_plans %>%
```

```
  left_join(prep_coverage)
nrow(hiv_prevention_benefits)

write_csv(hiv_prevention_benefits, path = "Data/hiv_prevention_plus_general_benefits.csv")
```

Distance Metric

```
## ---------------------------
## Script name: 2 - Distance Matrix.
##
## Author:Sam Powers
## Date Created: 2021-04-28
##
## ---------------------------
## Purpose of script: To calculate the pairwise distances between plans.
##
##
## ---------------------------
## load up the packages we will need:

library(tidyverse)
library(data.table)

## ---------------------------
## read in data:

hiv_all <- read_csv("hiv_prevention_plus_general_benefits.csv")

names(hiv_all)

# Select only the In Network Benefits
innet_all <-
  hiv_all %>%
  mutate(id = paste0("plan", 1:n())) %>% # create the ID
  select(-contains("Outof")) %>%
  mutate(
    DEDUCTInn = case_when(
      is.na(TEHBDedInnTier1IndividualA) ~  MEHBDedInnTier1IndividualA + DEHBDedInnTier1IndividualA,
      TRUE  ~ TEHBDedInnTier1IndividualA
    ),

    MOOPInn = case_when(
      is.na(TEHBInnTier1IndividualMOOPA) ~ as.numeric( MEHBInnTier1IndividualMOOPA + DEHBInnTier1IndividualMOOPA),
      TRUE ~ TEHBInnTier1IndividualMOOPA
    )
  ) %>%
  select(-contains("EHB"))


# Recode the benefits for the nested structure to work
benefits_recoded <-
  innet_all %>%
  select(id, contains("InnTier1")) %>%
  gather(label, value, -id)  %>%
  separate(label, c("benefit", "type"), sep = "_") %>%
  separate(type, c("coverage_type", "value_label"), "Inn") %>%
  spread(value_label, value) %>%
  select(-Tier1) %>%
  spread(coverage_type, Tier1A) %>%
  mutate(type = case_when(
    !is.na(Coins) ~ "coins",
     !is.na(Copay) ~ "copay",
    TRUE ~ "not covered"
    )
    ) %>%
  gather(type_vec, values, -id, -benefit) %>%
  mutate(benefit_code = paste0(benefit, "_", type_vec)) %>%
  select(-benefit, -type_vec) %>%
  spread(benefit_code, values) %>%
  mutate(across(contains(c("Copay", "Coins") ), as.numeric   ))


# Get the Administrative Information  ------------------------------------------------------------------
plan_info <-
  innet_all %>%
  select(id, YEAR, hios_id, ST, AREA, CARRIER, PLANNAME, METAL, PLANTYPE, NETWORKID)


# Get clustering Data  ---------------------------------------------------
plan_data_final <-
  innet_all %>%
  select(id, everything(),   -contains("InnTier1"), - YEAR, - hios_id, -ST, -AREA, -CARRIER, - PLANNAME, - METAL, - PLANTYPE, -
NETWORKID, - PREMI50   ) %>%
  left_join(benefits_recoded) %>%
  select( -contains("PV")) %>%                # Remove preventative care because it has no variance to it.
  mutate(PrEP_PA = as.character(PrEP_PA)) %>%
  mutate(PrEP_Coverage_Type = str_sub(str_to_lower(PrEP_Coverage_Type), 1, 5) ) %>%
  filter(!is.na(DEDUCTInn), !is.na(MOOPInn))

nrow(innet_all)    -
nrow(plan_data_final)

benefit_types <-
names(plan_data_final)[grepl("type", str_to_lower(names(plan_data_final))  )]
```

```
# Check for missing data
map_df(benefit_types,
       ~  tibble(var = .x,
                 na = sum(is.na(plan_data_final[, .x])))
       )

map_df(benefit_types,
       ~  tibble(var = .x,
                 na = sum(plan_data_final[, .x] == "not covered"  ))
)


map_df(c("DEDUCTInn", "MOOPInn"),
       ~  tibble(var = .x,
                 na = sum(is.na(plan_data_final[, .x]) ))
)



# Distance calculations ----------------------------------------------------
trial <-
  plan_data_final

# Create all pairwise combinations of plans.
combos <-
  CJ(1:nrow(trial), 1:nrow(trial), unique = TRUE) %>%
  filter(V2 > V1) %>%
  mutate(id = 1:n())

# Extract the numeric cols
numeric_cols <- trial %>%
  select(where(is.numeric)) %>%
  names()

# Extract the character cols
character_cols <- trial %>%
  select(where(is.character), -id) %>%
  names()

length(names(plan_data_final))
length(numeric_cols)
length(character_cols)

# Get the ranges for the numeric variables
max_na <- function(x) max(x, na.rm = TRUE)
min_na <- function(x) min(x, na.rm = TRUE)

ranges <-
  data.frame(t(apply(plan_data_final[,numeric_cols], 2, max_na ) -
                 apply(plan_data_final[,numeric_cols], 2, min_na)))

range_numeric <- ranges[rep(1, nrow(combos)), ] # create a matrix of the ranges with the right number of rows equivalent to the
number of pairwise combinations

numeric_distances <-   # Calculate the range-scaled difference for all of the pairwise combinations.
  abs(trial[c(combos$V1), numeric_cols] - trial[c(combos$V2), numeric_cols])/range_numeric %>%
  as_tibble()

categorical_distances <-  # check for equality in the character vals in all possible combos. Set to 0 if equal and 1 if else
  (1 -  (trial[c(combos$V1), character_cols] == trial[c(combos$V2), character_cols])  ) %>%
  as_tibble()

sum_na <- function(x) sum(x, na.rm  = TRUE)
non_na <- function(x) sum(!is.na(x))

