Understand the typology of HIV pre-exposure prophylaxis persistence among men having sex with men in the U.S.

Yi-No Chen Wendy Mo Wenkai Zhang

May 5, 2021

Abstract

Men having sex with men (MSM) are the population highly exploded to HIV risk in the United States (US). While pre-exposure prohylaxis (PrEP) medication for MSM at risk of HIV has been introduced and proved with high efficacy in randomized controlled trials, the effectiveness is greatly undermined due to sub-optimal adherence and persistence. This project implements two clustering methods, two-stage clustering and group-based trajectory modeling (GBTM), for the time series dataset to study the possible clusters of patterns that the patients may follow during the PrEP treatments. As the result, we find 6 possible clusters under two-stage clustering methods, and 7 possible clusters under GBTM. Both of them may help with informing the the common timings of PrEP cessations and re-initiation, and may potentially drive personalized PrEP persistence intervention.

1 Introduction

Men having sex with men (MSM) are the most affected population by HIV in the United States (US) [1]. Daily doses of Truvada (emtricitabine/tenofovir disoproxil fumarate) has been recommended as a pre-exposure prophylaxis (PrEP) medication for MSM at risk of HIV acquisition[2] due to its high efficacy in randomized controlled trials[3,4]. Suboptimal PrEP adherence (compliance to the daily dose in a prescribed interval) and persistence (maintaining PrEP use and adherence over time), which undermines its effectiveness[5], has been widely observed among MSM [6-8].

Common reasons for PrEP non-adherence and cessation, which includes lack of socio-economic support, low risk perception and PrEP awareness, side effects and other life style factors, has been described in existing literature[9]. However, knowledge of the root barriers to PrEP persistence without quantitative data regarding the magnitude and patterns of cessation does not allow for PrEP care providers to identify patient-specific intervention timing. As the result, the goal of this analysis is to cluster the common long-term patterns of the PrEP persistence among adult male PrEP users in the US. By using both non-parametric and parametric methods, we are hoping that the mined cluster structure would inform the common timings and durations of PrEP cessations and re-initiations. Such findings could potentially inform the development of personalized PrEP persistence interventions that target these important timings

2 Methods

Data Source

The Truvada prescription data of male PrEP users from Symphony Health (national PrEP pharmacy database of 80% of users in the US, 2012-2020) will be used. We included 160,739 male registered users who were aged between 18 and 65 years upon PrEP initiation, and had at least two years of

follow-up period since the initiation date. PrEP initiation date is defined as the earliest time when an user was prescribed with Truvada for at least 14 days. The raw dataset contains following fields: 1) Patient ID; 2) Start date of a prescription interval; 3) End date of a prescription interval. The raw data is formatted such that each row represent a prescription interval with start and end date, and a patient can have multiple prescription intervals (rows).

Data Pre-Processing

For stage-II clustering analysis, we constructed a weekly time series vector of proportions of days in the two weeks covered in PrEP medication (PDC) for each user. Each PDC time series would start on the date of the user's PrEP initiation and end at the two-year mark from his initiation date. This is result in a 160,739-by-103 data matrix . We particularly choose the two-year interval because PrEP users, on average, are retained in care for about 14 months. Since the randomized control studies have shown that a PDC below 57% (or < 4 pills per week) is associated with suboptimal efficacy against HIV infections, we treat PDC of 57% as the marker of PreP cessation in this study.

For the group-based trajectory modeling (GBTM) analysis, we constructed a biweekly time series vector of PDC instead for each user (160,739 by 52 matrix) in order to reduce the complexity of model estimation and the outcome correlation between time points (note: the GBTM assumes conditional independence in the sequential realization of the modeled outcome).

