# A Method for Clustering Flexible Longitudinal Trajectories with An Application to An HIV Prevention Trial

Jingwen Zhou

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2020

Committee:

Elizabeth Brown

James Hughes

Program Authorized to Offer Degree:

Biostatistics – Public Health

University of Washington

## Abstract

A Method for Clustering Flexible Longitudinal Trajectories with An Application to An HIV Prevention Trial

Jingwen Zhou

Chair of the Supervisory Committee:

Elizabeth Brown

Department of Biostatistics

Dapivirine (25 mg) in a silicone elastomer Vaginal Ring is a safe and effective method to prevent HIV-1 infection in healthy, sexually active, HIV-negative women. Understanding women's adherence to the dapivirine vaginal rings over time and exploring factors that might be associated with different patterns of adherence are helpful for supporting women who use the rings and achieving better levels of protection against HIV-1. In this analysis, we propose a trajectory clustering method that performs K-means clustering with generalized additive models implemented to produce cluster means. We identified six different adherence trajectories, using data from 1,356 HIV-uninfected women participated in the HIV open-label prevention extension (MTN-025/HOPE) trial. While most of the participants were clustered as high adherence (18% as *high adherence 1*; 18% as *high adherence 2*) and *medium adherence* (24%), we also identified

declining adherence trajectories (15% as *high – declining adherence*; 13% as *declining adherence*) and a *non-adherence* trajectory (12%). Baseline characteristics including non-oral contraceptive use and lack of menstrual bleeding in the past 3 months were found to be associated with being clustered as high or medium adherence. Being clustered as declining adherence was found to be associated with having experienced menstrual bleeding in the past 3 months and non-use of injectable contraceptive. Among participants who showed high adherence, contraceptive choice of intrauterine device was found to be correlated with being clustered as *high adherence 2.*

# Introduction

Women continue to bear the brunt of the HIV pandemic, accounting for more than half the number of people living with HIV worldwide [1]. In sub-Saharan Africa, four in five new infections among adolescents aged 15 – 19 years are in girls. Young women aged 15 – 24 years are twice as likely to be living with HIV compared to their male counterparts [2].

As of end of June 2019, 24.5 million people living with HIV were accessing antiretroviral therapy [2]. A recent trial named MTN-020/ASPIRE [3] within the Microbicide Trials Network found that dapivirine (25 mg) in a silicone elastomer Vaginal Ring (VR) is a safe and effective to prevent of HIV-1 infection in healthy, sexually active, HIV-negative women and that high levels of protection against HIV-1 can be achieved with regular and consistent use of the rings. Motivated by the MTN-020/ASPIRE trial, the HIV open-label prevention extension (MTN-025/HOPE) trial [4] was designed to assess the safety of and participant adherence to dapivirine VRs. In the MTN-025/HOPE trial, participants received a silicone elastomer VR containing 25 mg of dapivirine, to be replaced monthly, for a total period of 12 months of use. Participant adherence was determined by the amount of dapivirine released, which was calculated as the difference between the dapivirine concentration measured in the returned, used ring and the corresponding baseline level. For each participant, a sequence of monthly adherence measurements was obtained across the 12 months of dapivirine VR use, which forms an individual adherence trajectory. Understanding participants' adherence to the dapivirine VR over time and identifying factors that influence adherence patterns over time can help target the intervention to women who are most likely to benefit from it and support women who use the dapivirine VR in the future.

One way to understand how participants used the dapivirine VR is to identify subgroups (clusters) of participants with similar patterns of ring use through adherence trajectory clustering. For example, one subgroup of participants may exhibit consistent high adherence across the 12 months, while another subgroup may begin with high adherence but shift to low adherence after certain time points. We aim to identify these patterns and the subgroups of women who form them.

## Overview of Trajectory Clustering

Methods available for clustering trajectories can be separated into two families. The first family contains model-based methods which assume that the data were generated by a certain statistical model and attempt to recover the original model from the data. The recovered model then constructs clusters and assigns cluster memberships. Examples of model-based methods are latent class analysis or mixture modeling techniques. The second family comprises non-parametric methods that usually involve an extension of the classic K-means clustering algorithm. The pros and cons of both methods have been extensively studied and discussed [5 - 6]. The model-based methods are more generally preferred because of the potential to check clustering validity with formal tests, and the flexibility of including demographic or other exogenous variables as covariates. In contrast, non-parametric methods have some potential advantages, for example, no additional parametric assumptions within clusters are involved, and they do not require assumptions regarding the shape of the trajectory.

## Review of non-parametric methods

Several approaches have been developed for the extended use of K-means algorithm with longitudinal data. We will discuss three methods here which provide variant K-means-based clustering algorithms for clustering longitudinal data and have realizations on the R platform. They may be used in conjunction with different clustering outcomes of interest. In this analysis,

for example, participant adherence as outcome is determined by the continuous variable, the amount of dapivirine released.

The KmL package [5] implements a simple extension of the K-means algorithm to longitudinal data with adapted distance metrics. Clustering using the Euclidean distance and the Manhattan distance are available in the package. Let $Y$ denote the outcome of interest, participant adherence. Observations from the individual adherence trajectory for participant $i$ can be written as $Y_i = (y_{i1}, \dots, y_{it})$, where $y_{it}$ denotes the measurement at time $t$ $(t = 0, \dots, T)$.

The Euclidean distance between two trajectories $Y_i, Y_j$ is defined as $Dist(Y_i, Y_j) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_{it} - y_{jt})^2}$. The Manhattan distance between two trajectories $Y_i, Y_j$ is defined as $Dist_M(Y_i, Y_j) = \frac{1}{T}\sum_{t=1}^{T}|y_{it} - y_{jt}|$.

To address the issue of choosing the optimal cluster number $K$, the KmL package uses the Calinski and Harabasz criterion, $C(K)$, aiming at maximizing the between-cluster variance and minimizing the within-cluster variance. Let $n_k$ be the number of trajectories in cluster $k, k = 1, \dots, K$, $\overline{Y_k}$ be the mean trajectory of cluster $k$. Let $\bar{Y}$ be the mean trajectory of the whole data set. $v'$ denotes the transposition of vector $v'$. The between-variance matrix $B$ is defined as

$$B = \sum_{k=1}^{K} n_k(\overline{Y_k} - \bar{Y})(\overline{Y_k} - \bar{Y})'$$

the within-cluster variance matrix $W$ is defined as

$$W = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{Y}_{ki} - \overline{Y_k})(\boldsymbol{Y}_{ki} - \overline{Y_k})'$$

$C(K)$ is defined as the ratio of the trace of the between-variance matrix and the trace of the within-variance matrix. To deal with the missingness which could be caused by the difference in

time points of adherence measurements for two individual trajectories in the clustering process, the Gower adjustment is applied by calculating a weighted distance. The weight $w_{ijt}$ is 1 if both $Y_{it}$ and $Y_{jt}$ are available at time $t$, and 0 otherwise. For example, the Euclidean distance between two trajectories $Y_i, Y_j$ with Gower adjustment is given as

$$Dist_{Gower}(Y_i, Y_j) = \sqrt{\frac{1}{\sum w_{ijt}} \sum_{t=1}^{T} (y_{it} - y_{jt})^2 \cdot w_{ijt}}$$

This simple extension of the K-means algorithm shares the advantages of non-parametric methods with potentially higher efficiency compared with an existing mixture modeling procedure. However, several limitations exist for this method. Like other non-parametric approaches, a formal statistical test for the validity for the clustering structures is not applicable. In the KmL package, the built-in approaches to deal with missingness are not effective under certain circumstances where no distance measurement is possible without further assumptions. An example would be that, when calculating the distance between two trajectories where one of them has missing measurement only at time $t'$ while the other has only one measurement at time $t'$, the Gower adjustment ends up being zero which is not informative.

For trajectories with outcomes that are binary or count data, an updated algorithm from the KmL package is available in the R package kmlCov [7]. Because of the dependence between intra-group variance and mean for binary and counts data, deviance distances instead of classic distance metrics are implemented in this approach.

When describing the trajectory patterns in subgroups, one might be interested in identifying individual trajectories that follow the same trends. That is, the progress of a phenomenon is more important than the moment at which it occurs. The kmlShape package [8] implements a K-means

algorithm that uses the generalized Frechet distance. The Frechet distance is a shape-respecting distance measure that considers the flow of the two trajectories, both in vertical (i.e., clustering variable) and horizontal (i.e., time) directions. Let $P$ and $Q$ be two curves that are both continuous mapping from the time interval $[0, T]$ to $\mathbb{R}$. Let $\mathcal{A}$ denote all reparameterizations of the interval $t \in [0, T]$ and $\alpha$ and $\beta$ be two reparameterizations in $\mathcal{A}$. The Frechet distance between the curves $P, Q$ is defined as

$$Dist_F(P, Q) = \min_{\alpha,\beta} \max_{t\in[0,T]} d\left(P\left(\alpha(t)\right), Q\left(\beta(t)\right)\right)$$

Where $d$ is the Euclidean distance.

The generalized Frechet distance adds a time-scale parameter $\lambda$ to deal with the issue that the variable of interest and the time variable are not measured using the same units. The generalized Frechet distance between curves $P, Q$ is the Frechet distance obtained after a transformation in time variable, $A: t \to \lambda t$.

$$Dist_{gF} = \min_{\alpha,\beta} \max_{t\in[0,T]} d\left(P\left(\alpha\left(A(t)\right)\right), Q\left(\beta\left(A(t)\right)\right)\right)$$

The corresponding mean trajectories based on the generalized Frechet mean are calculated in the classification step. By using generalized Frechet distance, the K-means algorithm can easily deal with the issue of individuals having different number of repeated measurements at different sets of time points.

The kmlShape package tackles the complexity in the big data setting by allowing data size reduction from two perspective. One is to reduce the number of individuals by first identifying representative trajectories using classical clustering algorithm. The other one is to reduce the number of measurements without much loss of information.