combos$numeric_distance <- apply(numeric_distances, 1, sum_na) # sum of numeric distances
combos$numeric_denom <-  apply(numeric_distances, 1, non_na)   # denominator of numeric distances

combos$categoric_distance <- apply(categorical_distances, 1, sum_na) # sum of categorical distances
combos$categoric_denom <-  apply(categorical_distances, 1, non_na)   # denominator of categorical distances

combos$total_distance = combos$numeric_distance + combos$categoric_distance # total distance
combos$total_denom =  combos$numeric_denom + combos$categoric_denom         # total denominator

combos$distance = combos$total_distance/combos$total_denom  # average distance

# Create the distance matrix
dist_mat <-
  combos %>%
  select(V1, V2, distance) %>%
  spread(V2, distance) %>%
  as.matrix()

dist_mat <- dist_mat[,-1]

dist_mat <- cbind( rep(0, nrow(dist_mat)),
                   dist_mat)

dist_mat <- rbind(dist_mat,
                  rep(0, ncol(dist_mat))
)

dist_mat[is.na(dist_mat)] <- 0

dist_mat_final = dist_mat + t(dist_mat)
```

```
colnames(dist_mat_final) <- trial$id
rownames(dist_mat_final) <- trial$id

save(dist_mat_final, file = "distance_matrix_all_benefits.Rdata")
```

Clustering:

```
## ---------------------------
## Script name: 3 - Cluster.R
##
## Author:Sam Powers
## Date Created: 2021-04-28
##
## ---------------------------
## Purpose of script: To run the clustering trials with the complete distance matrix.
##
##
## ---------------------------
## load up the packages we will need:

library(tidyverse)
library(cluster)
library(fastcluster)
library(fpc)

## ---------------------------
## read in data:

load("distance_matrix_all_benefits.Rdata")


# Complete Linkage ----------------------------------------------------------

dist_mat_all <- dist_mat_final

dist_mat_all_use <- as.dist(dist_mat_all)

clust_all <- hclust(dist_mat_all_use)

# Cophenetic Distance
cc_all <-
  cophenetic(clust_all)

cc_cor_all  <-
  cor(dist_mat_all_use, cc_all)

cc_cor_all  #  0.7343951
# Index of Agreement

IOA <- function(distance, cophenetic_distance) {
  o_bar <- mean(distance)
  numerator <- sum((cophenetic_distance - distance)**2)
  denom <- sum((abs(cophenetic_distance - o_bar) + abs(distance - o_bar))**2 )

  ioa <- 1 - (numerator/denom)

  print(ioa)
}

IOA_all <-
  IOA(dist_mat_all_use, cc_all)

IOA_all #  0.4992368

# selection criteria ------------------------------------------------------

silhouette_df_all <-
  tibble(
    k = 2:20
  ) %>%
  mutate(
    sil = map(k, ~silhouette(cutree(clust_all, k = .x), dist_mat_all_use)),
    sil_sum = map(sil, ~summary(.x)),
    avg_sil = as.numeric(map_chr(sil_sum, ~.x["avg.width"][1][[1]]))
  )


ggplot(silhouette_df_all, aes(x  = k, y = avg_sil)) +
  geom_point() +
  geom_line() +
  labs(x = "Clusters", y = "Average Silhouette Distance") +
  scale_y_continuous(
    limits = c(0, .5),
    breaks = seq(0, .5, .1)
    ) +
  scale_x_continuous(limits = c(0, 20), breaks = seq(0, 20, 1))

silhouette_df_all
# Okay, so it loads into 2,3 or 4 clusters nicely.

complete_linkage <-
```

```
map_df(2:4,
~cutree(clust_all, k = .x) %>%
  data.frame() %>%
  rownames_to_column() %>%
  tibble() %>%
  rename(
    plan_id = rowname,
 #   cluster = `.`
  ) %>%
  mutate(algorithm = paste0("complete", .x))
) %>%
  spread(algorithm, 2)


table(complete_linkage$complete2)
table(complete_linkage$complete3)
table(complete_linkage$complete4)


# Try a PAM -----------------------------------------------------------------

# PAM Method
sil_width_all <- c(NA)
for(i in 2:10){
  pam_fit <- pam(dist_mat_all_use, diss = TRUE, k = i)
  sil_width_all[i] <- pam_fit$silinfo$avg.width
  print(sil_width_all)
}


plot(1:length(sil_width_all), sil_width_all,
     xlab = "Number of clusters",
     ylab = "Silhouette Width")
lines(1:length(sil_width_all), sil_width_all)

ggplot(tibble(k = 1:length(sil_width_all), avg_sil =sil_width_all ), aes(x  = k, y = avg_sil)) +
  geom_point() +
  geom_line() +
  labs(x = "Clusters", y = "Average Silhouette Distance") +
  scale_y_continuous(
    limits = c(0, .5),
    breaks = seq(0, .5, .1)
  ) +
  scale_x_continuous(limits = c(0, 10), breaks = seq(0, 20, 1))


# This is into 2 or 3 max.

pam_fit_2 <- pam(dist_mat_all_use, diss = TRUE, k = 2)
pam_fit_3 <- pam(dist_mat_all_use, diss = TRUE, k = 3)

pam_fit_2$silinfo$avg.width
pam_fit_3$silinfo$avg.width

pams <-
pam_fit_2$clustering %>%
data.frame() %>%
  rownames_to_column() %>%
  tibble() %>%
  rename(
    plan_id = rowname,
    pam2 = 2
  )  %>%
  left_join(

pam_fit_3$clustering %>%
  data.frame() %>%
  rownames_to_column() %>%
  tibble() %>%
  rename(
    plan_id = rowname,
    pam3 = 2
  )
)

# Try a different hierarchical joining method -------------------------------

clust_ward <- hclust(dist_mat_all_use, method = "ward.D2")
#plot(full_clust)

# Cophenetic Distance
cc_ward <-
  cophenetic(clust_ward)

cc_cor_ward  <-
  cor(dist_mat_all_use, cc_ward)

cc_cor_ward  #  0.6830316
# Index of Agreement

IOA <- function(distance, cophenetic_distance) {
  o_bar <- mean(distance)
  numerator <- sum((cophenetic_distance - distance)**2)
  denom <- sum((abs(cophenetic_distance - o_bar) + abs(distance - o_bar))**2 )
```

```
  ioa <- 1 - (numerator/denom)

  print(ioa)
}

IOA_ward <-
  IOA(dist_mat_all_use, cc_ward)

IOA_ward # 0.01360921 Ooh, this isnt good.