Two-Stage Clustering

The main purpose of the two-stage clustering strategy is to reduce the computational complexity of clustering PDC data by first breaking our large data set to a few smaller subsets. For the first stage, we partitioned the entire data set using the k-means algorithm (metric: euclidean distance) on the following three features: 1) the total duration of PrEP cessation; 2) the number of PrEP interval; and 3) the timing of the first PrEP cessation. All input data were standardized. We assessed cluster quality under various k values (k=2 $\sim$ 4) by examining the following indices: silhouette index, Davies-Bouldin index and Calinski-Harabasz index. Majority vote will be used to determine the optimal k value.

For the second stage, we further detect clusters of PDC time series within each of the stage-I clusters, using hierarchical agglomerative clustering algorithm under average linkage, in which the similarity is defined in dynamic time warping distance (DTW). DTW is selected because it provides flexible alignment against time shift[10]. In the mean time, we set a global Sakoe-Chibia band constraint [10] on the curve alignment to ensure that the final algorithm will deduce distinct cluster membership to PrEP users that have similar PDC curve shapes but very different timings of PrEP cessations. The band size is set at 10 weeks ($\sim$10% of the time series length) to be consistent with the general recommendation [11]. Sum of square error and silhouette index were evaluated to determine the optimal k-group solution. To assess whether the detected cluster structure is sensitive to the choice of clustering algorithm, we additionally implemented partition-around-medoids (PAM) algorithm under various k values, in lieu of the hierarchical clustering, for stage II as a sensitivity analysis. The aforementioned algorithms are chosen because they are the two most widely used methods for clustering time-series data, and may provide different clustering perspectives (graph-based vs partition).

Group-based Trajectory Modeling

GBTM, a specialized application of finite mixture models, is a statistical methodology for analyz-

ing developmental trajectories. It is particularly good for gathering individuals into meaningful subgroups by similar trajectories[12]. It identifies groups under different polynomial functions of time by maximum likelihood estimation. As we denotes i as the unit of analysis, t as time, and j as group, the general form data is: $Y_i = y_{i1}, y_{i2}, ..., y_{iT}$. The ultimate goal is to find a set of parameters, ω, that maximizes the probability of observing Y_i. These parameters also define the shape of the trajectories by a polynomial of time and the probability of group membership[13]. The construction of the likelihood function is the aggregation of the J conditional likelihood functions $P^j(Y_i)$ to form the unconditional probability: $P(Y_i) = \sum_j^J \pi_j P^j(Y_i)$.

In this analysis, we will assume that the PDC trajectory (Y_i) given membership in group j follows a censored normal distribution (max=1, min=0), where $\mu_{itj} = \beta_{0j} + \beta_{1j} Time_{it} + \beta_{2j} Time_{it}^2 + \beta_{3j} Time_{it}^3$. Akaike information criterion (AIC) and the Bayesian information criterion (BIC), the two commonly used model fitness statistics, are evaluated at various k values to identify the optimal k-group solution. Given the calculation formula used in the SAS procedure (PROC TRAJ), the less negative values of AIC and BIC indicate better model fit.

3 Results

Preliminary Data Exploration

Majority of the adult male PrEP users (63.99%) had only one PrEP prescription interval within the two-year follow-up period. The number of PrEP intervals also decreases exponentially (**Supp. table 1**). For the middle 50% of the PrEP users, their first PrEP cessation occurred between 4 and 48 weeks (median = 19 weeks) after the PrEP initiation; they overall spent between 36 and 93 weeks (median = 67 weeks) off PrEP during the two-year follow-up period. The distributions shown in **Supp. figure 1** and**2** suggest that it is common for the adult male PrEP users to discontinue PrEP within the first 5 months into the prophylaxis treatment, and not re-initiate afterward.

Stage-I Clustering: k-means

The global average of Silhouette index and Davies-Bouldin index suggest that the hyperparameter of k=4 results in better balance of cohesion within clusters and separation between clusters, than the other values of k. On the other hand, Calinski-Harabasz index indicates that k=3 may lead to slightly better cluster structure than k=4 (**Table 1**). However, since the improvement of the Calinski-Harabasz index is not significant, we concluded that k=4 is the optimal cluster solution for the stage-I k-mean partition. We use UMAP, which reduces the feature space to 2-dimensions while preserving global data structure, to further visualize data distribution under the optimal cluster solution (**Figure 1**). The cluster solution was able to generate clear cluster structure for Cluster C, and within-cluster cohesion for Cluster B. However, the denseness of Cluster A was rather uneven across the reduced dimensional space.

