A New Clustering Method for Longitudinal Data

Junyi Zhou
Department of Biostatistics and Health Data Science
Indiana University
Indianapolis
junyzhou@iu.edu

Ying Zhang
Department of Biostatistics
University of Nebraska Medical Center
Omaha
ying.zhang@unmc.edu

Wanzhu Tu
Department of Biostatistics and Health Data Science
Indiana University
Indianapolis
wtu1@iu.edu

Abstract

Longitudinal data clustering is challenging because the grouping has to account for the similarity of individual trajectories instead of the localized closeness at given time points. This paper puts forward a hierarchical agglomerative clustering method based on B-spline curve-fitting of repeatedly measured outcomes and a dissimilarity metric that quantifies the cost of merging two distinct groups of curves. Extensive simulations showed that the proposed method had superior clustering performance in terms of accuracy and efficiency. Importantly, the method is extendable to situations of multiple outcome clustering without much increased computational burden. To illustrate the use of the proposed clustering method, we analyzed data from two clinical studies.

***K*eywords** B-splines · Clustering · Dissimilarity metric · Longitudinal data · Multiple-outcomes

1 Introduction

Clustering is a data exploration technique that divides subjects into subgroups or clusters so that subjects within a cluster share a greater commonality than those from other clusters. The metric that one uses to quantify this commonality defines a specific clustering method. In the absence of objectively determined cluster labels, cluster analysis is typically considered an unsupervised learning technique to uncover hidden homogeneous subgroups. Fundamental ideas behind the modern clustering algorithms can be traced back for decades to the pioneering work on K-means [1], hierarchical clustering [2, 3], and density-based clustering [4]. With more efficient algorithms, modern cluster analysis has become an essential component of the data mining and machine learning toolbox. The method has been successfully applied to fields of pattern recognition, imaging processing, and clinical medicine.

The clustering of longitudinal data is generally a more challenging task because of the need to assess the similarity of longitudinal trajectories – localized proximity at given time points provides no assurance that temporal shapes of the trajectories will be similar. Clustering of longitudinal data tends to be simpler when outcomes are assessed regularly at

the same time points. For example, [5] proposed a K-means method for longitudinal data (KmL) that measured pairwise Euclidean distance between two longitudinal trajectories. Alternatively, one could fit functional curves by using a fixed number of basis functions and then applying the K-means clustering methods based on the pairwise Euclidean distance between two sets of fitted coefficients. Along this line, [6] and [7] proposed K-means clustering algorithms based on B-spline basis functions. In a similar spirit, [8] described an algorithm that used Fourier basis functions. Such K-means algorithms effectively alleviate the constraints on observational regularity required by KmL. Still, they tend to be quite inefficient with sparsely observed data.

When it comes to sparse and irregularly measured longitudinal data, model-based methods offer some advantages, especially considering the maturity of longitudinal modeling. Following such a strategy, several clustering methods have been developed; some were based on mixtures of linear models [9, 10, 11], others on mixed-effects models [12, 13, 14, 15]. When the models are correctly specified, these methods usually have respectable performance, even in sparsely observed data situations, because the model parameters are estimated with aggregated data from all subjects [12]. A significant issue with the model-based methods, however, is the lack of verifiability of the model assumptions [16]. Issues with identifiability could also arise when overfitting occurs [17]. In general, the computational burden of model-based methods tends to be heavy because different mixture models must be explored to determine the number of clusters according to the Bayesian Information Criterion (BIC) [18]. To remedy, [10] used a robust EM algorithm [19] to ascertain the number of mixtures of regression models automatically. More recently, [20] proposed a penalized regression method that selects the number of clusters simultaneously with parameter estimation. However, this method appears to be numerically inefficient, as we shall demonstrate in Section 3.1.

The current paper puts forward a new clustering method for longitudinal data with sparse and irregular observations. The method falls into the broad category of hierarchical agglomerative algorithms. In other words, it is a "bottom-up" program that combines trajectories when it goes up in the hierarchy if the "cost" for merging two trajectories is low [21]. We show that the proposed method is numerically efficient compared to the model-based methods, and it is robust against the variability of underlying cluster sizes. The algorithm exhibits improved accuracy in determining the number of clusters, in addition to the enhanced efficiency. The method is extendable to handle multiple-outcome longitudinal data, where some of the outcomes may not be easily distinguishable between clusters. Viewed as a whole, we believe that the proposed method provides a reliable clustering tool for longitudinal data. We further implemented the method in an R package and wrapped the main clustering function in an R Shiny app so that analysts not familiar with the R software can still access the clustering function through the interactive Shiny interface.

The rest of the paper is organized as follows: In Section 2, we describe the proposed clustering method and indices for determining the number of clusters. In Section 3, we present numerical results from an extensive simulation study; the performance of the proposed method were compared with competing methods. In Section 4, we present two real-data applications to demonstrate the use of the new clustering method. We end the paper in Section 5 with a few remarks. All technical details are included in the Appendix.

2 Methods

An established approach for clustering is to carry out the grouping in a hierarchical fashion [21] while using the number of clusters estimated by a modified gap statistic for cross-sectional data [22]. We proposed a new similarity metric based on the concept of "merging cost"; the metric is structurally analogues to the classic Chow test statistic [23]. We shall demonstrate the good operating characteristics of the proposed algorithm.

2.1 Modeling longitudinal trajectories

We first introduce the notation necessary for describing the temporal trajectories of longitudinally measured outcomes.