# selection criteria ------------------------------------------------------

silhouette_df_ward <-
  tibble(
    k = 2:20
  ) %>%
  mutate(
    sil = map(k, ~silhouette(cutree(clust_ward, k = .x), dist_mat_all_use)),
    sil_sum = map(sil, ~summary(.x)),
    avg_sil = as.numeric(map_chr(sil_sum, ~.x["avg.width"][1][[1]]))
  )

ggplot(silhouette_df_ward, aes(x  = k, y = avg_sil)) +
  geom_point() +
  geom_line() +
  labs(x = "Clusters", y = "Average Silhouette Distance") +
  scale_y_continuous(
    limits = c(0, .5),
    breaks = seq(0, .5, .1)
  ) +
  scale_x_continuous(limits = c(0, 20), breaks = seq(0, 20, 1))


ward_linkage <-
  map_df(2:5,
         ~cutree(clust_ward, k = .x) %>%
           data.frame() %>%
           rownames_to_column() %>%
           tibble() %>%
           rename(
             plan_id = rowname,
             #   cluster = `.`
           ) %>%
           mutate(algorithm = paste0("ward", .x))
  ) %>%
  spread(algorithm, 2)


# Try the average method --------------------------------------------------

clust_avg <- hclust(dist_mat_all_use, method = "average")
#plot(full_clust)

# Cophenetic Distance
cc_avg <-
  cophenetic(clust_avg)

cc_cor_avg  <-
  cor(dist_mat_all_use, cc_avg)

cc_cor_avg  #  0.7814216

# Index of Agreement

IOA <- function(distance, cophenetic_distance) {
  o_bar <- mean(distance)
  numerator <- sum((cophenetic_distance - distance)**2)
  denom <- sum((abs(cophenetic_distance - o_bar) + abs(distance - o_bar))**2 )

  ioa <- 1 - (numerator/denom)

  print(ioa)
}

IOA_avg <-
  IOA(dist_mat_all_use, cc_avg)

IOA_avg #   0.8680168 this great!

# selection criteria ------------------------------------------------------

silhouette_df_avg <-
  tibble(
    k = 2:10
  ) %>%
  mutate(
    cut_tree = map(k, ~cutree(clust_avg, k = .x)),
    sil = map(cut_tree, ~silhouette(.x, dist_mat_all_use)),
    sil_sum = map(sil, ~summary(.x)),
    avg_sil = as.numeric(map_chr(sil_sum, ~.x["avg.width"][1][[1]]))
  )

silhouette_df_avg$cut_tree[[1]]
```

```
ggplot(silhouette_df_avg, aes(x  = k, y = avg_sil)) +
  geom_point() + # 5 comes back around
  geom_line() +
  labs(x = "Clusters", y = "Average Silhouette Distance") +
  scale_y_continuous(
    limits = c(0, .5),
    breaks = seq(0, .5, .1)
  ) +
  scale_x_continuous(limits = c(0, 20), breaks = seq(0, 20, 1))

average_linkage <-
  map_df(2:5,
         ~cutree(clust_avg, k = .x) %>%
           data.frame() %>%
           rownames_to_column() %>%
           tibble() %>%
           rename(
             plan_id = rowname,
             #   cluster = `.`
           ) %>%
           mutate(algorithm = paste0("average", .x))
  ) %>%
  spread(algorithm, 2)

# Try the single method ---------------------------------------------------

clust_single <- hclust(dist_mat_all_use, method = "single")
#plot(full_clust)

# Cophenetic Distance
cc_single <-
  cophenetic(clust_single)

cc_cor_single  <-
  cor(dist_mat_all_use, cc_single)

cc_cor_single  #   0.3326598
# Index of Agreement

IOA <- function(distance, cophenetic_distance) {
  o_bar <- mean(distance)
  numerator <- sum((cophenetic_distance - distance)**2)
  denom <- sum((abs(cophenetic_distance - o_bar) + abs(distance - o_bar))**2 )

  ioa <- 1 - (numerator/denom)

  print(ioa)
}

IOA_single <-
  IOA(dist_mat_all_use, cc_single)

IOA_single #  0.4281735 t his eh!

# selection criteria ------------------------------------------------------
sil <- NULL

silhouette_df_single <-
  tibble(
    k = 2:20
  ) %>%
  mutate(
    sil = map(k, ~silhouette(cutree(clust_single, k = .x), dist_mat_all_use)),
    sil_sum = map(sil, ~summary(.x)),
    single_sil = as.numeric(map_chr(sil_sum, ~.x["avg.width"][1][[1]]))
  )

ggplot(silhouette_df_single, aes(x  = k, y = single_sil)) +
  geom_point() +
  geom_line() +
  labs(x = "Clusters", y = "Average Silhouette Distance") +
  scale_y_continuous(
    limits = c(-.5, .5),
    breaks = seq(-.5, .5, .1)
  ) +
  scale_x_continuous(limits = c(0, 20), breaks = seq(0, 20, 1))