Table 1: Stage-I cluster quality indices by k-group solution

K	Silhouette Index	Davies-Bouldin Index	Calinski-Harabasz Index
2	0.429	1.023	121,346
3	0.488	0.818	164,221
4	0.490	0.719	163,824

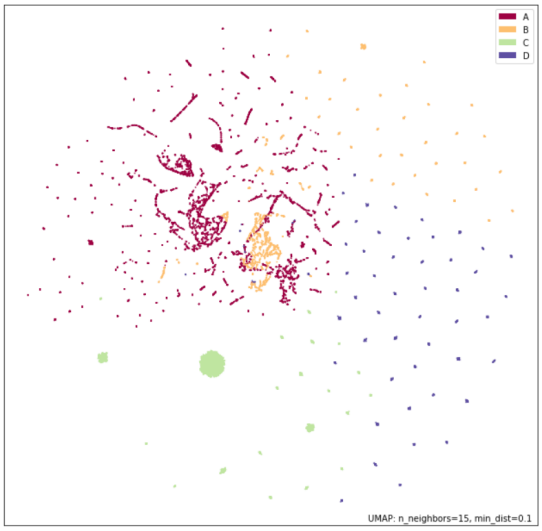

Figure 1: Data distribution by stage-I cluster membership in UMAP (sample size=25%)

Stage-II Clustering: hierarchical Clustering

Sample size selection

Due to limited memory, we used a subset of each stage-I cluster's data for the stage-II clustering. To determine the minimum sufficient sample size for each of the stage-I clusters, we computed the trend of silhouette index across various k values for each of the following sample size: 1%, 5%, 10% and 15%. We expected the change in the sample-specific silhouette trend to decrease as the sample size increases. **Figure 2** shows that the silhouette index trends converge once the sample size reaches at least 5% for stage-I cluster A and B. On the contrary, for stage-I cluster C and D, the silhouette index trends vary meaningfully from each other at all sample sizes, in which case we treated the sample size of 15% as the representative standard as it is the largest sample size our local machine can process. As the result, we determined that the following sample sizes are minimally sufficient for representing stage-I clusters: 5% (cluster A & B), 15% (cluster C & D).

Hyperparameter tuning: number of detected clusters

After finalizing the sufficient sample sizes, we evaluate the SSE across the number of clusters and the distribution of individual silhouette indices by cluster membership to determine the optimal number of clusters to generate. In **Figure 3,** the elbow method suggests that the optimal number of clusters could be within the following ranges: k=9 $\sim$ 11 for stage-I cluster A; k= 7 $\sim$ 9 for stage-I cluster B; k= 6 $\sim$ 7 for stage-I cluster C; k=8 $\sim$ 9 for stage-I cluster D.By examining the distributions of silhouette indices by stage-II cluster membership (**Supp. figure 4-7**), we determine that 11, 8, 6 and 8 are the optimal numbers of stage-II clusters for stage-I cluster A, B, C and D respectively. This is because these cluster solutions result in the most even distributions of silhouette across different stage-II cluster membership in their respective stage-I cluster samples.