Let $\mathcal{C} := \{(t_{ij}, Y_{ij})_{i=1,2,...,N, j=1,2,...,n_i}\}$ be the collection of observed data from N subjects, where the ith subject has n_i observations $Y_{i1}, \ldots, Y_{in_i}$, made at times $\mathbf{t}_i = \{t_{ij}\}_{j=1}^{n_i}$. The observation times are subject-specific and they do not have to be common across subjects.

Suppose that the longitudinal trajectory for the ith subject can be described by the following model

$$Y_{ij} = f_i(t_{ij}) + e_i(t_{ij}), \tag{1}$$

where $f_i(\cdot)$ is the fixed component, and $e_i(\cdot)$ is the random component with mean zero.

To retain the maximal flexibility, we avoid specifying a parametric functional form for $f_i(\cdot)$. Instead, we use B-splines to to approximate $f_i(\cdot)$ for $i = 1, 2, ..., N$. The splines are fitted using the observed data. Letting $\mathbf{X}_{ij} = [B_1(t_{ij}), B_2(t_{ij}), ..., B_p(t_{ij})]$ be the row vector of the B-spline basis functions evaluated at t_{ij}, we fit (1) by rewriting the model as

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}_i + e_i(t_{ij}), \tag{2}$$

where $\boldsymbol{\beta}_i = [\beta_{i1}, ..., \beta_{ip}]^T$ are the spline coefficients for $f_i(\cdot)$.

Writing $\mathbf{Y}_i = \begin{bmatrix} Y_{i1} & \cdots & Y_{in_i} \end{bmatrix}^T$ and $\mathbf{X}_i = \begin{bmatrix} \mathbf{X}_{i1}^T & \cdots & \mathbf{X}_{in_i}^T \end{bmatrix}^T$, we express model (2) as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{e}_i, \tag{3}$$

where the random component $\mathbf{e}_i$ satisfies (i) $E[\mathbf{e}_i] = \mathbf{0}_{n_i}$; (ii) $Var[\mathbf{e}_i] = G_i(\mathbf{t}_i)$; and (iii) $\mathbf{e}_i$ is independent of $\mathbf{e}_j$ for $i \neq j$.

2.2 Cost of merging two subgroups

Clustering is a process of partitioning an index set $c = \{1, 2, ..., N\}$ into mutually exclusive subsets $\{c_k : k = 1, 2, ..., K\}$, where

$$\bigcup_{k=1}^{K} c_k = c \text{ and } c_u \bigcap_{u \neq v} c_v = \phi,$$

such that subjects in the same set are homogeneous in some sense.

In longitudinal data clustering, we seek partitions $\{c_k\}_{k=1}^K$ where the subjects in cluster c_k share a *common* longitudinal trajectory. In other words, $c_k = \{i \in c : f_i(\cdot) = f_{c_k}(\cdot)\}$. This process requires fitting the fixed component given in (1).

We denote the observed data associated with cluster c_k as $\mathcal{C}_k$, i.e.,

$$\mathcal{C}_k := \{(\mathbf{X}_{ij}, Y_{ij})_{i \in c_k, j=1,...,n_i}\}, \quad k = 1, 2, ..., K.$$

Hence, $\mathcal{C} = \bigcup_{k=1}^K \mathcal{C}_k$. Similarly, we merge the observations for the two clusters, and let $\mathcal{C}_{u,v}$ be the union of $\mathcal{C}_u$ and $\mathcal{C}_v$, that is, $\mathcal{C}_{u,v} = \{\mathcal{C}_u, \mathcal{C}_v\}$.

To determine the appropriateness of clustering, one needs a metric to quantify the cost of combining $\mathcal{C}_u$ and $\mathcal{C}_v$. When the cost is low, the two subsets could be combined with a shared trajectory. When the cost is high, combining the subsets would leads to larger errors in the combined model, and thus should be avoided. Errors associated with statistical models are typically described by sum of squared residuals (SSR). We therefore contend that the following ratio of SSRs under the separate and combined models gives a quantification for the merging cost:

$$\mathcal{D}(\mathcal{C}_u, \mathcal{C}_v) = \frac{\big(\text{SSR}(\mathcal{C}_{u,v}) - \text{SSR}(\mathcal{C}_u) - \text{SSR}(\mathcal{C}_v)\big)/p}{(\text{SSR}(\mathcal{C}_u) + \text{SSR}(\mathcal{C}_v))/(n_u + n_v - 2p)}, \tag{4}$$

where

$$\text{SSR}(\mathcal{C}_k) = \sum_{i \in c_k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2.$$

In the calculation, $\hat{Y}_{ij} = \hat{f}_{c_k}(t_{ij})$ and β_is are assumed to be identical for $\forall i \in c_k$.

Alternatively, $\mathcal{D}(\mathcal{C}_u, \mathcal{C}_v)$ can also be viewed as a metric of *dissimilarity* between two sets of longitudinal data, $\mathcal{C}_u$ and $\mathcal{C}_v$. A larger $\mathcal{D}(\mathcal{C}_u, \mathcal{C}_v)$ value indicates greater dissimilarity between $\mathcal{C}_u$ and $\mathcal{C}_v$, and thus providing stronger evidence against merging the two subsets.