# Join them together ------------------------------------------------------
cluster_assignments <-
complete_linkage %>%
  left_join(pams) %>%
  left_join(average_linkage) %>%
  left_join(ward_linkage)

write_csv(cluster_assignments, path = "cluster_assignments.csv")
```

Plots and Tables:

```
## ---------------------------
## Script name: 4 - Plots and Tables
##
## Author:Sam Powers
## Date Created: 2021-04-28
##
## ---------------------------
## Purpose of script: To create the plots and tables for the results section of my thesis.
##
##
## ---------------------------
## set working directory

setwd()

## ---------------------------
## load up the packages we will need:

library(tidyverse )
library(urbnmapr)
library(sf)


## ---------------------------
## read in data:

plan_data_final <- read_csv("plan_data_final.csv")
cluster_asignments <- read_csv("cluster_assignments.csv")
clustered_plans <- read_csv("clustered_plans_final.csv")

# Number of Plans ------------------------------------------------------------

nrow(plan_data_final)

 # Table 1: Benefits ---------------------------------------------------------------------
benefit_labels <-
  tibble(
    benefit  =
      c(
        "SP",
        "DT",
        "PV",
        "OM",
        "OS",
        "PC",
        "AB",
        "ER",
        "IH",
        "GD",
        "HA",
        "OH",
        "IB"
      ),
    label =
      c(

        "Specialist",
        "Diagnostic Tests",
        "Preventative Care",
        "Outpatient Mental Health",
        "Outpatient Substance",
        "Primary Care Physician",
        "Ambulance",
        "Emergency Room",
        "Inpatient Physician",
        "Generic Drugs",
        "Habilitation Services",
        "Outpatient Physician Care",
        "Inpatient Birth"
      )
  )

table1_benefit_codes <-
plan_data_final %>%
  select(contains(c("Coins", "Copay")   ), -contains("PrEP")) %>%
  gather(benefit, value ) %>%
  separate(benefit, c("benefit", "type"), sep = "_" ) %>%
  group_by(benefit, type) %>%
  filter(!is.na(value)) %>%
  summarize(count = n(),
            low = quantile(value, .25, na.rm = TRUE),
            med = quantile(value, .5, na.rm = TRUE),
            hi = quantile(value, .75, na.rm = TRUE)
  ) %>%
  mutate(pct = count /nrow(plan_data_final)   ) %>%
  group_by(benefit) %>%
  mutate( `Not Covered` = 1- sum(pct)) %>%
  spread(type, pct) %>%
  gather(type, pct, -c(benefit:hi)) %>%
  arrange(benefit) %>%
  filter(!is.na(pct)) %>%
```

```
  ungroup() %>%
  mutate(count =
           case_when( type == "Not Covered" ~ round(nrow(plan_data_final)* pct),
                      TRUE ~ as.numeric(count)
           )
  ) %>%
  mutate(
    across( c(low, med, hi), ~case_when(type == "Not Covered" ~ NA_real_, TRUE ~ .x) )
  ) %>%
  distinct() %>%
  mutate(pct = round(pct*100, 2)) %>%
  select(benefit, type, count, pct, low, med, hi) %>%
  split(.$benefit) %>%
  map_dfr(., ~rbind( c(unique(.x$benefit), NA, NA, NA, NA, NA, NA), .x   )
  ) %>%
  left_join(benefit_labels) %>%


  mutate(
    across(c(low, med, hi), ~case_when(type == "Copay" ~ paste0("$",.x),
                                       type == "Coins" ~ paste0(.x, "%")

       )
    ),

    type = case_when(
      is.na(type)  ~ paste0(label, " (", benefit, ")")  ,
      TRUE ~ paste0("   ",type)
    )
  ) %>%
  select(-benefit, -label) %>%
  mutate(across(everything(), as.character),
         across(everything(), ~case_when(is.na(.x) ~ "",
                                         TRUE ~ .x))
         ) %>%
  select(Benefit = 1, `Number of Plans` = 2, `Percent` = 3, Q1 = 4, Median = 5, Q3 = 6)

table1_benefit_codes
write_csv(table1_benefit_codes, path = "table1_benefits.csv")


plan_data_final %>%
  select(DT_Copay) %>%
  filter(!is.na(DT_Copay)) %>%
  summarize( mean(DT_Copay == 0))



# Table 1: PrEP Characteristics ------------------------------------------

table1_prep <-
plan_data_final %>%
  select(contains("PrEP")) %>%
  gather(var, level) %>%
  group_by(var, level) %>%
  summarize(count = n()) %>%
  ungroup() %>%
  mutate(level =
    case_when(
      is.na(level) ~ "Not Covered",
      level == "TRUE" ~ "Requires PA",
      level == "FALSE" ~ "No PA",
      TRUE ~ str_to_title(  str_replace_all(level, "_", " "))
      ),

      var = case_when(
        var == "PrEP_PA" ~ "Prior Authorization",
        var == "PrEP_Coverage_Type" ~ "Coverage Strategy",
        var == "PrEP_Covered" ~ "Coverage",
        TRUE ~ str_replace_all(var, "PrEP_|", "")
      )
    ) %>%
  arrange(var) %>%
  mutate(
    pct = round(count/nrow(plan_data_final) *100,2 )
  ) %>%
  split(.$var) %>%
  map_df(.,  ~rbind(  c( unique(.x$var), NA, NA, NA), .x)

  ) %>%
  mutate(across(everything(), as.character),
         across(everything(), ~case_when(is.na(.x) ~ "", TRUE ~ .x)),
         level = case_when(
           level == "" ~ paste0(level, var),
           TRUE ~ paste0("   ", level)

         )
         ) %>%
  select(Characteristic = level, `Number of Plans` = count, Percent = pct)

write_csv(table1_prep, path = "table1_prep.csv")