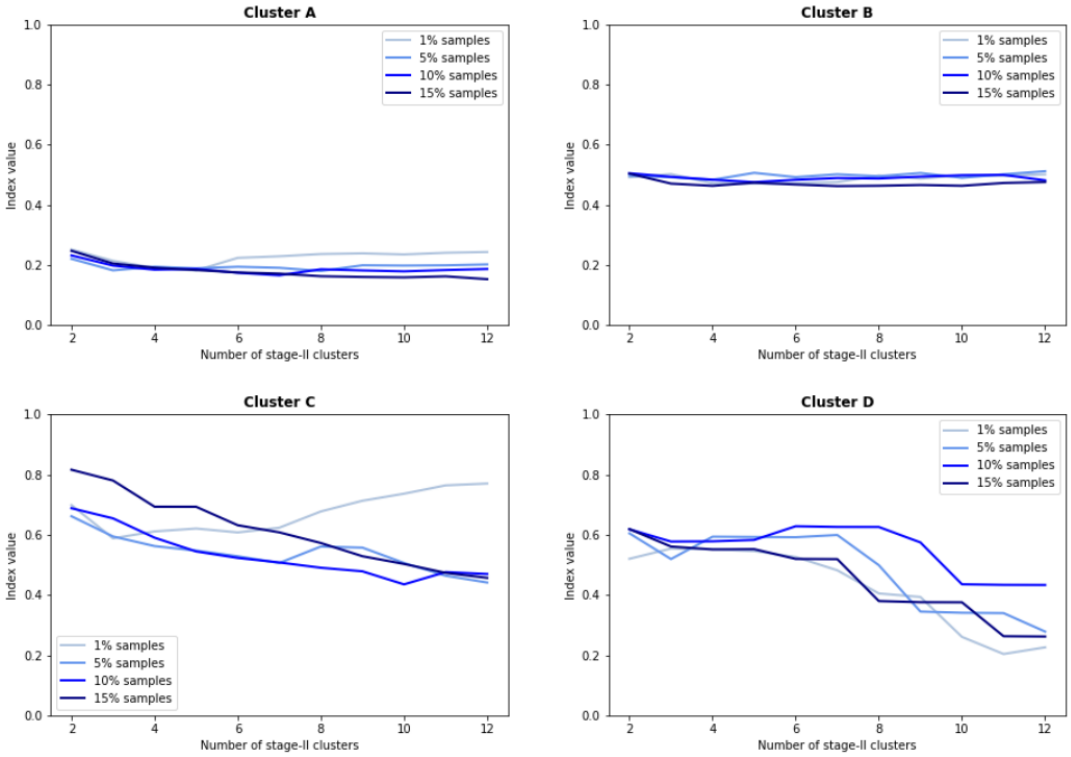

Figure 2: Average silhouette indices by number of stage-II clusters, in stage-I-clustered sample datasets