Remark 1. The numerator of $\mathcal{D}(\cdot, \cdot)$ is conceptually similar to the "merging cost" proposed by [21] under the classical linear models. We note that for any $\mathcal{C}_u, \mathcal{C}_v$, $\mathcal{D}(\mathcal{C}_u, \mathcal{C}_v) \geq 0$, meaning that merging actions would not incur negative costs. See Proposition 1 in the Appendix.

Remark 2. We also note that the structure of $\mathcal{D}(,)$ is parallel to that of the Chow-test statistic [23], which was constructed to test whether two datasets would result in the same coefficients in *linear models of the same structure*. Viewed from this perspective, two subsets with the smallest $\mathcal{D}(,)$ value among all possible pairs of subsets are least likely to be rejected for homogeneity by the Chow-test. Hence sequentially merging two subsets with the smallest $\mathcal{D}(,)$ value gives rise to an agglomerative hierarchical clustering algorithm.

Remark 3. Metric $\mathcal{D}(,)$ is readily extendable to situations of longitudinal clustering with multiple outcomes: A weighted sum of $\mathcal{D}$ values from different outcomes represents the *overall* merging cost

$$\mathcal{D}(\mathcal{C}_u, \mathcal{C}_v) = \sum_h W_h \mathcal{D}(\mathcal{C}_u^{(h)}, \mathcal{C}_v^{(h)}),$$

where W_h is the standardized weight for outcome h, and $\mathcal{C}_k^{(h)} = \{(\mathbf{X}_{ij}, Y_{ij}^{(h)})_{i\in c_k, j=1,\ldots,n_i}\}$ contains data in cluster k for outcome h.

For the purpose of maintaining computing efficiency, we propose an *ad-hoc* method to determine the weights by expanding the Chow-test statistics as defined above,

$$W_h^* = \frac{(\text{SSR}(\mathcal{C}^{(h)}) - \sum_{i=1}^N \text{SSR}(\mathcal{C}_i^{(h)}))/((N-1)p)}{\sum_{i=1}^N \text{SSR}(\mathcal{C}_i^{(h)})/(|\mathcal{C}^{(h)}| - Np)}, \tag{5}$$

where $\mathcal{C}^{(h)}$ includes all subjects, and $\mathcal{C}_i^{(h)}$ is the "bottom-level cluster" that contains only observations from the ith subject as long as the data allow for fitting an individual-specific spline for calculating $\text{SSR}(\mathcal{C}_i^{(h)})$. Some pre-processing steps are needed when data from an individual are not sufficient to fit the spline model, as we shall illustrate in Section 2.4. The outcome-specific weight W_h can be chosen as $W_h = W_h^* / \sum_i W_i^*$.

2.3 Number of Clusters

Choosing an optimal number of clusters is an essential component of all clustering methods. Since there is no universally accepted criteria to define the "optimality" of data partitioning, solutions tend to be problem-specific. Two approaches are frequently used in the literature. One is to optimize the difference of within-cluster (or intra-cluster) dissimilarity and between-cluster (or inter-cluster) dissimilarity, such as the CH index [24] and Silhouette statistic [25]. The other is the Gap statistic that maximizes the "gap" between within-cluster dissimilarity and its expectation under the hypothesis that data are fully homogeneous [22].

In this work, we explore both approaches in the context of longitudinal data clustering. For a given a partition of K clusters, the within-cluster dissimilarity can be defined as

$$W(K) = \sum_{k=1}^K \sum_{i\in c_k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2 = \sum_{k=1}^K \text{SSR}(\mathcal{C}_k),$$

and the between-cluster dissimilarity as

$$B(K) = \sum_{k=1}^K \sum_{i\in c_k} \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \tilde{Y}_{ij})^2,$$

where $\tilde{Y}_{ij} = \hat{f}(t_{ij})$ is the estimated outcome using all data, for which $\hat{f}(\cdot)$ is the spline estimator of the population mean function. CH index at K partitions is therefore defined by

$$\text{CH}(K) = \frac{B(K)/(K-1)}{W(K)/(\sum_i n_i - Kp)}$$

for longitudinal data. The optimal number of clusters corresponds to the maximum value of CH index, i.e., $\hat{K} = \arg\max_K \text{CH}(K)$.

Similarly, we propose a new Gap statistic using the between-cluster dissimilarity, defined as

$$B'(K) = \sum_{k=1}^{K} \{\text{SSR}(\mathcal{C}) - \text{SSR}(\mathcal{C}_k) - \text{SSR}(\mathcal{C}_{-k})\},$$

where $\mathcal{C}_{-k}$ represents data from all subjects except those in set c_k. The new Gap statistic is therefore

$$\text{Gap}_b(K) = B'(K) - E[B'(K)]. \tag{6}$$

The expectation in equation (6) can be calculated under certain data homogeneity assumptions. For example, [22] assumed that the observed data are uniformly distributed in the sample space for clustering cross-sectional data and they used the bootstrap method to compute the expectation. For longitudinal data, however, this approach appears to be numerically intractable. To maintain numerical efficiency, we propose an ad-hoc procedure to compute the expectation under the following homogeneous longitudinal spline model:

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}_i + e_{ij} \tag{7}$$

where $E[e_{ij}] = 0$ and $Var[e_{ij}] = \sigma^2$. Under this model,

$$E[B'(K)] = \sum_{k=1}^{K} E[\text{SSR}(\mathcal{C}) - \text{SSR}(\mathcal{C}_k) - \text{SSR}(\mathcal{C}_{-k})] = Kp\sigma^2,$$

following the result of Proposition 2 in the Appendix. Hence, we have

$$\text{Gap}_b(K) = B'(K) - Kp\sigma^2,$$

where σ^2 can be estimated by fitting the B-spline model with all the data. Following the recommendation of [22], we use the first turning point as the chosen number of clusters.