# Table 1: Admin ---------------------------------------------------------
```

```
plan_data_final %>%
  select(
    PREMI27,
    DEDUCTInn,
    MOOPInn
  ) %>%
  gather(label, value) %>%
  group_by(label) %>%
  summarize(
    low = quantile(value, .25, na.rm = TRUE),
    med = quantile(value, .5, na.rm = TRUE),
    hi = quantile(value, .75, na.rm = TRUE)
  )

mean(plan_data_final$MULTITIERED)
sum(plan_data_final$MULTITIERED)

names(plan_data_final)



# Cluster Sizes ------------------------------------------------------------

clusters_considered <-
names(cluster_assignments)[c(2, 5, 6, 7, 10, 11, 12)]


cluster_sizes <-
  map_df(clusters_considered,

         ~  cluster_assignments %>%
           group_by(cluster = !!sym(.x)   ) %>%
           count() %>%
           mutate(cluster_method = .x) %>%
           spread(cluster, n)
  )

cluster_sizes %>%
  arrange(cluster_method) %>%
  mutate(
    number_clust = str_sub(cluster_method, -1, -1),
    method = str_to_sentence(str_sub(cluster_method, 1, -2))
  ) %>%
  mutate(across(where(is.numeric), as.character))  %>%
  mutate(across(everything(), ~case_when(is.na(.x) ~"", TRUE ~.x))  ) %>%
  select(Method = method, Clusters = number_clust, 2:6) %>%
  write_csv(., path = "cluster_sizes.csv")


clustered_plans %>%
  filter(average5 == 5) %>%
  View()

(141 + 8)/nrow(clustered_plans)


3527/(13534 + 3527)
6499/(10562 + 6499)


# Plan Characteristic Graphics ---------------------------------------------

clusters_considered2 <- c("noclusters1", clusters_considered )[c(1, 5, 2, 3, 4, 7, 8)]
clusters_considered2

## Coinsurance Prevalence
coins_copay_data_prep <-
  clustered_plans %>%
  select(contains("type"), all_of(clusters_considered2), -PLANTYPE) %>%
  mutate(
    across(where(is.character), str_to_lower),
    across(contains("type"), ~ case_when(.x == "copay" ~ 0, TRUE ~ 1))
  )

cluster_results <- map_df(  clusters_considered2,
                            ~ coins_copay_data_prep %>%
                              group_by(cluster = !!sym(.x ) )%>%
                              summarize(across(contains("type"), mean), n = n()) %>%
                              mutate(grouping_type = .x) %>%
                              select(grouping_type, cluster, everything())
) %>%
  gather(benefit, coins_pct, -grouping_type, -cluster, -n) %>%
  separate(benefit, c("benefit", NA), sep = "_") %>%
  mutate(cluster = paste0("Cluster ", cluster)) %>%
  mutate(
    grouping_type =
      factor(
        grouping_type,
        levels = c(
          "noclusters1",
          "average2",
          "complete2",
          "ward2",
          "ward3",
```

```
          "pam2",
          "pam3"

        )
      )
  )

ggplot(cluster_results, # %>%
         #filter(grepl("pam|noclusters", grouping_type)),
       aes(
         x = benefit,
         y  = coins_pct,
         fill = as.factor(cluster)
       )) +
  geom_col(position = "dodge") +
  geom_text(aes(x = "IB", y = .7, label = n), hjust = .5       ) +
  coord_cartesian(clip = "off") +
  facet_grid(grouping_type ~ cluster) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Benefit", y = "Plans With Coinsurance, %") +
  scale_y_continuous(limits = c(0, 1),
                     labels = function(x) paste0(round(x*100), "%")
                     ) +
  theme(
    legend.position = "none"
  )



# Ward's 3 Cluster Solution ------------------------------------------------


# Get Copay/Coins Amounts
coins_copay_amounts <-
  clustered_plans %>%
  select(contains(c("_Coins", "_Copay")   ), all_of(clusters_considered2)) %>%
  mutate(
    across(contains("type"), ~ case_when(.x == "copay" ~ 0, TRUE ~ 1))
  )


copay_coins_vals <- map_df(  clusters_considered2,
                             ~ coins_copay_amounts %>%
                               group_by(cluster = !!sym(.x ) )%>%
                               summarize(across(contains(c("_Coins", "_Copay")),
                                                list( low = ~quantile(.x, .25, na.rm = TRUE),
                                                      med = ~quantile(.x, .5, na.rm = TRUE),
                                                      hi = ~quantile(.x, .75, na.rm = TRUE)

                                                      # sd = ~sd(.x, na.rm = TRUE)
                                                )
                               )
                               ) %>%
                               mutate(cluster_strategy = .x) %>%
                               select(cluster_strategy, everything())
)  %>%
  gather( benefit, number, -cluster_strategy, -cluster
  ) %>%
  separate(benefit, c("benefit", "type", "bound"), sep = "_")



ward3_benefit <-
  cluster_results %>%
  filter(grouping_type == "ward3" | grouping_type == "noclusters1") %>%
  mutate(type = "Plans With Coins, %")  %>%
  filter(!benefit == "PrEP") %>%
  mutate(coins_pct = coins_pct *100)

ward3_vals <-
copay_coins_vals %>%
  filter(cluster_strategy == "ward3" | cluster_strategy == "noclusters1") %>%
  mutate(cluster = paste0("Cluster ", cluster),
         type = case_when(
           type == "Coins" ~ "Coins Amount, %",
           type == "Copay" ~ "Copay Amount, $"
         )


          ) %>%
  spread(bound, number)  %>%
  rename(
    grouping_type = cluster_strategy
  )

ward3_facts <-
  ward3_benefit %>%
  bind_rows(ward3_vals) %>%
mutate(cluster =
  case_when(
    grouping_type == "noclusters1" ~ "No Cluster",
    cluster == "Cluster 1" ~ "Cluster 2",
    cluster == "Cluster 2" ~ "Cluster 1",
    TRUE ~ cluster
      )
```