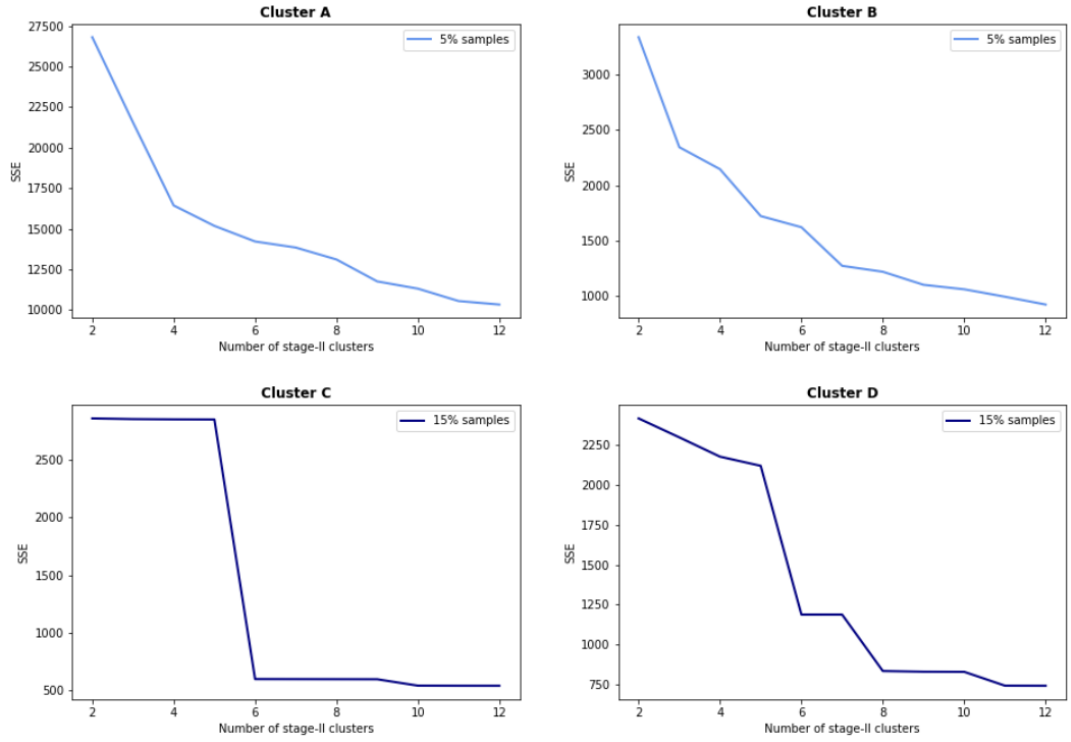

Figure 3: Sum of square error by number of stage-II clusters

Extraction of common longitudinal PDC patterns

In total, the 2-stage clustering approach generates 33 clusters, which may be trivial for coherent interpretations. Many clusters (e.g., Cluster C.4 – C.6) may be considered as outlying patterns due to low occurrence frequency (**Supp. figure 12**). Therefore, we extract the DTW barycenter averaging (DBA) centroid, or average time series through which the squared distance is minimized, from each of these clusters as their prototypes (i.e., representative feature of a cluster). Then, we use hierarchical clustering method (with average link) to merge the similar centroids, defined by DTW distance. We determined that k=6 is the optimal cluster solution via the elbow method on SSE (**Supp. figure 13**). For each of merged cluster of centroids, a new DBA centroid weighted by the relative frequencies of the member centroids in the cluster is computed (**Figure 4**).

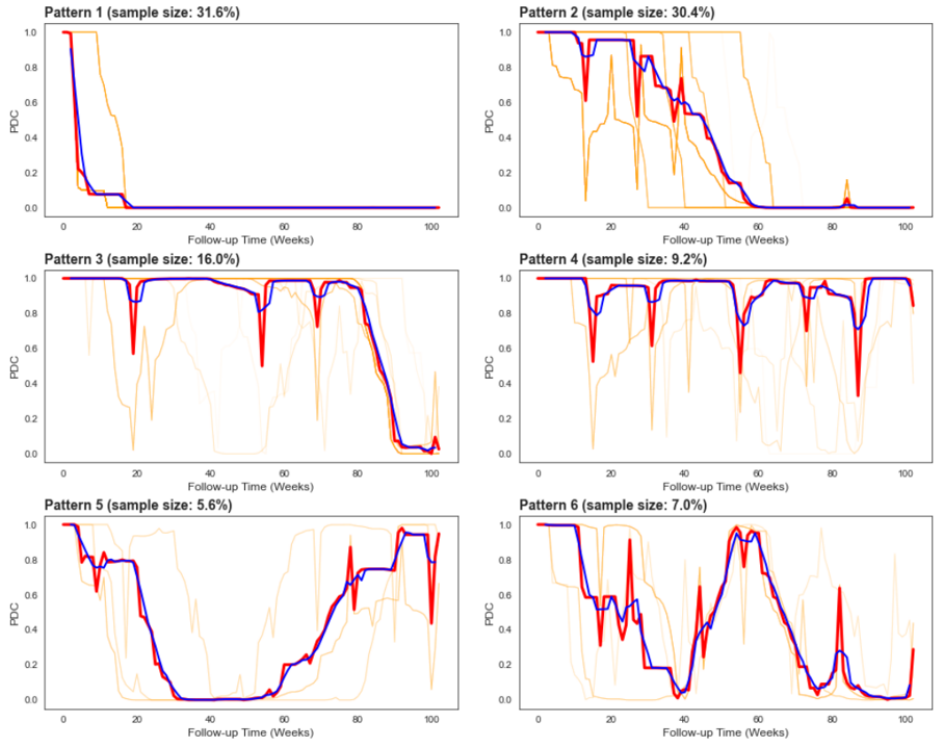

Figure 4: Distribution of extracted DBA centroids and final prototype patterns.