In this analysis of adherence trajectory clustering, the moment at which the pattern of ring use changes is of interest. Thus, the approach implemented in the kmlShape package might not be the best choice.

## Review of model-based methods

There are several model-based methods available for studying the developmental trajectories. Nagin [9] proposed a longitudinal analytic method called group-based trajectory modeling (GBTM). The method has been widely applied in varied science fields including physical aggression [10], criminal behavior [11], cortisol level [12] and HIV [13 - 14]. GBTM is an application of the finite mixture models which assumes that the population is composes of a mixture of distinct groups of individuals following similar trajectories of a single outcome over time [15]. GBTM assumes that the population distribution of the outcome conditioning on age (or time) rises from a finite mixture of order $K$, where $K$ is the number of groups set apriori. Under the assumption of GBTM, population variability is considered to be fully captured by differences across groups in the shape and level of their trajectories. Thus, for a given subgroup $k$, conditional independence is assumed for the repeated measurements at time points $t = 1, \dots, T$. The model parameters are estimated through maximum likelihood estimation. The likelihood for individual $i$ can be written as

$$P(Y_i|age_i, K, \beta_1, \dots, \beta_k, \pi_1, \dots, \pi_k) = \sum_{k=1}^{K} \pi_k \cdot \prod_{t=1}^{T} p(Y_{it}|age_{it}, k; \beta_k)$$

Where $\pi_k$ is the unknown parameter stands for the probability of membership in group $k$ and $p(.)$ is the conditional distribution of the outcome given the group membership $k, k = 1, \dots, K$ which is indexed by unknown parameter vector $\beta_k$. Two key outputs from GBTM are the shape

of the group trajectories indexed by the estimated parameters, and the posterior probability of trajectory group membership which determines the clustering.

Another commonly used model-based method for trajectory clustering is the latent class mixed modeling (LCMM). LCMM assumes that the population is heterogeneous consisting $K$ latent classes of trajectories. LCMM is different from GBTM in the way that the within-group variability is modeled [16]. While GBTM fixes within-group variability of individuals to zero, LCMM allows individual variability within groups by including random effects. Thus, LCMM provides additional information about how closely individual trajectories resemble the mean trajectory for a subgroup. A prior probability $\pi_{ik}$ of latent class membership is defined by a logistic model based on a set of covariates $X_{c_i}$, where $c_i = k$ if the *i*th individual belongs to the *k*th latent class. The latent class membership model can be written as

$$\pi_{ik} = \frac{\exp\left(\xi_{0k} + X_{c_i}\xi_{1k}\right)}{\sum_{l=1}^{K}\exp(\xi_{0l} + X_{c_i}\xi_{1l})}$$

The trajectory outcome $Y_i$ conditioning on the latent class membership is described by a linear mixed model based on time and covariates associated fixed or random effects. Conditional on class $k$, the model is defined for individual $i$ at time $j$ as:

$$Y_{ij}|_{c_i=k} = X_{2ij}\beta + X_{3ij}\gamma_k + Z_{ij}b_{ik} + \epsilon_{ij}$$

Here $X_{2ij}$ is a vector of covariates associated with common fixed effect $\beta$ across $K$ latent classes. $X_{3ij}$ is a vector of covariates associated with class-specific fixed effects $\gamma_k$. $Z_{ij}$ is a vector of covariates associated with individual random effects $b_i|_{\mathrm{c_i=k}}$ called $b_{ik}$ whose distribution is class-specific. $\epsilon_{ij}$ is a vector of random variables for individual measurement errors. The parameter vectors are estimated using maximum likelihood estimation method. Posterior

clustering for each individual is then defined as the latent class $k$ that generates the largest posterior class membership probability.

The R package lcmm [17 - 18] is available to perform the LCMM for continuous trajectory outcome. Posterior clustering of each individual trajectory as well as estimated mean trajectories for each subgroup can be obtained and visualized. Prediction of cluster membership for a new dataset with a specified profile of covariates is also available. Some built-in methods are available for model evaluation, including the selection of the number of latent classes.

This thesis proposes a trajectory clustering method (KGAM) that can be viewed as a variant of K-means algorithm that utilizes a generalized additive model (GAM) when constructing the cluster means. KGAM keeps the simplicity of the non-parametric approach and incorporates the model-based approach with an attempt to better address the issues of data missingness caused by the difference in time points of adherence measurements for different participants. Details about this method are introduced in the following sections. The method was applied to the MTN-025/HOPE trial to identify adherence pattern of dapivirine VR use. Additionally, associations between baseline characteristics and the cluster membership assignments were also explored.

# Methods

## Study population

This thesis is motivated by the HIV Open-label Prevention Extension (MTN-025/HOPE) trial, a multi-site, open-label, randomized, Phase 3B trial. The MTN-025/HOPE trial aimed at further assessing the safety of and participant adherence to dapivirine (25mg) in a silicone elastomer VR as a prevention of HIV-1 infection in healthy, sexually active, HIV-negative women. A total of 1,572 eligible HIV-uninfected MTN-020/ASPIRE participants were contacted, among which 1,456 women agreed to continue their participation in the HOPE trial and were enrolled at 14

sites in Malawi, South Africa, Uganda, and Zimbabwe. After enrollment, participants received a silicone elastomer VR containing 25 mg of dapivirine, to be replaced monthly, for a total period of 12 months of use. The ring(s) were returned in the next follow-up visits for adherence measurement defined as the amount of dapivirine released. Follow-up visits occurred monthly for the first 3 months, and quarterly thereafter for a total period of 12 months. For each participant, a sequence of monthly adherence measurements was obtained across the 12-month follow-up, which forms an individual adherence trajectory. Besides adherence data, participant characteristics including demographic variables, contraceptive choice, and sexual behaviors were collected at study enrollment.

## A K-means variant Using Generalized Additive Model (KGAM).

### Introduction to K-means

K-means is an unsupervised clustering algorithm which belongs to the family of Expectation-Maximization (EM) algorithms [19]. Specifically, two phases operate in turn to reach the optimal clustering: during the *Expectation* phase, the cluster means are calculated; then in the *Maximization* phase participants are assigned to the closest cluster. The alternation of the two phases is repeated until no further changes occur in the clusters or a pre-specified maximum number of iterations is reached. The distance between participant $i, i = 1, \dots, n$ and the cluster mean is defined as the sum of squared difference between observations from the adherence trajectory , $Y_i = (y_{i0}, \dots, y_{i,11}), i = 1, \dots, n$ and the cluster mean, $\widehat{Y_k} = (\widehat{y_{k0}}, \dots, \widehat{y_{k,11}}), k = 1, \dots, K$ at each time point $t, t = 0, \dots, 11$ .

$$Dist_k = \sum_t (y_{it} - \widehat{y_{kt}})^2$$

For each participant, the cluster with the minimum value of the sum of squared differences over the $K$ clusters is considered to be the closest cluster.

**Introduction to GAM**

A generalized additive model (GAM) is a generalized linear model (GLM) in which part of the linear predictors are represented by the sum of smoothing functions of the predictor variables. In this analysis, time in study (month) is the predictor; participant adherence measured by the amount of dapivirine released is the outcome. Consider a GLM which describes the relationship between the exponential family distributed response variable $Y$, with mean $\mu$, and the predictors via the model

$$g(\mu) = X^*\beta^* + \sum_j L_j f_j$$

Here $g$ is a known monotonic link function. $X^*$ represents the data matrix of the parametric model components, if any, associated with the coefficients $\beta^*$. Specifically, the $X^*$ term includes baseline characteristics of interest in this analysis. The $f_j$ are unknown smoothing functions of predictors, and the $L_j$ are known linear functions that depend on the predictors. In a GAM model, the $L_j f_j$ term can be written as $f_j(t)$. Overly wiggly smoothing functions might be introduced to ensure low bias when fitting the data. To avoid overfitting, models are fitted by the penalized least squares method. In a penalized least squares estimate, the squared residual term is modified by the addition of a penalty for each smooth function, penalizing for its "wiggliness". To balance between penalizing wiggliness and penalizing badness of fit, each penalty is multiplied by an associated smoothing parameter. In particular, the model is estimated by finding the smoothing function which minimizes

$$\left|\left|y - f\right|\right|^2 + \lambda \int f''(x)^2 dx$$

Where $\lambda$ are positive smoothing parameters; the second term measures the wiggliness of the smoothing function. Note that larger $\lambda$ leads to more smoothness.

This approach fits GAM models when constructing the cluster means during the K-means process. By fitting a GAM model, we can obtain predicted values of participant adherence depending on time in study (month) as predictor, as well as other baseline characteristics as covariates of interest. The predicted values of participant adherence will then be averaged to obtain the cluster means and be used to calculate the distance between individual trajectories and cluster means. Using model prediction values instead of the mean of observations to form the cluster means and to calculate the distance between two trajectories allows difference in time points of adherence measurements among different participants. As we are dealing with adherence trajectories, we would want to account for correlation between repeated measures overtime for the same participant. Participant ID as random effects term is added in the GAM model fit. GAM is chosen here for more flexibility in terms of identifying non-linear regression effect in time. In this analysis of adherence trajectory clustering, the GAM models are performed with R package mgcv [20]. Restricted maximum likelihood (REML) method [21] is chosen for estimating smoothing parameters, and thin plate regression splines [22] are used to represent the smoothing functions.

**Number of clusters**

In K-means algorithm, the number of clusters, denoted here as $K_{ini}$, needs to be pre-specified. However, it is possible for the proposed approach to conclude in fewer clusters than $K_{ini}$: if during an *E-M* iteration, there are less than $M$ percent of individuals in one or more clusters, the smallest cluster will be dropped before reassigning the cluster membership. $M$ is a user-specified number that depends on the sample size as well as how tight one wants the clustering to be. The larger $M$ is, the tighter the clusters will be. The number of final clusters is denoted as $K_{fin}$ $(K_{fin} \leq K_{ini})$.