```
  )%>%
  mutate(
    cluster  = factor(cluster, levels = c("No Cluster", "Cluster 1", "Cluster 2", "Cluster 3"))
  )

# Copay/Coins Amounts
ggplot(ward3_facts, aes(x = benefit, color = cluster, fill = cluster)) +
  geom_col(aes(y = coins_pct) ) +
  geom_point(aes(y  = med)) +
  geom_errorbar(aes(ymin = low, ymax = hi)) +
  facet_grid(type ~ cluster, scales = "free_y", switch = "y") +
 # scale_y_continuous(position = "right") +
  coord_cartesian(clip = "off") +
  labs( y = "", x = "") +
  theme(
    legend.position = "none",
    strip.background = element_blank(),
          strip.placement = "outside"
  )


## How they handle PrEP
clustered_plans$PrEP_PA
prep_info <-
  clustered_plans %>%
  select(contains("PrEP"), all_of(cluster_types)) %>%
  mutate(
    across(where(is.character), str_to_lower),
    across(contains("type"), ~ case_when(.x == "copay" ~ 0, TRUE ~ 1)),
    across(contains("Covered"), ~ case_when(.x == "covered" ~ 0, TRUE ~ 1)),
    across(contains("_PA"), ~ case_when(.x == "FALSE" ~ 0, TRUE ~ 1)),
    across(contains("Tier"), ~ case_when(.x == "non_preferred_specialty" ~ 1, TRUE ~ 0))
  )
unique(prep_info$pam2)


prep_results <- map_df(  clusters_considered2,
                         ~ prep_info %>%
                           group_by(cluster = !!sym(.x ) )%>%
                           summarize(across(contains("PrEP"), mean), n = n()) %>%
                           mutate(grouping_type = .x) %>%
                           select(grouping_type, cluster, everything())
) %>%
  gather(benefit, pct, -grouping_type, -cluster, -n) %>%
  filter(
    grouping_type == "ward3" | grouping_type == "noclusters1"
  ) %>%
  mutate(
    cluster = paste0("Cluster ", cluster),
  cluster =
            case_when(
              grouping_type == "noclusters1" ~ "No Cluster",
              cluster == "Cluster 1" ~ "Cluster 2",
              cluster == "Cluster 2" ~ "Cluster 1",
              TRUE ~ cluster
            ),

  benefit = case_when(
    benefit == "PrEP_PA" ~ "Prior Auth",
    benefit == "PrEP_Coverage_Type" ~ "Coinsurance",
    benefit == "PrEP_Covered" ~ "Not Covered",
    benefit == "PrEP_Tier"  ~ "Specialty",
    TRUE ~ str_replace_all(benefit, "PrEP_|", "")

  ),

  benefit = factor(benefit, levels = c("Not Covered", "Coinsurance", "Prior Auth", "Specialty" )),
  cluster  = factor(cluster, levels = c("No Cluster", "Cluster 1", "Cluster 2", "Cluster 3"))

)


 prep_results

ggplot(prep_results %>%
         filter(grepl("ward|noclusters", grouping_type)),
       aes(
         x = benefit,
         y  = pct,
         fill = as.factor(cluster)
       )) +
  geom_col(position = "dodge") +
  # geom_text(aes(x = "IB_type", y = .7, label = n), hjust = .5       ) +
  coord_cartesian(clip = "off") +
  facet_grid(~cluster ) +
  theme(axis.text.x = element_text(angle = 90)) +
  theme(
    legend.position  = "none"
  ) +
  labs(y = "Plans With Characteristic, %",  x = "" ) +
  scale_y_continuous( labels = function(x) paste0(x*100,"%"))


## Admin Vars
admin_vars <-
```

```
deduct_prem %>%
  gather(cluster_type, cluster_number,-contains(money_vars)) %>%
  mutate(
    cluster_strategy = str_sub(cluster_type, 1,-2),
    num_clusters = str_sub(cluster_type,-1,-1)
  ) %>%
  filter(cluster_type == "ward3" |
           cluster_type == "noclusters1") %>%
  mutate(
  cluster = paste0("Cluster ", cluster_number),
  cluster = case_when(
    cluster_type == "noclusters1" ~ "No Cluster",
              cluster == "Cluster 1" ~ "Cluster 2",
              cluster == "Cluster 2" ~ "Cluster 1",
              TRUE ~ cluster
            ),

    cluster  = factor(
      cluster,
      levels = c("No Cluster", "Cluster 1", "Cluster 2", "Cluster 3")
    )

  )  %>%
  rename(Deductible = DEDUCTInn,
         `Premium/Month` = PREMI27,
         MOOP = MOOPInn) %>%
  select(Deductible, `Premium/Month`, MOOP, cluster) %>%
  gather(money,
         amount,
         -cluster)

  ggplot(admin_vars, aes(x = cluster, y = amount, fill = cluster)) +
  geom_boxplot() +
  facet_grid(money ~ cluster, scales = "free")+
    labs(x = "", y = "") +
    scale_y_continuous(labels = function(x) paste0("$", x)) +
    theme(
      legend.position = "none"
    )