Sensitivity analysis: PAM for stage-II clustering

We used the same strategies discussed in the previous sections to determine the sufficient sample sizes (cluster A: 5%; cluster B: 10%; cluster C: 5%; cluster D: 10 %) and optimal number of clusters (cluster A: 11; cluster B: 8; cluster C: 7; cluster D: 4). The intermediate results are included in supp. figures 14 - 23. We then extracted the DBA centroid from each of the 30 detected clusters, and merged them using hierarchical agglomerative clustering algorithm to obtain the interpretable number of pattern clusters. The elbow method on the SSE suggests the optimal k=6. Pattern 1 - 5 are similar to those detected from the stage-II hierarchical clustering analysis (**Figure 5**).

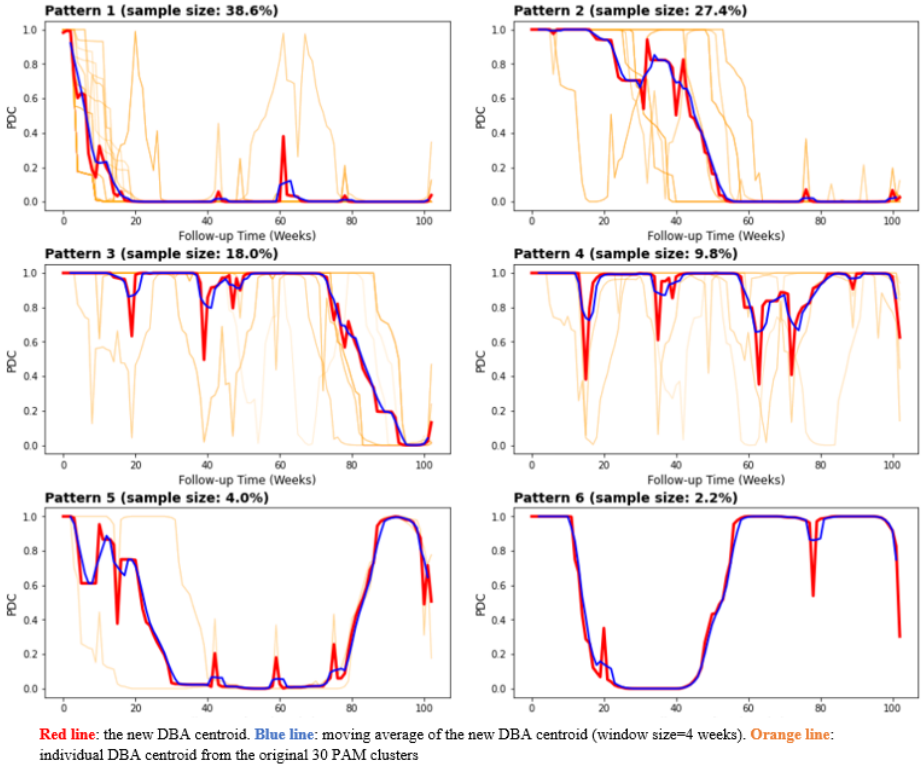

Figure 5: Distribution of extracted DBA centroids and final prototype patterns.