**Procedures of KGAM**

1. Initial cluster membership assignment. Participant $i, i = 1, \dots, n$ is randomly assigned to cluster $k, k = 1, \dots, K_{ini}$. With equal probability, $n_k$ participants are assigned to cluster $k$.
2. During each *E-M* alternation, each participant is assigned to the closest cluster.
    a. Obtain cluster-specific mean trajectories. Within each cluster:
        i. Build a GAM model.
        ii. Produce predictions of participant adherence at Month 0 – 11 given the original covariate values for model fit.
        iii. Calculate the values of mean adherence trajectory, $\widehat{y_{kt}}$, at Month $t, t = 0, \dots, 11$ by averaging over predictions of adherence for $n_k$ participants.
    b. Calculate the distance matrix and assign cluster membership.

        FOR participant $i, i = 1, \dots, n$

            FOR cluster $k, k = 1, \dots, K, distance_k = \sum_t (y_{it} - \widehat{y_{kt}})^2$

            New cluster membership $k_i = \min_k distance_k$

    c. Check the dropping group condition.

            IF any $n_k < M * n, k = 1, \dots, K$

            Remove the smallest cluster

            Repeat STEP b

            ELSE continue to the next *E-M* alternation.
3. Convergence. The *E-M* alternation stops if no participant is moved or maximum number of iterations is reached.

## Performance evaluation

The final clustering results will be evaluated and compared using the adapted Akaike's Information Criterion (AIC) [23] and Bayesian Information Criterion (BIC) [24]. Both AIC and BIC are parametric criteria that measure the goodness of fit of a statistical model when likelihood methods are implemented. AIC and BIC criteria are commonly used for comparing several non-nested models and choosing the best subsets of predictors. They are different in the size of the penalty and are recommended to use together in model selection. As AIC and BIC measure the information lost from the truth, smaller AIC and BIC indicate that information contained in the data are better explained and thus, more appropriate patterns of adherence trajectories are identified.

Suppose we now have a final clustering results with $K_{fin}$ clusters. Let $k, k = 1, \dots, K_{fin}$ denote the final clusters; $n_k$ be the number of participants assigned to cluster $k$; $m_{kj}, j = 1, \dots, n_k$ be the number of repeated measures for the *j*th participant in cluster $k$; and $n_{tot,k} = \sum_{j=1}^{n_k} m_{kj}$ be the number of total observations in cluster $k$. A GAM model is built for each final cluster. For the model of cluster $k$, let $p_k$ be the number of estimated parameters; $\widehat{L_k}$ be the maximum value of the likelihood function.

**AIC**

The AIC value of the cluster-specific model is calculated using Akaike's formula

$$AIC_k = -2 \log \widehat{L_k} + 2p_k$$

**Adapted AIC value for the final clustering results**

Since the $K_{fin}$ clusters contain all the participants exclusively, $\log \hat{L}$, the log of the maximum value of the likelihood function for the final clustering results will be the sum of the cluster-specific values, that is, $\log \hat{L} = \sum_{k=1}^{K_{fin}} \log \widehat{L_k}$. The total number of estimated parameters in final

clustering results is $p = \sum_{k=1}^{K_{fin}} p_k$. Thus, an adapted AIC expression for the final clustering results generated by KGAM can be written as

$$AIC_{kgam} = -2\log\hat{L} + p = \sum_{k=1}^{K_{fin}} AIC_k$$

The $AIC_{kgam}$ uses the likelihood of all the model coefficients evaluated at the penalized maximum likelihood estimation. The number of estimated parameters to be used is the corrected effective degrees of freedom as introduced in Wood's paper [22].

**BIC**

BIC introduces a larger penalty term compared with AIC, which accounts for the sample size besides the number of parameters. In the classic expression for BIC, the sample size to use is the total number of observations assumed to be independent [25]. In this analysis however, information associated with the model is affected by the within-participant correlation. Thus, more appropriate BIC penalty for mixed effects models should be applied here.

**Adapted BIC value for the final clustering results**

Delattre [26] proposes a BIC penalty term by deriving a Laplace approximation of the likelihood under a model with both fixed effects and random effects presented. Consider the GAM model for cluster $k, k = 1, \dots, K_{fin}$, Delattre's BIC expression is

$$BIC_k = -2\log\widehat{L_k} + \dim(\theta_{R,k})\log n_k + \dim(\theta_{F,k})\log n_{tot,k}$$

Where $\dim(\theta_{R,i})$ is the number of variance components of the individual random effects. and $\dim(\theta_{F,i})$ is the number of population parameters with fixed effects in the model. The new BIC penalty terms depend on both the number of participants and the number of repeated measures per participant. Thus, adapting from Delattre's work, a BIC expression for the final clustering results generated by KGAM can be written as

$$BIC_{kgam} = \sum_{k=1}^{K_{fin}} BIC_k$$

In consistent with the notations in AIC expression, we have $p_k$ equals the sum of $\dim(\theta_{R,k})$ and $\dim(\theta_{F,k})$.

# Results

## Application to MTN-025/HOPE Trial

Of the 1,456 eligible HIV-uninfected MTN-020/ASPIRE participants who agreed to continue their participation in the MTN-025/HOPE trial, 91 participants had no adherence information over the 12-month follow-up period and were excluded in the following analysis. The major reason for the missingness was that they never accepted a ring during the follow-up visits. Other reasons for the missingness include ring usage not applicable due to pregnancy, being permanently discontinued from product and ring not returned.

We performed the KGAM on the MTN-025/HOPE trial data in two settings. In the first setting, time in study (month, continuous) as a single predictor along with participant ID as random effects term were included in the GAM model fit for producing cluster means. In the second setting, additional participant characteristics were considered (Table 1). In preliminary analysis, there has been evidence that participants' baseline characteristics including age less than 25 years old (binary), use of oral contraceptive (binary), use of intrauterine device (IUD) (binary), having primary sex partner in the past 3 months (binary), experiencing menstrual bleeding in the past 3 months (binary), having anal sex in the past 3 months (binary), use of injectable contraceptive (binary) are significantly associated with participant adherence to the dapivirine VR. Thus, to explore the influence of participant characteristics on identifying the adherence trajectory patterns, we added the above baseline covariates into GAM model fit for producing cluster means.

Table 1. Baseline characteristics

| Baseline characteristics | N (%) |
|---|---|
| N | 1,365 |
| <25 years old | 164 (12.0%) |
| Oral contraceptive users | 188 (13.8%) |
| IUD users | 256 (18.8%) |
| Had primary sex partner | 1,321 (96.8%) |
| Any menstrual bleeding | 1,107 (81.2%) |
| Any anal sex | 89 (6.6%) |
| Injectable contraceptive users | 531 (38.9%) |

To explore the clustering results under different conditions and to test the algorithm's ability to drop redundant clusters, we ran the algorithm with initial number of clusters $K_{ini}$ equals 5, 6 or 7 and the dropping cluster criterion $M$ equals 0.1 or 0.15. Overall, 2 (covariates setting) $\times$ 3 (initiated number of clusters) $\times$ 2 (dropping cluster criterion) $=$ 12 combinations were evaluated. For each combination, KGAM was repeated 5 times with different randomly generated initial cluster membership. The adapted AIC ($AIC_{kgam}$) and BIC ($BIC_{kgam}$) criteria were used to evaluate the performance of KGAM results. Table 2 presents the best clustering results under each combination. The covariates setting, the clustering conditions including $K_{ini}$ and $M$ are indicated in columns 1 – 2; the number of final clusters, $K_{fin}$ is listed in column 3; the values of performance evaluation criteria $AIC_{kgam}$, $BIC_{kgam}$ are presented in columns 4 – 5. Both $AIC_{kgam}$ and $BIC_{kgam}$ suggested the combination of $K_{ini} = 7, M = 0.1$ and covariate setting where seven baseline covariates were added into GAM model fit for producing cluster means.

Table 2. Performance evaluation by $AIC_{kgam}$ and $BIC_{kgam}$.

| Covariates setting | Clustering conditions | $K_{fin}$ | $AIC_{kgam}$ | $BIC_{kgam}$ |
|---|---|---|---|---|
| Setting 1* | $K_{ini} = 5, M = 0.1$ | 5 | 41734 | 42908 |
| | $K_{ini} = 5, M = 0.15$ | 3 | 43057 | 44982 |
| | $K_{ini} = 6, M = 0.1$ | 5 | 41733 | 42898 |
| | $K_{ini} = 6, M = 0.15$ | 4 | 42122 | 44135 |
| | $K_{ini} = 7, M = 0.1$ | 5 | 41749 | 42696 |
| | $K_{ini} = 7, M = 0.15$ | 4 | 42407 | 43585 |
| Setting 2** | $K_{ini} = 5, M = 0.1$ | 4 | 42414 | 43747 |
| | $K_{ini} = 5, M = 0.15$ | 3 | 43061 | 45061 |
| | $K_{ini} = 6, M = 0.1$ | 5 | 41762 | 43063 |
| | $K_{ini} = 6, M = 0.15$ | 3 | 43061 | 45061 |
| | $K_{ini} = 7, M = 0.1$ | 6 | 41368 | 42160 |
| | $K_{ini} = 7, M = 0.15$ | 3 | 43061 | 45061 |

**In setting 1, time in study (month) as a single predictor along with participant ID as random effects term were included in the GAM model fit for producing cluster means **In setting 2, seven baseline covariates were added into GAM model fit for producing cluster means, including age less than 25 years old, use of oral contraceptive, use of intrauterine device (IUD), having primary sex partner in the past 3 months, experiencing menstrual bleeding in the past 3 months, having anal sex in the past 3 months and use of injectable contraceptive.*

We identified 6 clusters of dapivirine VR adherence (Table 3). The majority of participants were clustered as high adherence (36%) and *medium adherence* (24%). Of participants who showed high adherence, 51% were clustered by KGAM as *high adherence 1* where a tendency of achieving higher adherence at each quarterly visit at Month 3, 6 and 9 were indicated; 49% were clustered by KGAM as *high adherence 2* where they showed steady high adherence up to the quarterly visit at Month 9 with a bump of increase at Month 6, and had a tendency of achieving even higher adherence after Month 9. Among 28% participants who presented a trend of declining adherence, 46% presented consistent high adherence until Month 6 with declining adherence afterward and were clustered by KGAM as *high-declining adherence*. 54% presented

a declining adherence over the 12 months period and were clustered by KGAM as *declining adherence*. Finally, 12% were clustered as *non-adherence*. The predicted adherence trajectory along with the observed individual adherence trajectories were presented separately in Figure 1.