# Table 1's for the Clusters ----------------------------------------------

## Benefits
clustered_table1_benefit_codes <-
 map(1:3,

   ~ clustered_plans %>%
    filter(ward3 == .x) %>%
    group_by(ward3) %>%
    mutate(denom = n()) %>%
    select(contains(c("Coins", "Copay")   ), -contains("PrEP"), ward3, denom) %>%
    gather(benefit, value,  -ward3, -denom ) %>%
    separate(benefit, c("benefit", "type"), sep = "_" ) %>%
    group_by(ward3, denom, benefit, type) %>%
    filter(!is.na(value)) %>%
    summarize(count = n(),
              low = quantile(value, .25, na.rm = TRUE),
              med = quantile(value, .5, na.rm = TRUE),
              hi = quantile(value, .75, na.rm = TRUE)
    ) %>%
    mutate(pct = count /denom   ) %>%
    ungroup() %>%
    select(-denom) %>%
    group_by(ward3, benefit) %>%
    mutate( `Not Covered` = 1- sum(pct)) %>%
    spread(type, pct) %>%
    gather(type, pct, -c(ward3,benefit:hi)) %>%
    arrange(benefit) %>%
    filter(!is.na(pct)) %>%
    ungroup() %>%
    mutate(count =
             case_when( type == "Not Covered" ~ round(nrow(plan_data_final)* pct),
                        TRUE ~ as.numeric(count)
             )
    ) %>%
    mutate(
      across( c(low, med, hi), ~case_when(type == "Not Covered" ~ NA_real_, TRUE ~ .x) )
    ) %>%
    distinct() %>%
    mutate(pct = round(pct*100, 2)) %>%
    select(ward3, benefit, type, count, pct, low, med, hi) %>%
    split(.$benefit) %>%
    map_dfr(., ~rbind( c(unique(.x$ward3), unique(.x$benefit), NA, NA, NA, NA, NA, NA), .x   )
    ) %>%
    left_join(benefit_labels) %>%
    mutate(
      across(c(low, med, hi), ~case_when(type == "Copay" ~ paste0("$",.x),
                                         type == "Coins" ~ paste0(.x, "%")

      )
      ),

      type = case_when(
```

```
        is.na(type)  ~ paste0(label, " (", benefit, ")")  ,
        TRUE ~ paste0("   ",type)
      )
    ) %>%
    select(-benefit, -label) %>%
    mutate(across(everything(), as.character),
           across(everything(), ~case_when(is.na(.x) ~ "",
                                           TRUE ~ .x))
    ) %>%
    select(Cluster = 1, Benefit = 2, `Number of Plans` = 3, `Percent` = 4, Q1 = 5, Median = 6, Q3 = 7)
 )


  walk(1:3,
  ~  write_csv(clustered_table1_benefit_codes[[.x]], path = paste0("cluster", .x, "_table1_benefits.csv")  )
  )



### PrEP Characteristics

ward_table1_prep <-
 map(1:3,

  ~ clustered_plans %>%
    filter(ward3 == .x)  %>%
    select(contains("PrEP")) %>%
    gather(var, level) %>%
    group_by(var, level) %>%
    summarize(count = n()) %>%
    ungroup() %>%
    mutate(level =
             case_when(
               is.na(level) ~ "Not Covered",
               level == "TRUE" ~ "Requires PA",
               level == "FALSE" ~ "No PA",
               TRUE ~ str_to_title(  str_replace_all(level, "_", " "))
             ),

           var = case_when(
             var == "PrEP_PA" ~ "Prior Authorization",
             var == "PrEP_Coverage_Type" ~ "Coverage Strategy",
             var == "PrEP_Covered" ~ "Coverage",
             TRUE ~ str_replace_all(var, "PrEP_|", "")
           )
    ) %>%
    arrange(var) %>%
    mutate(
      pct = round(count/nrow(clustered_plans %>% filter(ward3 == .x) ) *100,2 )
    ) %>%
    split(.$var) %>%
    map_df(.,  ~rbind(  c( unique(.x$var), NA, NA, NA), .x)

    ) %>%
    mutate(across(everything(), as.character),
           across(everything(), ~case_when(is.na(.x) ~ "", TRUE ~ .x)),
           level = case_when(
             level == "" ~ paste0(level, var),
             TRUE ~ paste0("   ", level)

           )
    ) %>%
    select(Characteristic = level, `Number of Plans` = count, Percent = pct)
 )

ward_table1_prep

walk(1:3,
     ~  write_csv(ward_table1_prep[[.x]], path = paste0("cluster", .x, "_table1_prep.csv")  )
)



# Admin Variables ------------------------------------------------------------

ward_admin_table1 <-
map(1:3,

~ tibble(
  label = "Monetary Variables",
  Q1 = "",
  Median = "",
  Q3 = ""
) %>%
  bind_rows(

clustered_plans %>%
  filter(ward3 == .x) %>%
  select(
    PREMI27,
    DEDUCTInn,
    MOOPInn
  ) %>%
```

```
  gather(label, value) %>%
  group_by(label) %>%
  summarize(
    Q1 = as.character(quantile(value, .25, na.rm = TRUE)),
    Median = as.character(quantile(value, .5, na.rm = TRUE)),
    Q3 = as.character(quantile(value, .75, na.rm = TRUE) )
  ) %>%
  mutate(
    label = case_when(
      label == "DEDUCTInn" ~ "   Deductible",
      label == "PREMI27" ~ "   Premium/Month (Age 27)",
      label == "MOOPInn" ~ "   Maximum Out of Pocket"
     )
    )
) %>%
  bind_rows(
    tibble(
    label = c("", "", "Plan Factors", "   Contains 2+ Tiers"),
    Q1 = c("", "n", "", clustered_plans %>% filter(ward3 == .x)  %>% pull(MULTITIERED) %>% sum()  ),
    Median = c("", "%", "", paste0(round(clustered_plans %>% filter(ward3 == .x)  %>% pull(MULTITIERED) %>% mean() *100, 2), "%")
),
    Q3 = c("", "", "", "")
    )
  )

)


walk(1:3,
     ~  write_csv(ward_admin_table1[[.x]], path = paste0("cluster", .x, "_table1_admin.csv")  )
)