Group-based trajectory modeling

AIC and BIC both suggest that k=7 and 8 are the two best fitted hyperparameter values among the examined k-group solutions. Based on the parsimonious modeling principle, we selected the 7-group solution. To visualize the PDC trajectories in the same scale as the pattern prototypes computed in the two-stage clustering analysis, we converted the biweekly time series of predicted and observed mean PDC back to the weekly time series format (**Figure 6**). Moving average smoothing is applied to refine the underlying logic of the detected trajectory. The congruence between the trajectory of model predicted mean trajectory and that of the observed group mean (weighted by the posterior probabilities of cluster membership) also suggest good model fit. Figure 6 also shows that majority of the PrEP users are one-time users, and the common timings of PrEP cessation include .

Table 2: Model fitness statistics by k-group solution

K	BIC	AIC	False convergence
2	-1,490,957	-1,490,917	Yes
3	-541,238	-541,178	No
4	-1,215,666	-1,215,585	Yes
5	-471,022	-470,921	No
6	-471,047	-470,926	No
7	-437,626	-437,484	No
8	-435,879	-435,717	No
9	-409,027	-930,408	Yes

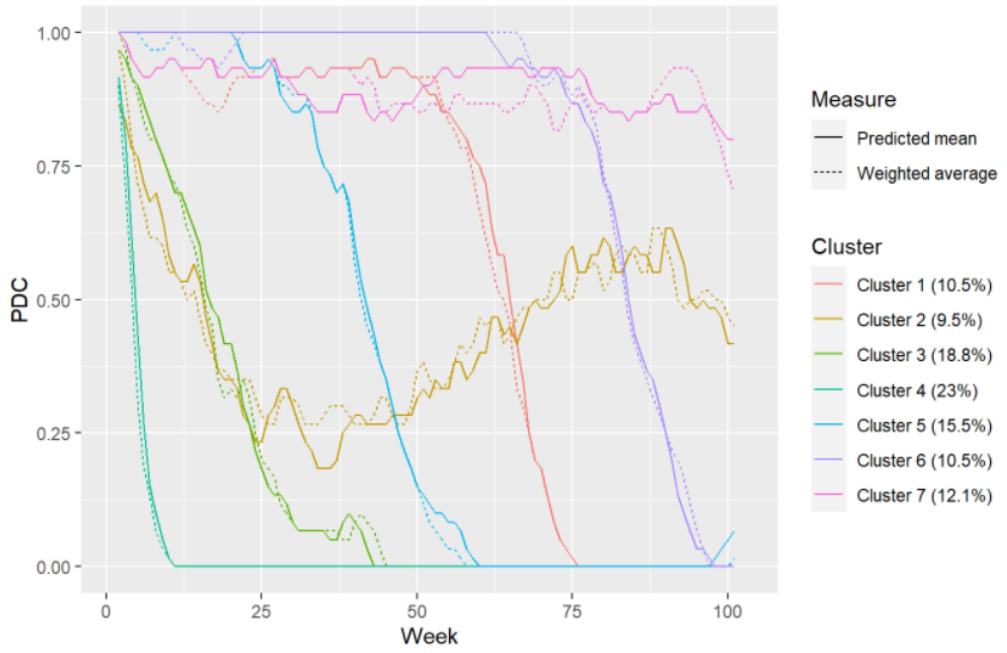

Figure 6: Weighted average vs. predicted mean PDC over time, censored normal mixture model (moving average smoothing, with window size of 4 weeks).

4 Discussion

In this clustering exercise, we adopted methods from both the non-parametric (e.g., hierarchical clustering, PAM) and statistical modeling (i.e., GBTM) traditions to discover the longitudinal patterns of PrEP PDC. We found that the discovered structure of average PDC trend pattern (or DBA) is robust against different algorithms. Pattern 1-5 discovered in the main analysis (Figure 4) corresponds well with the pattern 1-5 in the sensitivity analysis (Figure 5). These shared patterns account for over 90% of the male PrEP users. However, the within-cluster variance of average PDC time series appears greater in the sensitivity analysis. This is perhaps because the prototypes used in PAM (i.e., medoid) and the final cluster merging process (i.e., DBA) are not consistent.