Figure 1. Six adherence trajectory clusters. Yellow line represents the predicted trajectory based on a GAM model in the corresponding cluster. Horizontal lines indicate target release by design (4 mg) and cutoff applied for non-use in previous studies (0.9 mg). Observed values can be less than zero due to measurement error.

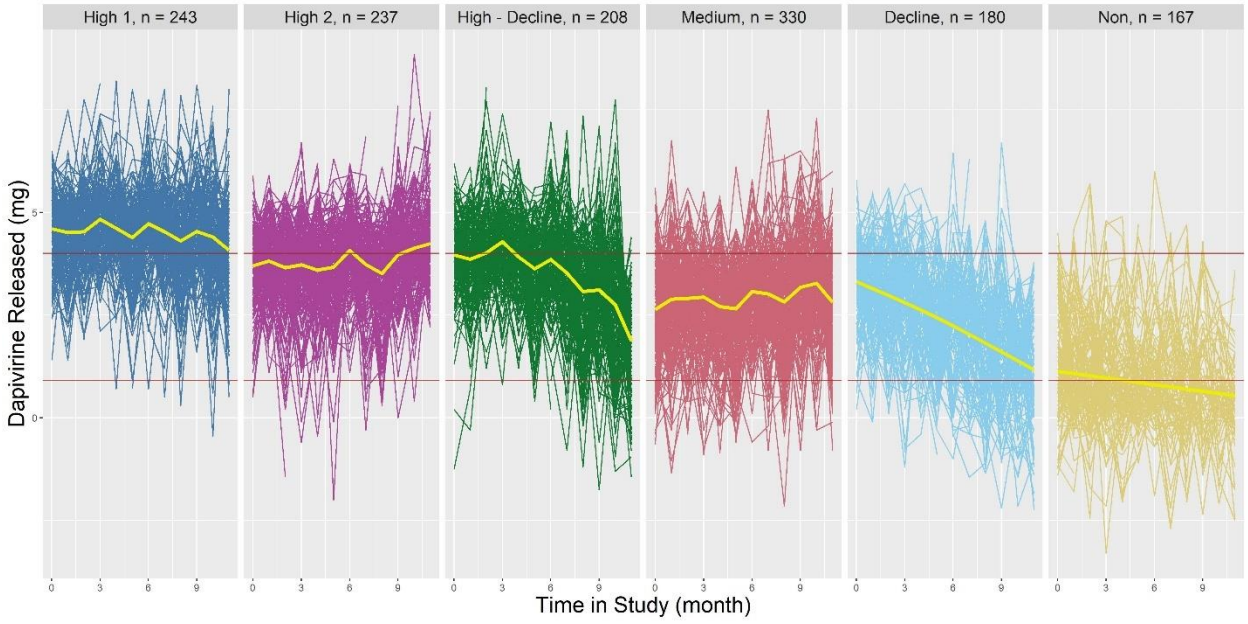


Table 3. Six adherence trajectory clusters and baseline characteristics

| Baseline characteristics | High adherence 1 (n = 243) | High adherence 2 (n = 237) | Medium adherence (n = 330) | High - declining adherence (n = 208) | Declining adherence (n = 180) | Non-adherent (n = 167) |
|---|---|---|---|---|---|---|
| <25 years old | 25 (10.3%) | 27 (11.4%) | 33 (10.0%) | 21 (10.1%) | 25 (13.9%) | 33 (19.8%) |
| Oral contraceptive users | 24 (9.9%) | 33 (13.9%) | 38 (11.5%) | 24 (11.5%) | 31 (17.2%) | 38 (22.8%) |
| IUD users | 22 (9.1%) | 47 (19.8%) | 94 (28.5%) | 42 (20.2%) | 35 (19.4%) | 16 (9.6%) |
| Had primary sex partner | 236 (97.1%) | 228 (96.2%) | 318 (96.4%) | 205 (98.6%) | 175 (97.2%) | 159 (95.2%) |
| Any menstrual bleeding | 170 (70.0%) | 186 (78.8%) | 287 (87.0%) | 171 (82.2%) | 153 (85.0%) | 140 (83.8%) |
| Any anal sex | 14 (5.8%) | 14 (6.0%) | 18 (5.5%) | 11 (5.3%) | 10 (5.6%) | 22 (13.3%) |
| Injectable contraceptive users | 137 (56.4%) | 102 (43.0%) | 74 (22.4%) | 75 (36.1%) | 71 (39.4%) | 72 (43.1%) |

## Associations between cluster membership and baseline characteristics

We assessed the associations between the seven baseline characteristics and the adherence trajectory cluster membership identified by the KGAM method. Associations were assessed individually using univariable logistic regression models with group membership as outcome, adjusting for study site. Associations significant at $p < 0.10$ were included in a multivariable regression model. This analysis will help us better understand questions include, (1) What factor(s) is (are) associated with being high or medium adherence? (2) Is there any factor that distinguishes high adherence and declining adherence? (3) Among participants who showed high adherence, is there any factor associated with being clustered in *high adherence group 1* instead of *high adherence group 2*?

Non-oral contraceptive use and lack of menstrual bleeding in the past 3 months were identified to be correlated with being clustered as high or medium adherence (Table 4). Specifically, oral contraceptive users had 0.38 ($p < 0.005$, 95% Confidence Interval (CI): [0.15, 0.55]) times lower odds of being clustered as high or medium adherence compared with non-users. Compared with participants lack of menstrual bleeding, those who had experienced menstrual bleeding in the past 3 months had 0.33 ($p < 0.01$, 95% CI: [0.10, 0.51]) times lower odds of being clustered as high or medium adherence. Correlates of being clustered as high or medium adherence were also identified for oral contraceptive non-users ($p = 0.01$) and participants who were lack of menstrual bleeding ($p = 0.02$) in the past 3 months in multivariable model adjusting for study site.

Table 4. Participant baseline characteristics in high or medium adherence groups versus other groups

| Baseline covariates | High or medium adherence (n = 810) | Others (n = 555) | OR (p-value) |
|---|---|---|---|
| <25 years old | 85 (10.5%) | 79 (14.2%) | 0.80 (p=0.20) |
| Oral contraceptive user | 95 (11.7%) | 93 (16.8%) | 0.62 (p<0.005) |
| IUD user | 163 (20.1%) | 93 (16.8%) | 1.15 (p=0.35) |
| Had primary sex partner | 782 (96.5%) | 539 (97.1%) | 0.77 (p=0.42) |
| Any menstrual bleeding | 643 (79.4%) | 464 (83.6%) | 0.67 (p<0.01) |
| Any anal sex | 46 (5.7%) | 43 (7.7%) | 0.77 (p=0.26) |
| Injectable contraceptive user | 313 (38.6%) | 218 (39.3%) | 1.09 (p=0.50) |

*OR = Odds Ratio estimation of being clustered as high or medium adherence. Estimations from univariable models, comparing women age less than 25 years old to age older than 25 years old, oral contraceptive users to non-users, IUD users to non-users, women with a primary sex partner in the past 3 month to women without, women who had anal sex to women who had not and injectable contraceptive users to non-users.*

Correlates of being clustered as declining adherence include use of IUD, having experienced menstrual bleeding in the past 3 months and non-use of injectable contraceptive (Table 5). More precisely, the odds of being clustered as declining adherence was 0.47 ($p = 0.04$, 95% CI: [0.02, 1.16]) times higher for IUD users compared with non-users. For injectable contraceptive users however, the odds of being clustered as declining adherence was 0.44 ($p < 0.001$, 95% CI: [0.25, 0.58]) times lower compared with non-users. The odds of being clustered as high adherence for women who have experienced menstrual bleeding in the past 3 months was 0.84 ($p < 0.001$, 95% CI: [0.29, 1.65]) times higher compared with women who were lack of menstrual bleeding. In multivariable model adjusting for study site, having experienced menstrual bleeding ($p = 0.01$) and non-use of injectable contraceptive ($p < 0.01$) were found to be associated with being clustered as declining adherence.

Significant associations between high adherence and the use of IUD, use of oral contraceptive, having had menstrual bleeding in the past 3 months and the use of injectable contraceptive were identified by Wald test (Table 6) in univariable models. Specifically, the odds of being clustered as *high adherence 2* was 1.99 ($p < 0.001$, 95% CI: [0.70, 4.39]) and 0.70 ($p = 0.07$, 95% CI: [-0.05, 2.08]) times higher for IUD users and oral contraceptive users respectively, compared with non-users. For injectable contraceptive users however, the odds of being clustered as *high adherence 2* was 0.51 ($p < 0.001$, 95% CI: [0.28, 0.68]) times lower compared with non-users. Participants who had menstrual bleeding in the past 3 months had 0.74 ($p = 0.02$, 95% CI: [0.11, 1.75]) times higher odds of being clustered as *high adherence 2* compared with participants who were lack of menstrual bleeding. Correlates of being clustered as *high adherence 2* including use of IUD ($p < 0.005$) were found in multivariable model adjusting for study site.