Following [22], we choose

$$\hat{K} = \min\{K : \text{Gap}_b(K) > \text{Gap}_b(K-1), \text{Gap}_b(K) > \text{Gap}_b(K+1)\}, \text{for } K \geq 2,$$

as the optimal number of clusters.

Remark 4. Numerical experiments showed that $\text{Gap}_b(K)$ is more sensitive to pick out clusters with extreme cluster size compared to the Gap Statistic based on within-cluster dissimilarity in addition to its convenience in computing as bootstrap is no longer required. We therefore recommend using the modified Gap statistic to determine the number of clusters in practice.

Remark 5. Both $\text{CH}(K)$ and $\text{Gap}_b(K)$ can be extended to longitudinal clustering with multiple outcomes by adding the SSR for each individual outcome.

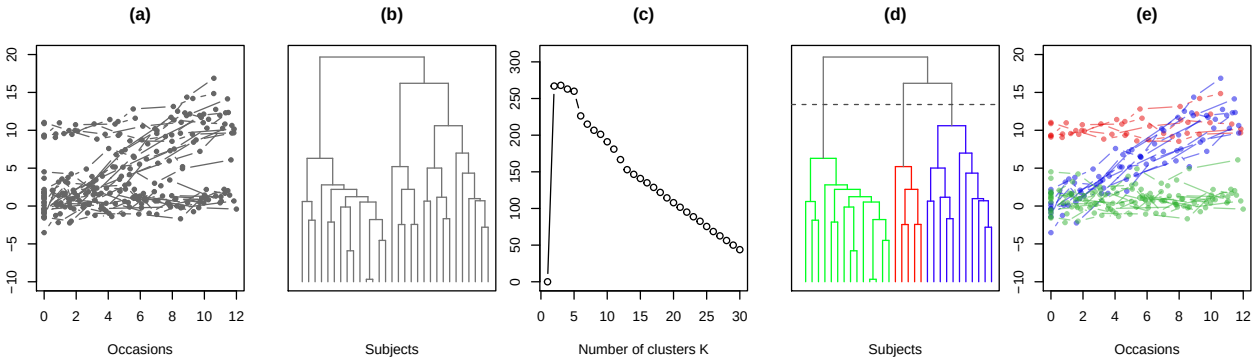

Figure 1: Steps in the clustering algorithm: (a) Original sample data. (b) Dendrogram from the hierarchical clustering algorithm. (c) Graph of Gap_b statistic. (d) Determine the number of clusters according to the Gap_b statistic. (e) Final color-coded clustering results.

2.4 The Clustering Algorithm

We propose a clustering algorithm that uses a greedy search strategy with a hierarchical merging mechanism for longitudinal data. The algorithm is structured in an agglomerative (i.e., bottom-up) manner such that the size of subclusters increases as the algorithm proceeds. The process results in more accurate estimation of the mean function for each subcluster and the dissimilarity measure between subclusters and thus enhancing the clustering performance for sparse longitudinal data situations.

The essential steps are as follows:

1. Baseline subclusters: Assuming that for the ith subject, the observed data $\{(t_{ij}, Y_{ij}) : j = 1, ..., n_i\}$ (e.g., Figure 1 (a)) allow for the fitting of a B-spline with p basis functions, we start with each subject as a baseline subcluster, fit the B-spline model, and calculate SSRs and Gap_b.
2. Bottom-up merging process: Conduct a greedy search to merge two subclusters with the minimal cost metric $\mathcal{D}$ into one cluster, and then update SSRs and Gap_b after each merge. This process is repeated in a hierarchical manner to the top level where only one cluster is left. During the process, the subclusters merged in the previous steps stay in the same cluster. The process is depicted in a dendrogram (Figure 1 (b)).
3. Determining the number of clusters: Plot the sequential Gap_b statistics against its respective number of clusters during the merging process, and determine the optimal number of clusters by the first turning point in the plot as shown in Figure 1 (c).
4. Conclusion: Summarize the final clusters from the hierarchically structured clustering results with the number determined in Step 3, as shown in the color-coded dendrogram in Figure 1 (d) and the subject-level clustering results in Figure 1 (e).

The modeling assumption in Step 1 is commonly used in the longitudinal data clustering [6, 7, 20]. When the assumption is not met, one can add a pre-processing step to form the baseline subclusters for situations where some individual subjects cannot form a baseline subcluster by themselves: Let $c^{(-)} = \{i \in c : n_i \leq p\}$ be the set of subjects whose longitudinal observations are not sufficient for the fitting of a subject-specific B-spline model with p basis functions.

Let $c^{(+)} = \{i \in c : n_i > p\}$ be the set of subjects with sufficient data to fit a subject-specific B-spline. Each subject in $c^{(+)}$ results in an initial subcluster. The idea of the pre-processing is to empty $c^{(-)}$ by combining the "most" similar subjects such that the combined one is sufficient to fit a baseline subcluster. The process can be accomplished through another greedy search strategy as described below.