Overall, over half of the adult male PrEP users follow a steadily decreasing PDC patterns with no re-initiation (Pattern 1 and Pattern 2 in Figure 4 and 5). Among those, the common cessation occurred at around week 10 or week 50. About 9% of the adult male PrEP users follow the pattern of a relatively stable PrEP users with occasional short-term drop off from medication (Pattern 4). About 6 – 12% had an average PDC time series characterized by long-term breaks ($>$ 20 weeks) before re-initiating on PrEP again. On the other hand, GBTM discovered a cluster structure for the population-level trends similar to that for the average PDC trend patterns detected by our non-parametric methods. However, the model estimated lower proportions of users belonging to the group characterized by PrEP discontinuation about 10 weeks into the treatment. GBTM was also able to cluster more subgroups among the one-time PrEP users, which helps reveal more common timings of PrEP cessations. Compared to the other methods that use DBA as the prototypes, one limitation with GBTM is that the discovered trajectory only speaks about population average over time, which may not shed light on the longitudinal trend of an average individual in a cluster.

In conclusion, the similar findings across all our methods suggest that the discovered structure of PrEP persistence pattern is robust and likely to exist beyond our sample data (i.e., contains true structural information as opposed to the result of random noises).

5 Reference

[1] Diagnoses of HIV Infection in the United States and Dependent Areas, 2018 (Updated), in HIV Surveillance Report. 2018, U.S. Centers for Disease Control and Prevention: Atlanta, GA.

[2] Preexposure prophylaxis for the prevention of HIV infection in the United States—2017 update: a clinical practice guideline. 2018, U.S. Public Health Service: Atlanta, GA.

[3] Grant, R.M., et al., Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. New England Journal of Medicine, 2010. 363(27): p. 2587-2599.

[4] Molina, J.-M., et al., On-demand preexposure prophylaxis in men at high risk for HIV-1 infection. N Engl J Med, 2015. 373: p. 2237-2246.

[5] Grant, R.M., et al., Uptake of pre-exposure prophylaxis, sexual practices, and HIV incidence in men and transgender women who have sex with men: a cohort study. The Lancet. Infectious diseases, 2014. 14(9): p. 820-829. [6] Amico, K.R., et al., Study product adherence measurement in the iPrEx placebo-controlled trial: concordance with drug detection. Journal of acquired immune deficiency syndromes (1999), 2014. 66(5): p. 530-537.

[7] Chan, P.A., et al., Retention in care outcomes for HIV pre-exposure prophylaxis implementation programmes among men who have sex with men in three US cities. J Int AIDS Soc, 2016. 19(1): p. 20903.

[8] Montgomery, M.C., et al., Adherence to Pre-Exposure Prophylaxis for HIV Prevention in a Clinical Setting. PLoS One, 2016. 11(6): p. e0157742.

[9] Sidebottom, D., A.M. Ekström, and S. Strömdahl, A systematic review of adherence to oral pre-exposure prophylaxis for HIV - how can we improve uptake and adherence? BMC infectious diseases, 2018. 18(1): p. 581-581.

[10] Cassisi, C., et al., Similarity measures and dimensionality reduction techniques for time series data mining. Advances in data mining knowledge discovery and applications'(InTech, Rijeka, Croatia, 2012, 2012: p. 71-96

[11] Sardá-Espinosa, A., Comparing time-series clustering algorithms in r using the dtwclust package. R package vignette, 2017. 12: p. 41.

[12] Nagin. Group-based trajectory modeling: An overview. 65, 2014. doi: 10.1159/000360229. URLhttps://www.karger.com/DOI/10.1159/000360229.

[13] Gary Sweeten.Group-Based Trajectory Models, pages 1991–2003. Springer New York, NewYork, NY, 2014. ISBN 978-1-4614-5690-2. doi: 10.1007/978-1-4614-5690-2479. URLhttps://doi.org/10.1007/978-1-4614-5690-2$_4$79.