Table 5. Participant baseline characteristics in high adherence groups versus declining adherence groups

| Baseline covariates | High adherence (n = 480) | Declining adherence (n = 388) | OR (p-value) |
|---|---|---|---|
| <25 years old | 52 (10.8%) | 46 (11.9%) | 1.08 (p=0.72) |
| Oral contraceptive user | 57 (11.9%) | 55 (14.2%) | 1.25 (p=0.28) |
| IUD user | 69 (14.4%) | 77 (19.8%) | 1.48 (p=0.04) |
| Had primary sex partner | 464 (96.7%) | 380 (97.9%) | 1.72 (p=0.23) |
| Any menstrual bleeding | 356 (74.2%) | 324 (83.5%) | 1.84 (p<0.001) |
| Any anal sex | 28 (5.8%) | 21 (5.4%) | 0.91 (p=0.75) |
| Injectable contraceptive user | 239 (49.8%) | 146 (37.6%) | 0.56 (p<0.001) |

*OR = Odds Ratio estimation of being clustered as declining adherence. Estimations from univariable models, comparing women age less than 25 years old to age older than 25 years old, oral contraceptive users to non-users, IUD users to non-users, women with a primary sex partner in the past 3 month to women without, women who had anal sex to women who had not and injectable contraceptive users to non-users.*

Table 6. Participant baseline characteristics in high adherence group

| Baseline covariates | High adherence 1 (n = 243) | High adherence 2 (n = 237) | OR (p-value) |
|---|---|---|---|
| <25 years old | 25 (10.1%) | 27 (11.4%) | 1.02 (p=0.95) |
| Oral contraceptive user | 24 (9.9%) | 33 (13.9%) | 1.70 (p=0.07) |
| IUD user | 22 (9.1%) | 47 (19.8%) | 2.99 (p<0.001) |
| Had primary sex partner | 236 (97.1%) | 228 (96.2%) | 0.71 (p=0.51) |
| Any menstrual bleeding | 170 (70%) | 186 (78.5%) | 1.74 (p=0.02) |
| Any anal sex | 14 (5.8%) | 14 (5.9%) | 1.01 (p=0.97) |
| Injectable contraceptive user | 137 (56.4%) | 102 (43%) | 0.49 (p<0.001) |

*OR = Odds Ratio estimation of being clustered as high adherence 2. Estimations from univariable models, comparing women whose age were less than 25 years old to those older than 25 years old, oral contraceptive users to non-users, IUD users to non-users, women who had primary sex partner in the past 3 month to women who did not, women who had anal sex to women who did not and injectable contraceptive users to non-users.*

## Discussion

In this analysis we proposed a trajectory clustering method called KGAM and presented the applications of KGAM method to the HIV open-label prevention extension (MTN-025/HOPE) trial. We identified six trajectories of participant adherence to the dapivirine (25 mg) in a silicone elastomer Vaginal Ring (VR), which was found to be a safe and effective prevention of HIV-1 infection in healthy, sexually active, HIV-negative women in the previous trial, MTN-020/ASPIRE. Specifically, most participants in the HOPE trial were clustered as high or medium adherence through the 12-month period of ring use. However, we also identified declining adherence trajectories and a non-adherence trajectory. Non-use of oral contraceptive and lack of menstrual bleeding in the past 3 months were found to be associated with being clustered as high or medium adherence in logistic regression models. Correlates of being clustered as declining adherence include having experienced menstrual bleeding in the past 3 month and non-use of injectable contraceptive. Among participants who follow high adherence pattern, use of IUD was associated with being clustered as high adherence with a trend of increasing ring use at the end of the 12-month follow-up period.

The KGAM method presents some advantages compared with existing methods for trajectory clustering. KGAM provides scope for choosing the optimal cluster member. KGAM inherits the flexibility of the non-parametric clustering algorithms with relaxed assumptions on trajectory shape. By incorporating model fit to produce cluster means during the K-means process, KGAM can deal with missingness caused by difference in time points of adherence measurements for different participants and can account for factors that are considered to be associated with the outcome of trajectory. Meanwhile, the use of GAM models provides the potential for better fits of the relationship between participant adherence and time in study.

There are several limitations of the KGAM methods. As with all clustering algorithms, the determination of the optimal cluster number is still an open issue. KGAM enables dropping redundant clusters along the clustering process by setting a user-defined criterion. In the application to the MTN-025/HOPE trial data, we also examined different combinations of the initiated cluster number and the dropping criterion to find optimal clustering results. In this analysis, we evaluated the clustering performance using parametric criteria AIC and BIC. However, there are many other criteria exist (global posterior probability [27], Davies & Bouldin [28]). Also, one might consider choosing the number of clusters based on clinical relevance rather than an index.

In conclusion we presented a method for longitudinal trajectory clustering that incorporates advantages from both non-parametric clustering methods with the model-based techniques. In the applications to the MTN-025/HOPE trial, we identified distinct patterns of participant adherence to the dapivirine VR and found participants' baseline characteristics that were associated with trajectory clustering membership. Our results might be useful for understanding participant adherence to dapivirine VRs in this open-label trial and for planning future trials that implement the use of dapivirine VRs.

# Appendix A: Best clustering results in different combinations

Figure 2. Five adherence trajectory clusters. $K_{ini} = 5$, $M = 0.1$, baseline covariates not included in GAM model fit.

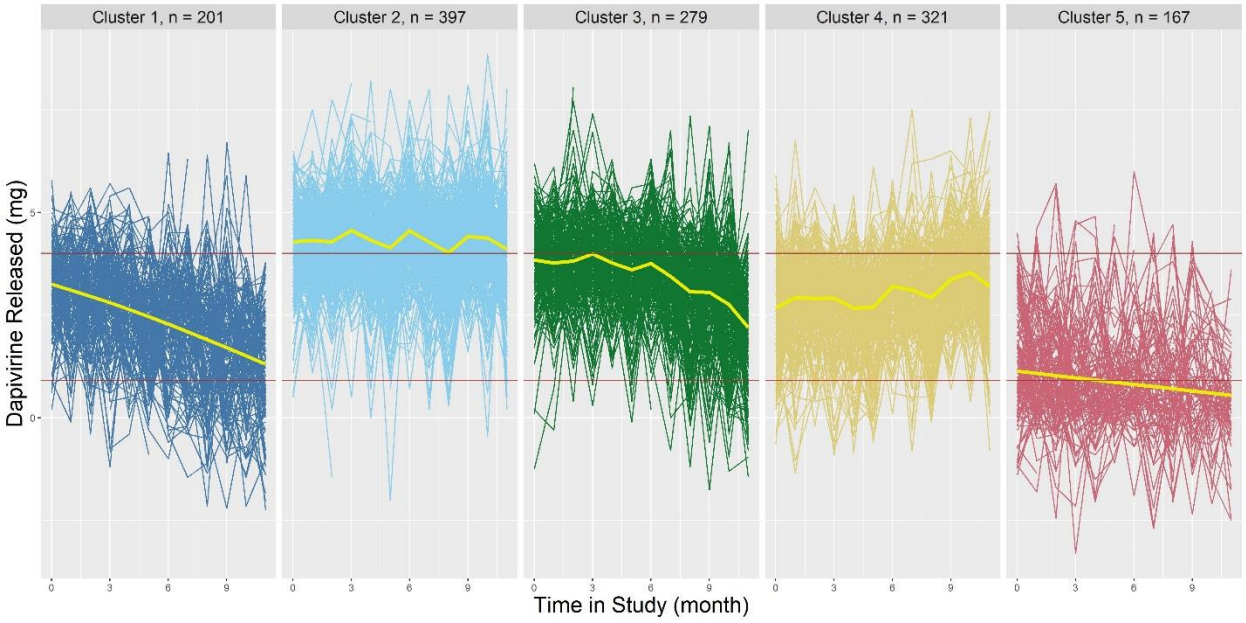



Figure 3. Three adherence trajectory clusters. $K_{ini} = 5$, $M = 0.15$, baseline covariates not included in GAM model fit.

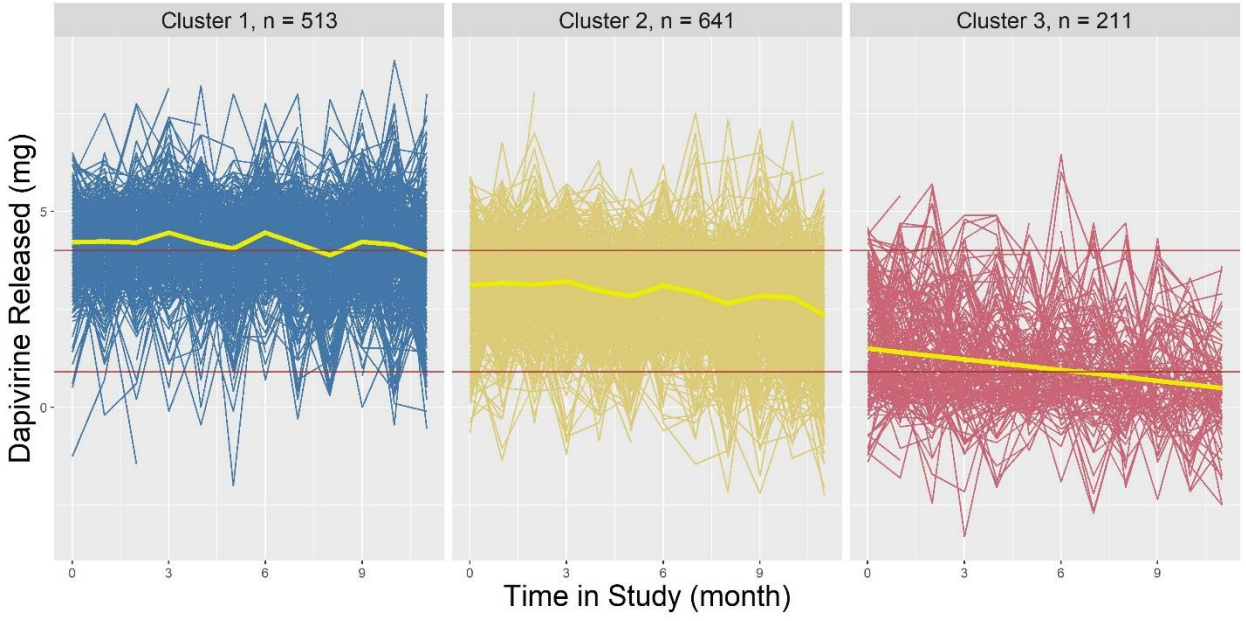


Figure 4. Five adherence trajectory clusters. $K_{ini} = 6$, $M = 0.1$, baseline covariates not included in GAM model fit.