1. For each $i \in c^{(-)}$, merge the data from this subject with that of subject j that satisfies $n_i + n_j > p$ to fit a B-spline with p basis functions. The mean squared residual, $d(i, j)$, is calculated for the fitted model:

$$d(i,j) = \frac{\sum_{k=1}^{n_i} \left(Y_{ik} - \hat{Y}_{ik}\right)^2 + \sum_{k=1}^{n_j} \left(Y_{jk} - \hat{Y}_{jk}\right)^2}{n_i + n_j - p}.$$

2. Identify (u, v) such that

$$(u, v) = \arg \min_{i \in c^{(-)}, j} d(i, j).$$

   If $v \in c^{(-)}$, update $c^{(-)}$ by removing subjects u and v, and update $c^{(+)}$ by adding the two subjects to the combined longitudinal observations as a new baseline subcluster in $c^{(+)}$; otherwise, update $c^{(-)}$ by removing subject u, and update $c^{(+)}$ by merging data from subject u into that from $v \in c^{(+)}$.

3. Repeat Steps 1 and 2 until $c^{(-)}$ becomes empty. Then modified subjects in $c^{(+)}$ constitute the baseline subclusters for Step 1 in the clustering algorithm.

3 Simulation Studies

We examined the operating characteristics of the proposed method and compared it against existing methods through an extensive simulation study. We first assessed method's performance in a relatively straightforward setting involving one outcome and less sparse observations [20]; this is a setting where many existing clustering methods are readily applicable. We then examined the proposed method's performance in situations of multiple outcomes and sparse observations, where the existing methods are expected to have difficulties.

Metrics of assessment. Key metrics include: (1) the empirical probability of identifying the correct number of clusters, $\Pr(\hat{K} = K)$, and (2) the overall accuracy of clustering, which is frequentlyquantified by the rate of pairwise agreement, often referred to as the Rand index (RI) [26], defined as

$$\text{RI} = \frac{TP + TN}{TP + FP + TN + FN},$$

where TP (true positive) is the number of pairs classified to the same cluster that are actually from the same cluster; TN (true negative) is the number of pairs classified into different clusters that are actually from different clusters; FP (false positive) is the number of pairs classified into the same cluster that are from different clusters; and FN (false negative) is the number of pairs classified into different clusters that are from the same cluster. Theoretically, RI takes a value between 0 and 1, but it is unlikely to be near 0 even for a random classification.

A more intuitive metric is the adjusted Rand Index (ARI) proposed by [27]:

$$\text{ARI} = \frac{RI - E[RI]}{\max(RI) - E[RI]},$$

where the expected value of RI is calculated based on the completely random classification, and $\max(RI)$ is the highest value that a classification can achieve. ARI is close to zero for a poor clustering method, but it is bounded from above by 1. ARI approaches to 1 as a clustering method becomes more accurate. ARI is applied to assess the overall accuracy of clustering whether or not the true number of clusters is used. When the correct number of clusters is used, we also explored a clustering-specific accuracy (CSA) measure to quantify the percentage of subjects correctly assigned to the right clusters.

For each setting, the simulation was repeated 100 times and we reported the average number of identified clusters, the empirical probability of correct identification of the cluster number, the average of ARI, the average of CSA, and the average of computing time.

3.1 Case 1: Single Outcome with Non-sparse Observations

For this simulation, we considered the following longitudinal model (1),

$$Y_{ij} = f_i(t_{ij}) + r_i(t_{ij}) + \varepsilon_{ij}. \tag{8}$$

We used a setting previously explored by [20], which had four clusters with the mean trajectories $f_{c_1}(t) = \cos(2\pi t), f_{c_2}(t) = 1 - 2\exp(-6t), f_{c_3}(t) = -1.5t$, and $f_{c_4}(t) = 1.5 - 1.5t$, respectively. We assumed a random error $\varepsilon_{ij} \sim \mathcal{N}(0, 0.4^2)$. However, instead of using 10 evenly spaced observation times in $[0, 1]$, we simulated $\{t_{ij}\}_{j=1}^{10}$ from a continuous uniform distribution $\mathcal{U}(0, 1)$ to add more variability in observation time. We required the interval between two adjacent observations to be larger than 0.06. We modeled the random component in model (8) with a random quadratic function

$$r_i(t) = b_{i0} + b_{i1}t + b_{i2}t^2, \tag{9}$$

where the random coefficients following a multivariate normal distribution

$$\begin{bmatrix} b_{i0} \\ b_{i1} \\ b_{i2} \end{bmatrix} \sim MVN\left(\mathbf{0}, \boldsymbol{\sigma_b}\boldsymbol{\rho_b}\boldsymbol{\sigma_b}\right) \tag{10}$$

with the following correlation coefficient matrix

$$\boldsymbol{\rho_b} = \begin{bmatrix} 1 & 0.4 & -0.3 \\ 0.4 & 1 & -0.2 \\ -0.3 & -0.2 & 1 \end{bmatrix}, \tag{11}$$

and diagonal variance matrices $\boldsymbol{\sigma_b} = \text{diag}(0.1, 0.2, 0.2)$ and $\boldsymbol{\sigma_b} = \text{diag}(0.2, 0.4, 0.4)$, respectively, to indicate low and high noise scenarios. We generated data for 100 subjects for each of the two cluster sizes: A balanced case with 25 subjects for each cluster and an unbalanced case with 5, 25, 25, and 45 for the four clusters.