# Percentage of plans ------------------------------------------------------
clust_sizes <- c(3565, 6329, 7167)
clust_sizes/sum(clust_sizes) # 0.2089561 0.3709630 0.4200809



# Map ----------------------------------------------------------------------

rating_county <- read_csv("rating_area_county.csv") %>%
  rename(AREA = rating_area_id)

nrow(rating_county) # 3190
length(unique(rating_county$AREA)) # 502
length(unique(rating_county$fips_code)) # 3142

rating_county %>%
  select(AREA, fips_code) %>%
  group_by(fips_code) %>%
  mutate(count = n()) %>%
  filter(count >1 )

restrictiveness_data <-
  clustered_plans %>%
  select(id, ward3) %>%
  left_join(plan_info %>%
              select(id, AREA)
  ) %>%
  group_by(AREA, ward3) %>%
  count() %>%
  spread(ward3, n, fill = 0) %>%
  rename(restrictive  = `2`, less_restrictive = `1`, not_restrictive = `3`) %>%
  mutate( total =  restrictive +  less_restrictive + not_restrictive  ) %>%
  mutate(across(contains("restrictive"), ~.x/total*100)) %>%
  left_join(rating_county)

states_sf <- get_urbn_map("states", sf = TRUE)
counties_sf <- get_urbn_map("counties", sf = TRUE)


restrictiveness_map_data <-
  restrictiveness_data %>%
  rename(county_fips = fips_code) %>%
  left_join(counties_sf) %>%
  st_as_sf()


ggplot(restrictiveness_map_data) +
  geom_sf(aes(fill = restrictive), size = .05, color = "white")  +
  geom_sf(data = states_sf, fill = NA, size = .1, color = "white")  +
  scale_fill_binned(
    low  = "grey",
    high = "dark red",
    space = "Lab",
    n.breaks = 10,
    labels = function(x) paste0(x,"%"),

    guide = guide_coloursteps(even.steps = TRUE,
                              barheight = unit(0.1, "in"),
                              barwidth = unit(4, "in"),
```

```
                              show.limits = TRUE,
                              title.position = "top",
                              limits = c(0, 100)
    )
  ) +

  theme_void() +
  theme(
    legend.position = "bottom"
  )+
  labs(fill = "Most Restrictive Plans in Rating Area, %")



ggplot(restrictiveness_map_data) +
  geom_sf(aes(fill = less_restrictive), size = .05, color = "white")  +
  geom_sf(data = states_sf, fill = NA, size = .1, color = "white")  +
  scale_fill_binned(
    low  = "grey",
    high = "dark blue",
    space = "Lab",
    n.breaks = 10,
    labels = function(x) paste0(x,"%"),


  guide = guide_coloursteps(even.steps = TRUE,
                            barheight = unit(0.1, "in"),
                            barwidth = unit(4, "in"),
                            show.limits = TRUE,
                            title.position = "top",
                            limits = c(0, 100)
      )
  ) +

  theme_void() +
  theme(
    legend.position = "bottom"
  ) +
  labs(fill = "Moderately Restrictive Plans in Rating Area, %")

ggplot(restrictiveness_map_data) +
  geom_sf(aes(fill = not_restrictive), size = .05, color = "white")  +
  geom_sf(data = states_sf, fill = NA, size = .1, color = "white")  +
  scale_fill_binned(
    low  = "grey",
    high = "dark green",
    space = "Lab",
    n.breaks = 10,
    labels = function(x) paste0(x,"%"),

    guide = guide_coloursteps(even.steps = TRUE,
                              barheight = unit(0.1, "in"),
                              barwidth = unit(4, "in"),
                              show.limits = TRUE,
                              title.position = "top",
                              limits = c(0, 100)
    )
  ) +

  theme_void() +
  theme(
    legend.position = "bottom"
  )+
  labs(fill = "Least Restrictive Plans in Rating Area, %")


# Map Part 2 ------------------------------------------------------------
restrict_categories <-
restrictiveness_map_data %>%
  tibble() %>%
  ungroup() %>%
  mutate(
    across( contains("restrictive"), ~as.numeric(.x > 50)
            )
    ) %>%
  gather(restrictiveness, indicator, -AREA, -c(total:geometry))  %>%
  filter(indicator == 1) %>%
  mutate(
    restrictiveness = str_to_title(str_replace_all(restrictiveness, "_", " ")),
    restrictiveness = factor(restrictiveness, levels =
                               c("Restrictive", "Less Restrictive", "Not Restrictive"))
  ) %>%
  st_as_sf()



ggplot(restrict_categories) +
  geom_sf(aes(fill = restrictiveness), size = .05, color = "white", alpha = .8)  +
  geom_sf(data = states_sf, fill = NA, size = .1, color = "grey")  +
  scale_fill_manual(values = c("dark red", "dark blue", "dark green"),
                    guide = guide_legend(title.position = "top") ) +

  theme_void() +
  theme(
    legend.position = "top"
```

```
  )+
  labs(fill = "")


# rankings ------------------------------------------------------------------
restrictive_areas <-
clustered_plans %>%
  select(id, ward3) %>%
  left_join(plan_info %>%
              select(id, AREA)
  ) %>%
  group_by(AREA, ward3) %>%
  count() %>%
  spread(ward3, n, fill = 0) %>%
  rename(restrictive  = `2`, less_restrictive = `1`, not_restrictive = `3`) %>%
  mutate( total =  restrictive + less_restrictive + not_restrictive  ) %>%
  mutate(across(contains("restrictive"), ~.x/total*100))

restrictive_areas %>%
  arrange(-restrictive) %>%
  View()

restrictive_areas %>%
  arrange(-less_restrictive) %>%
  View()

restrictive_areas %>%
  arrange(not_restrictive) %>%
  View()

restrictive_areas %>%
  arrange(-not_restrictive) %>%
  View()
```