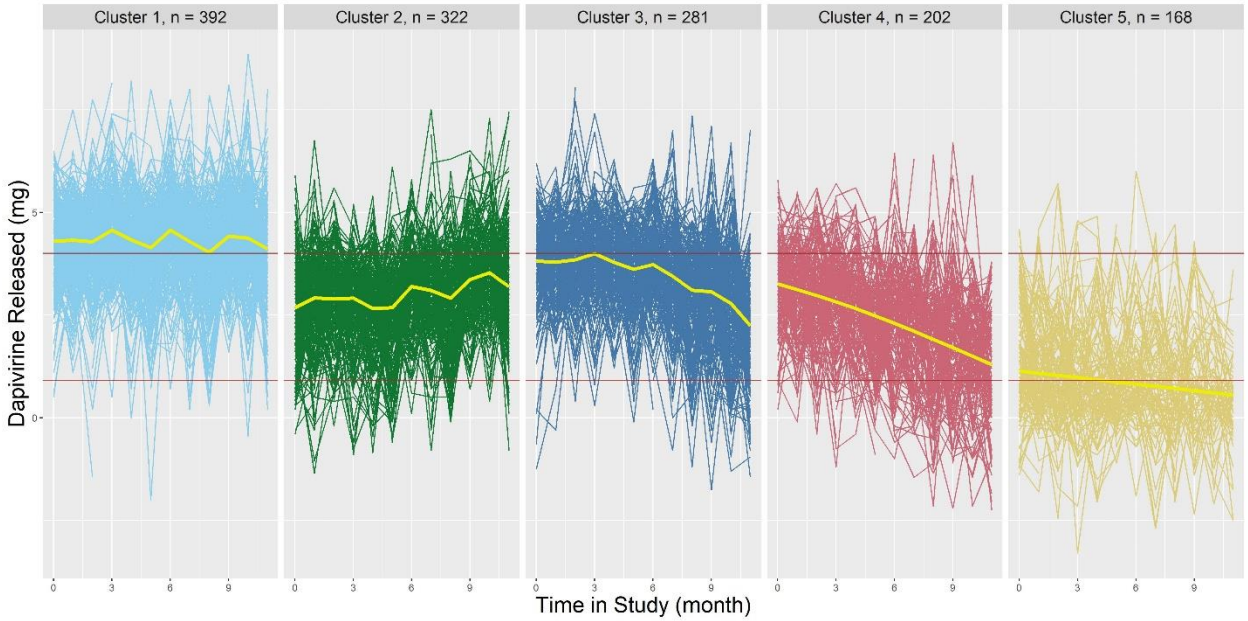



Figure 5. Four adherence trajectory clusters. $K_{ini} = 6$, $M = 0.15$, baseline covariates not included in GAM model fit.

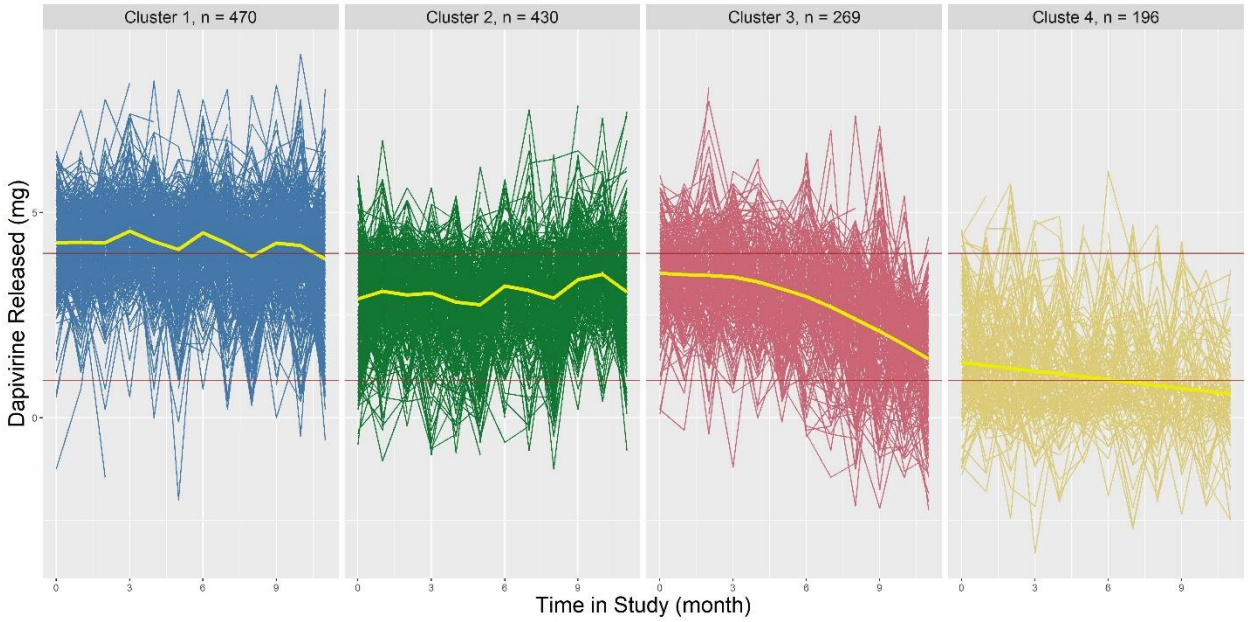


Figure 6. Five adherence trajectory clusters. $K_{ini} = 7$, $M = 0.1$, baseline covariates not included in GAM model fit.

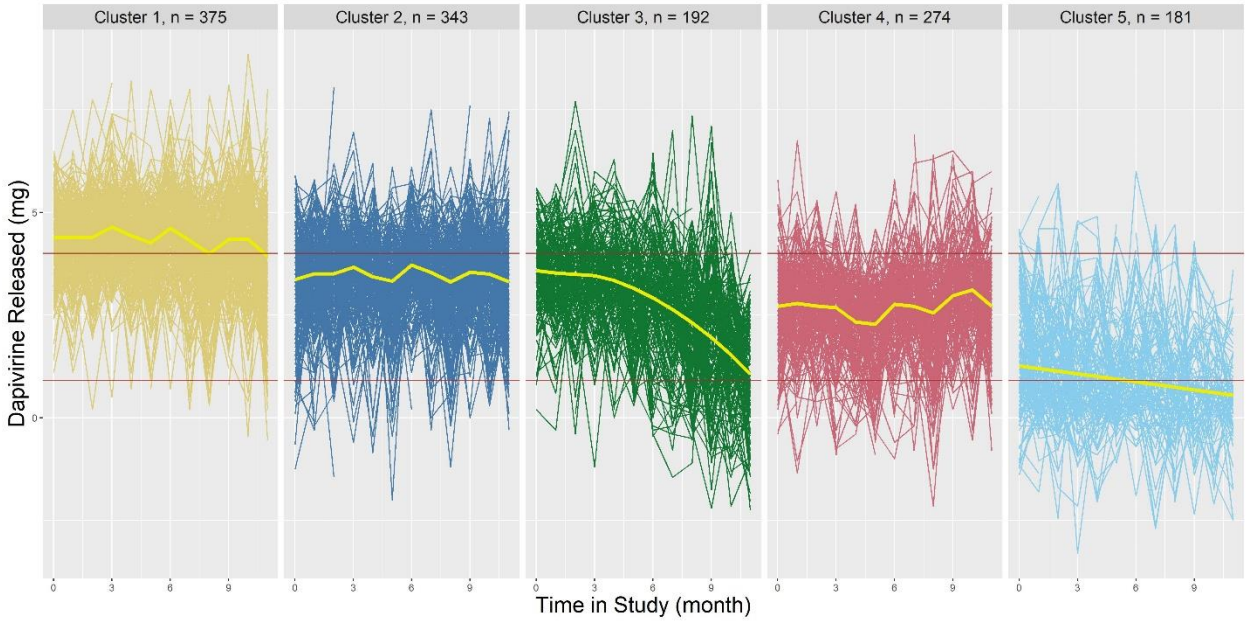



Figure 7. Four adherence trajectory clusters. $K_{ini} = 7$, $M = 0.15$, baseline covariates not included in GAM model fit.