Table 1: Comparison of clustering performance among different methods with 100 replications: 1. $\hat{K}$: the average of the identified cluster numbers; 2. Pr($\hat{K}$=4). The empirical probability of the correct identification for right number of clusters; 3. ARI: the average of ARI; 4. CSA: the average of CSA; 5. Computation time: The average computing time in seconds.

	N=(25,25,25,25)					**N=(5,25,25,45)**				
	$\hat{K}$	Pr($\hat{K}$=4)	ARI	CSA	Time	$\hat{K}$	Pr($\hat{K}$=4)	ARI	CSA	Time
Low Noise										
Proposed	4	1	0.994	0.998	0.79	3.96	0.96	0.994	0.999	0.72
PAM (B-splines)	1.05	0	< 0.01	-	3.74	1.04	0	< 0.01	-	14.24
KmL	4	1	0.992	0.997	35.45	2.23	0.07	0.558	0.994	35.74
GM (B-splines)	5.34	0.21	0.937	0.997	1.05	5.02	0.58	0.844	0.939	0.95
UReMix	4.32	0.69	0.969	0.999	0.57	3.04	0.04	0.916	0.875	0.49
NPG	3.86	0.83	0.95	0.999	328.97	4.06	0.94	0.968	0.985	334.22
High Noise										
Proposed	4	0.98	0.923	0.972	0.8	3.62	0.44	0.884	0.975	0.79
PAM (B-splines)	1.05	0	< 0.01	-	14.7	1.04	0	< 0.01	-	14.16
KmL	2.6	0.27	0.505	0.956	35.78	2.01	0	0.443	-	35.6
GM (B-splines)	5.26	0.22	0.917	0.989	0.94	4.96	0.45	0.817	0.937	1.01
UReMix	4.87	0.33	0.857	0.969	0.68	3.73	0.48	0.787	0.824	0.61
NPG	4.07	0.43	0.744	0.865	330.56	4.58	0.45	0.889	0.96	335.91

We compared the proposed method with five competing methods: (1) Partitioning Around Medoids (PAM) with B-splines [6, 7, 20]; (2) Gaussian Mixtures (GM) with B-splines [20]; (3) the K-means method for longitudinal data (KmL) by [28] and [29]; (4) unsupervised regression mixtures (UReMix) by [10]; and (5) non-parametric pairwise grouping (NPG) by [20]. The methods PAM and KmL are algorithm-based, whereas GM, UReMix, and NPG are model-based. Gap statistic [22] and BIC [18] were used to determine the number of clusters for PAM and GM, respectively. For KmL, the CH index was used as the recommended method among the alternatives provided in the R package `kml` for the optimal number of clusters. For UReMix, the number of mixtures was automatically determined through a robust EM algorithm [19]. The number of clusters in NPG was decided by properly tuning the hyperparameters. The results are summarized in Table 1.

Table 1 shows that the proposed method substantially outperforms the competing methods. Overall, none of the competing methods consistently yield comparable results to the proposed method in the simulated settings. PAM and GM failed to identify the correct number of clusters. KmL performed well when the cluster sizes were balanced and the noise level was low; performance deteriorated in other cases. UReMix had outstanding computational efficiency owing to the explicit updating rule of the EM algorithm. But its performance on clustering accuracy left much to be desired, especially when the cluster sizes were unbalanced. NPG showed a strength when cluster sizes were unbalanced or noise low, but its computing time was 400 times more than that of the proposed method in all tested settings. The proposed method performed well in terms of identifying the correct number of clusters and overall accuracy ARI and CSA in all settings except for situations where the cluster size is highly unbalanced and the noise level is high; in those

situations the performance was less than optimal but still better than most of the competing methods. The computational efficiency of the new method offered a great advantage when clustering large volume of longitudinal data.

3.2 Case 2: Multiple Outcomes with Sparse Observations

For this simulation, we considered a specific multivariate longitudinal model,

$$Y_{ij}^{(h)} = f_i^{(h)}(t_{ij}) + r_i(t_{ij}) + \sigma_{ih} + \varepsilon_{ijh}, \tag{12}$$

where $Y_{ij}^{(h)}$ denoted the h^{th}outcome for subject i on the j^{th} occasion, and $f_i^{(h)}(\cdot)$ the h^{th} fixed component or the h^{th} mean trajectory for the i^{th} subject. Herein, we considered five outcomes with four underlying clusters. The mean trajectories $f^{(h)}(\cdot), h = 1, ..., 5$ for each cluster were given in Table 2 and graphically illustrated in Figure 2 (A). For this setting, the clusters were mostly distinguished by $Y^{(1)}$ and $Y^{(2)}$ and less distinguished by $Y^{(3)}$ and $Y^{(4)}$. $Y^{(5)}$ is completely non-informative in identifying these clusters.

Table 2: The five cluster-specific mean trajectories used in Simulation 2

Cluster 1	Cluster 2	Cluster 3	Cluster 4
$f_{c_1}^{(1)}(t) = 8t - 0.6t^2,$	$f_{c_2}^{(1)}(t) = 20 - 6t + 0.3t^2,$	$f_{c_3}^{(1)}(t) = 0,$	$f_{c_4}^{(1)}(t) = 20,$
$f_{c_1}^{(2)}(t) = t,$	$f_{c_2}^{(2)}(t) = -t,$	$f_{c_3}^{(2)}(t) = -7t + 0.5t^2,$	$f_{c_4}^{(2)}(t) = -20 + t,$
$f_{c_1}^{(3)}(t) = -10 + 6t - 0.4t^2,$	$f_{c_2}^{(3)}(t) = -10 + 6t - 0.4t^2,$	$f_{c_3}^{(3)}(t) = 0.2t,$	$f_{c_4}^{(3)}(t) = 0.2t,$
$f_{c_1}^{(4)}(t) = -1 + t,$	$f_{c_2}^{(4)}(t) = -1 + t,$	$f_{c_3}^{(4)}(t) = -1 + t,$	$f_{c_4}^{(4)}(t) = 10 + 2t - 0.2t^2,$
$f_{c_1}^{(5)}(t) = -2t + 0.1t^2$	$f_{c_2}^{(5)}(t) = -2t + 0.1t^2$	$f_{c_3}^{(5)}(t) = -2t + 0.1t^2$	$f_{c_4}^{(5)}(t) = -2t + 0.1t^2$

The random component $r_i(\cdot)$ was modeled by following the same equation (9) as in Case 1 with $\boldsymbol{\sigma_b} = \text{diag}(2, 0.3, 0.06)$ and the same $\boldsymbol{\rho_b}$. σ_{ih} for $h = 1, ..., 5$ were the additional random effects that described the correlations among the outcomes and they were generated from the multivariate normal distribution

$$\begin{bmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \sigma_{i3} \\ \sigma_{i4} \\ \sigma_{i5} \end{bmatrix} \sim MVN \left(\mathbf{0}, \boldsymbol{\sigma} \begin{bmatrix} 1 & 0.5 & 0.3 & -0.1 & 0 \\ 0.5 & 1 & 0.2 & 0.1 & 0 \\ 0.3 & 0.2 & 1 & 0.1 & 0 \\ -0.1 & 0.1 & 0.1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \boldsymbol{\sigma} \right),$$

where $\boldsymbol{\sigma} = \text{diag}(3, 3, 3, 3, 3)$. As a pure noise, outcome 5 was set to be totally independent of the other four outcomes. The random measurement error ε_{ijh} was simulated from either a normal distribution $\mathcal{N}(0, \varepsilon^2)$ or a uniform distribution $\mathcal{U}(-\varepsilon/2, \varepsilon/2)$. To evaluate the performance of the proposed method with different noise levels, we chose $\varepsilon = 2$ or 4.

The sparse and irregular observations were simulated as follows. For subject i, we sampled the number of observations n_i from a discrete uniform distribution between 4 and 12. Then the observation times $\{t_{ij}\}_{j=1}^{n_i}$ were selected as the order statistics of the n_i random observations from $\mathcal{U}(0, 11)$ with the first observation fixed at 0 and the interval between

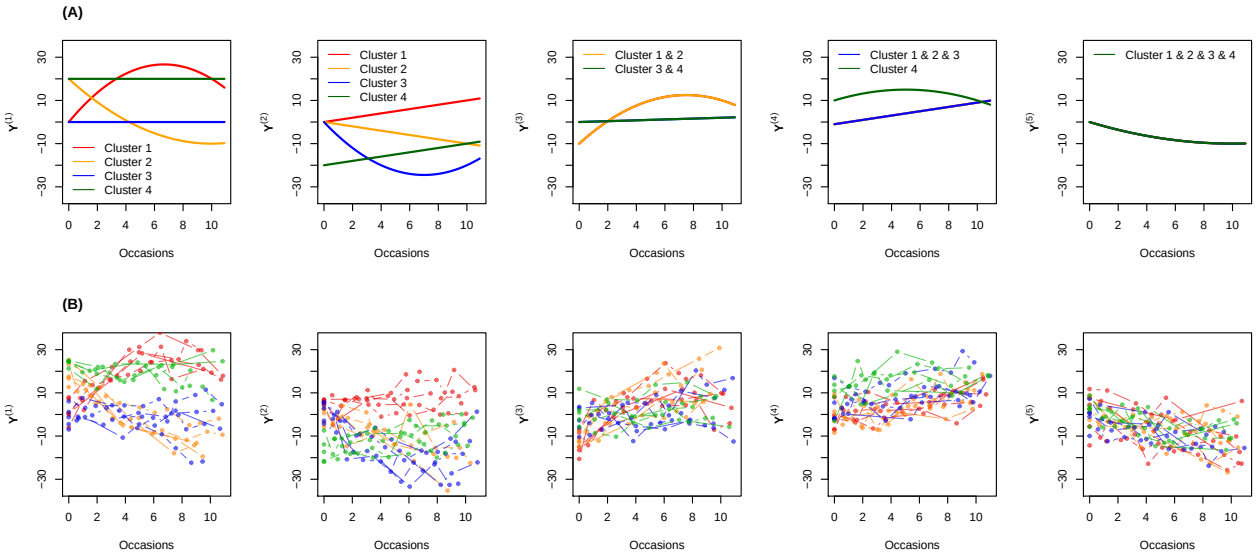

Figure 2: (A) Five mean functions used in the simulation. (B) Sample trajectories in five outcomes from each cluster with $\varepsilon_{ijh} \sim \mathcal{N}(0, 4^2)$.

two adjacent observations set to be greater than 0.5. Some sample trajectories color-coded by clusters were plotted in Figure 2 (B) for each outcome.