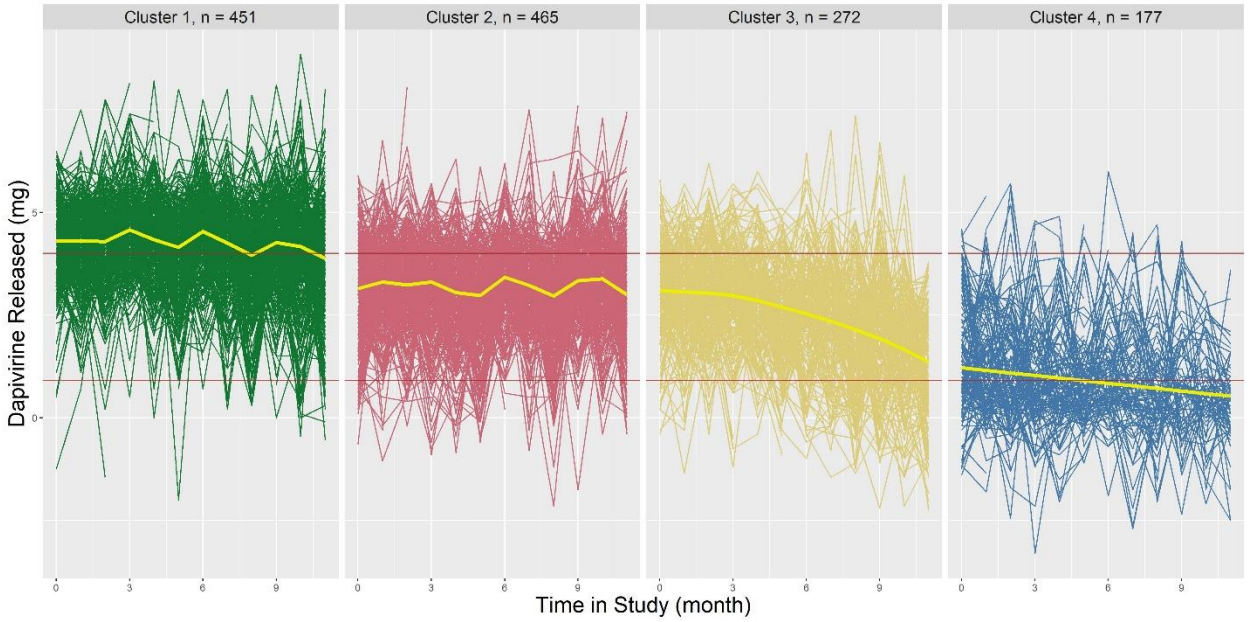


Figure 8. Four adherence trajectory clusters. $K_{ini} = 5$, $M = 0.1$, baseline covariates included in GAM model fit.

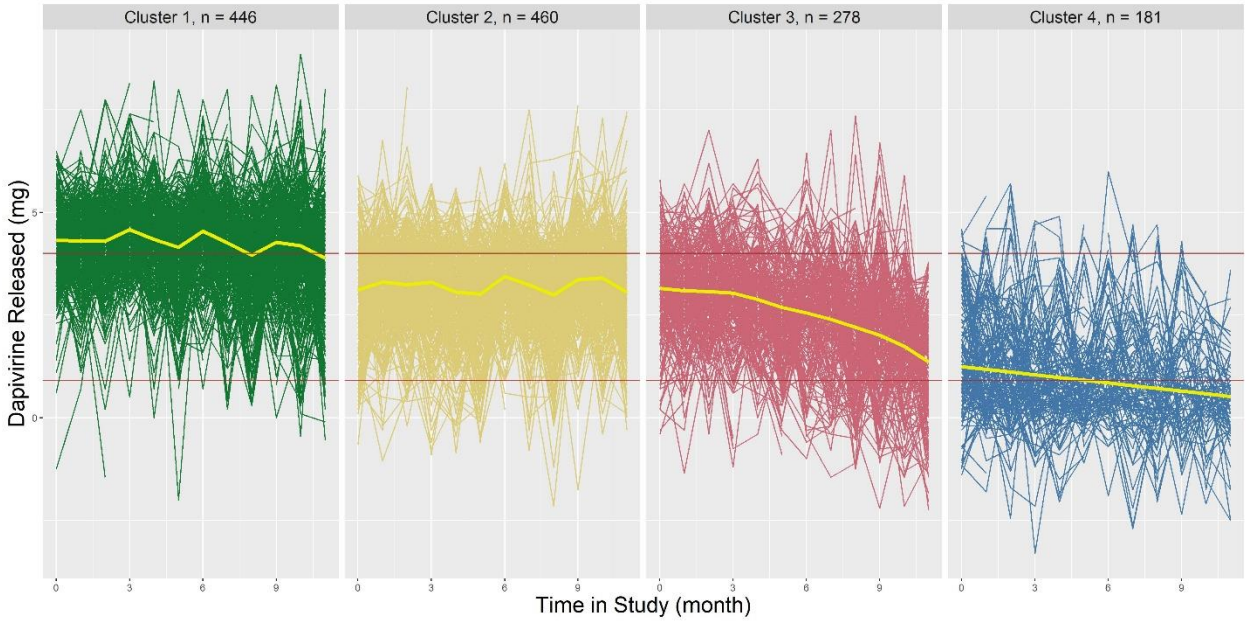



Figure 9. Three adherence trajectory clusters. $K_{ini} = 5$, $M = 0.15$, baseline covariates included in GAM model fit.

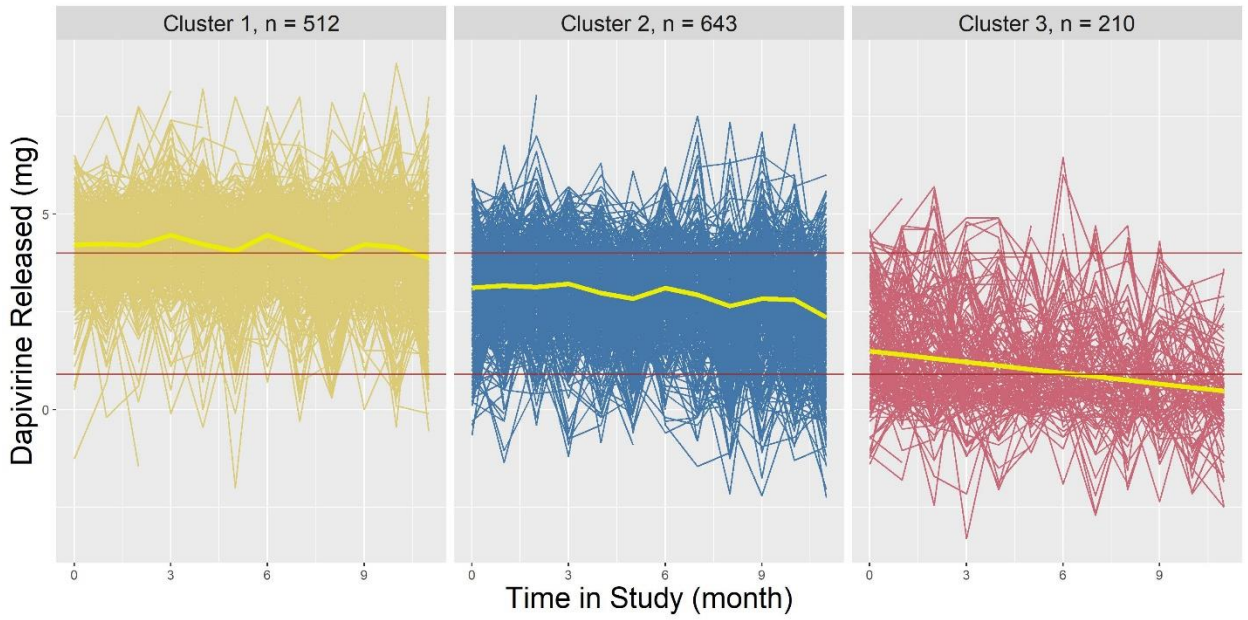


Figure 10. Five adherence trajectory clusters. $K_{ini} = 6$, $M = 0.1$, baseline covariates included in GAM model fit.

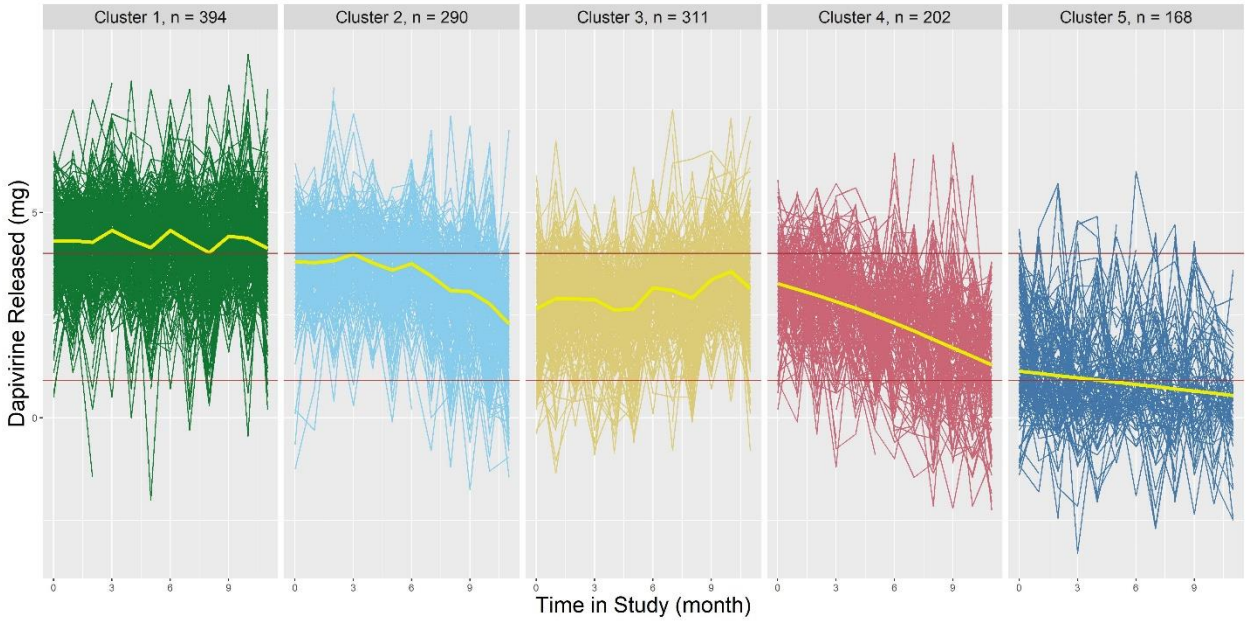



Figure 11. Three adherence trajectory clusters. $K_{ini} = 6$, $M = 0.15$, baseline covariates included in GAM model fit.

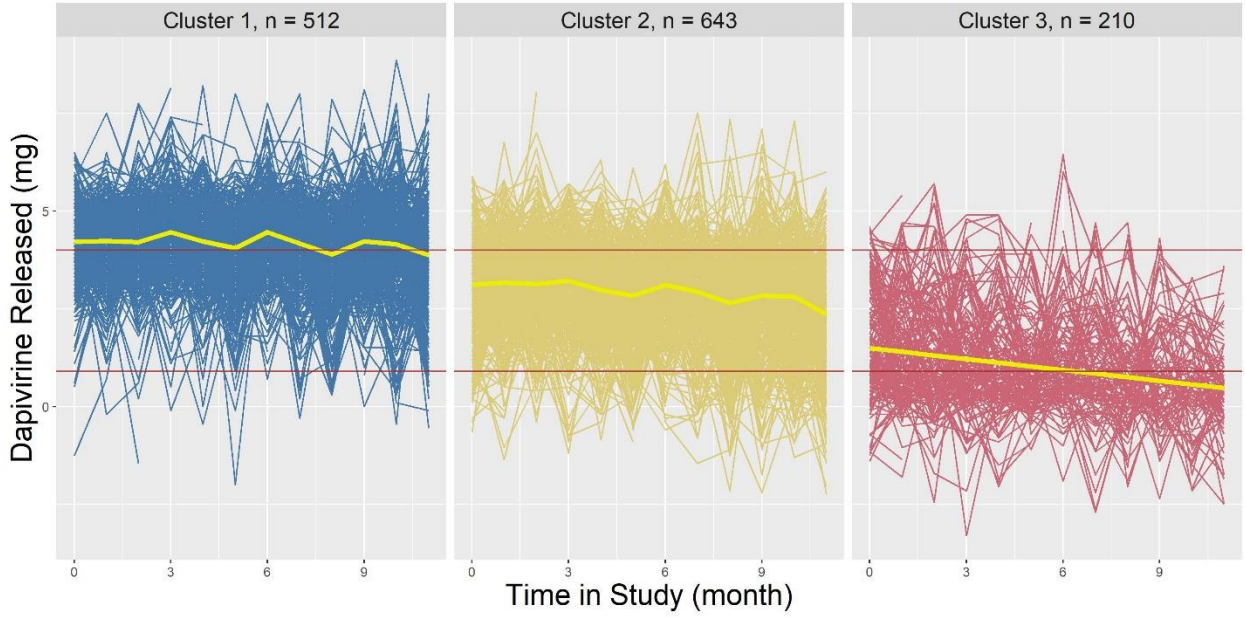
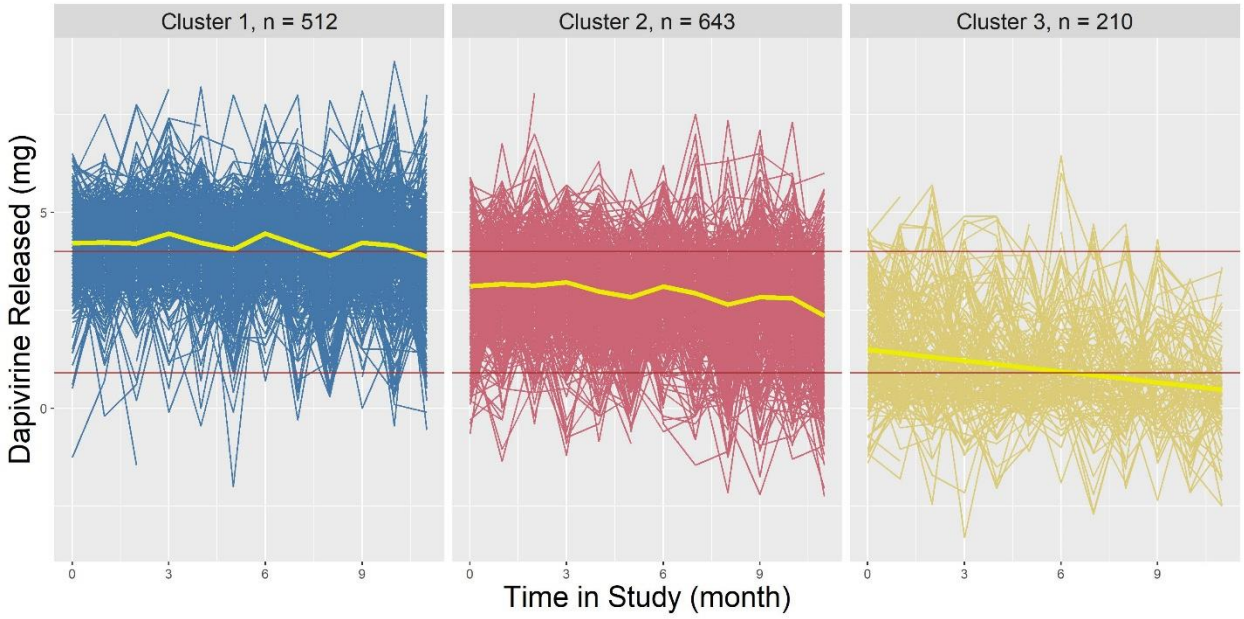


Figure 12. Three adherence trajectory clusters. $K_{ini} = 7$, $M = 0.15$, baseline covariates included in GAM model fit.



1. Avert (last reviewed 21 May 2019) 'Women and Girls, HIV and AIDS' [accessed 1 June 2020]; https://www.avert.org/professionals/hiv-social-issues/key-affected-populations/women
2. Fact Sheet – World AIDS DAY 2019. UNAIDS; June 2019. https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf
3. Baeten JM, Palanee-Phillips T, Brown ER, et al. Use of a Vaginal Ring Containing Dapivirine for HIV-1 Prevention in Women. *N Engl J Med*. 2016;375(22):2121-2132. doi:10.1056/NEJMoa1506110
4. ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000 Feb 29 -. Identifier NCT02858037, Trial to Assess the Continued Safety of and Adherence to a Vaginal Ring Containing Dapivirine in Women; 2016 August 8 [cited 2020 June 1]. Available from: https://clinicaltrials.gov/ct2/show/study/NCT02858037
5. Genolini, Christophe, and Bruno Falissard. "KmL: k-means for longitudinal data." *Computational Statistics* 25.2 (2010): 317-328.
6. Magidson, Jay, and Jeroen Vermunt. "Latent class models for clustering: A comparison with K-means." Canadian Journal of Marketing Research 20.1 (2002): 36-43.
7. Subtil, Fabien, et al. "An alternative classification to mixture modeling for longitudinal counts or binary measures." *Statistical Methods in Medical Research* 26.1 (2017): 453-470.
8. Genolini, Christophe, et al. "kmlShape: an efficient method to cluster longitudinal data (time-series) according to their shapes." *Plos one* 11.6 (2016): e0150738.
9. Nagin, Daniel S., and Daniel NAGIN. *Group-based modeling of development*. Harvard University Press, 2005.
10. Nagin, Daniel, and Richard E. Tremblay. "Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency." *Child development* 70.5 (1999): 1181-1196.
11. Kreuter, Frauke, and Bengt Muthén. "Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling." *Journal of Quantitative Criminology* 24.1 (2008): 1-31.
12. Van Ryzin, Mark J., et al. "Identifying atypical cortisol patterns in young children: the benefits of group-based trajectory modeling." *Psychoneuroendocrinology* 34.1 (2009): 50-61.
13. Sagaon-Teyssier, Luis, et al. "A group-based trajectory model for changes in pre-exposure prophylaxis and condom use among men who have sex with men participating in the ANRS IPERGAY Trial." *AIDS Patient Care and STDs* 32.12 (2018): 495-510.
14. Pyra, Maria, et al. "Patterns of oral PrEP adherence and HIV risk among Eastern African women in HIV serodiscordant partnerships." *AIDS and Behavior* 22.11 (2018): 3718-3725.
15. Nagin, Daniel S., and Candice L. Odgers. "Group-based trajectory modeling in clinical research." *Annual review of clinical psychology* 6 (2010): 109-138.
16. Frankfurt, Sheila, et al. "Using group-based trajectory and growth mixture modeling to identify classes of change trajectories." *The Counseling Psychologist* 44.5 (2016): 622-660.
17. Proust-Lima C, Philipps V, Liquet B (2017). "Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm." *Journal of Statistical Software*, 78(2), 1–56. doi: 10.18637/jss.v078.i02.
18. Proust-Lima C, Philipps V, Diakite A, Liquet B (2020). *lcmm: Extended Mixed Models Using Latent Classes and Latent Processes*. R package version: 1.9.1, https://cran.r-project.org/package=lcmm.

19. Celeux, Gilles, and Gérard Govaert. "A classification EM algorithm for clustering and two stochastic versions." *Computational statistics & Data analysis* 14.3 (1992): 315-332.
20. Wood S (2017). *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC.
21. Wood, Simon N. "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1 (2011): 3-36.
22. Wood, Simon N. "Thin plate regression splines." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1 (2003): 95-114.
23. Akaike, Hirotugu. "A new look at the statistical model identification." *IEEE transactions on automatic control* 19.6 (1974): 716-723.
24. Schwarz, Gideon. "Estimating the dimension of a model." *The annals of statistics* 6.2 (1978): 461-464.
25. Wit, Ernst, Edwin van den Heuvel, and Jan-Willem Romeijn. "'All models are wrong...': an introduction to model uncertainty." *Statistica Neerlandica* 66.3 (2012): 217-236.
26. Delattre, Maud, Marc Lavielle, and Marie-Anne Poursat. "A note on BIC in mixed-effects models." *Electronic journal of statistics* 8.1 (2014): 456-475.
27. Bolstad, William M., and James M. Curran. *Introduction to Bayesian statistics*. John Wiley & Sons, 2016.
28. Davies, David L., and Donald W. Bouldin. "A cluster separation measure." *IEEE transactions on pattern analysis and machine intelligence* 2 (1979): 224-227.