The performance under different cluster size combinations was also explored: In setting S0 we considered a balanced case wherein the ratio of four cluster sizes was 1:1:1:1; S1 represented a case where one cluster was extremely small, with the cluster size ratio being 1:13:13:13; and S2 represented a case where one cluster was much larger than the other clusters, with the cluster size ratio being 1:10.33:1:1. Settings S1 and S2 were designed to assess whether the clustering algorithm was influenced by an extreme small or large cluster, a situation likely to occur in studies of rare or overly abundant disease phenotypes. Furthermore, to explore the influence of sample size, simulations with total sample size of 200 and 400 were conducted. When the total sample size was 200, all four clusters had 50 subjects in S0; one cluster had 5, and the rest had 65 subjects each in S1; and one cluster had 155, and the rest had 15 subjects each in S2.

For the clustering algorithm, we used cubic B-splines [30, 31] to model the individual functional curve fitting, as described in the method section, in comparison with the existing clustering methods [6, 7, 15, 20]. Three internal knots were selected at the first, second, and third sample quartiles of all observation times to form B-spline basis functions. This resulted in a total of 7 basis functions in the B-spline curve fitting [32]. For subjects without sufficient observations, the pre-processing data steps were implemented, as described in Section 2.4.

The simulation results with respect to the identification of number of clusters based on the Gap_b statistic and CH index in all simulations settings are presented in Table 3. The CH index seemed to have a decent chance of identifying the right cluster number in Setting S0, when measurement error was uniformly distributed or normally distributed with smaller variability. However, in unbalanced settings of S1 and S2, the CH index often failed to identify the correct cluster number. The suboptimal performance of the CH index makes it a less attractive choice for determining the cluster number for longitudinal data. In contrast, the Gap_b statistic had a superior and robust performance in detecting

Table 3: Mean of $\hat{K}$ and $\Pr(\hat{K}=4)$ resulted by Gap_b statistic and CH index in 100 replicates

	Gap_b				**CH**			
	Uniform		**Normal**		**Uniform**		**Normal**	
	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=2$	$\varepsilon=4$
N=200								
S0	3.86(0.84)	3.9(0.9)	3.89(0.89)	3.97(0.91)	3.55(0.72)	3.5(0.72)	3.37(0.61)	2.18(0.01)
S1	3.64(0.66)	3.66(0.67)	3.67(0.67)	3.81(0.77)	2.5(0)	2.48(0)	2.36(0)	2.11(0)
S2	3.82(0.62)	3.62(0.57)	3.79(0.74)	3.84(0.84)	3.18(0.16)	3.36(0.15)	2.96(0.12)	2.7(0.1)
N=400								
S0	3.98(0.98)	3.99(0.99)	4.01(0.99)	4.02(0.94)	3.77(0.88)	3.65(0.82)	3.36(0.6)	2.2(0)
S1	3.81(0.81)	3.9(0.9)	3.91(0.91)	3.96(0.94)	2.45(0)	2.38(0)	2.27(0)	2.05(0)
S2	4.11(0.74)	3.94(0.81)	3.74(0.78)	3.93(0.93)	3.46(0.16)	3.01(0.07)	2.97(0.12)	2.57(0.07)

Note: The numbers in parentheses are $\Pr(\hat{K}=4)$

Table 4: Values of ARI and accuracy with number of clusters determined by Gap_b

	ARI				**CSA**			
	Uniform		**Normal**		**Uniform**		**Normal**	
	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=2$	$\varepsilon=4$	$\varepsilon=2$	$\varepsilon=4$
N=200								
S0	0.933	0.944	0.941	0.930	0.991	0.989	0.989	0.981
S1	0.952	0.954	0.953	0.946	0.991	0.984	0.985	0.969
S2	0.874	0.898	0.916	0.934	0.988	0.989	0.982	0.983
N=400								
S0	0.979	0.981	0.980	0.960	0.994	0.994	0.993	0.987
S1	0.978	0.984	0.984	0.970	0.994	0.995	0.995	0.989
S2	0.898	0.936	0.956	0.969	0.992	0.994	0.985	0.989

the right cluster number in both balanced and unbalanced settings as evidenced by the higher empirical probability $\Pr(\hat{K}=4)$, especially in the balanced setting S0. The performance improved when sample size increased from 200 to 400. The Gap_b statistic had a clear advantage over the CH index. We therefore opted to use it in the clustering algorithm.

With the number of clusters determined by the Gap_b, we present the values of ARI and CSA in Table 4. The overall accuracy of the proposed algorithm is excellent with ARI values above 0.9 in most cases, and CSA above 0.9 in all cases. As expected, increasing sample size from 200 to 400 yielded better clustering acuracy. In the data setting with $\varepsilon=2$, the average estimated importance weights W_h for the five outcomes were 0.419, 0.295, 0.097, 0.125, and 0.064 with uniformly distributed measurement error and 0.411, 0.278, 0.111, 0.126, and 0.074 with normally distributed measurement errors, respectively. When $\varepsilon=4$, the corresponding weights were 0.418, 0.286, 0.103, 0.126, and 0.067 with uniformly distributed measurement error and 0.384, 0.266, 0.124, 0.135, and 0.090 with normally distributed measurement errors, respectively. The estimated weights reflected the order of importance of the multiple outcomes in differentiating the clusters, as shown in Figure 2.

In summary, the proposed clustering algorithm delivered a more reliable performance in determining subject's membership based on longitudinal observations, in comparison with the existing methods. The numerical efficiency of the algorithm offers an added advantage. These characteristics have made the proposed algorithm an ideal clustering tool for longitudinal data, especially when big data and/or multiple outcomes are involved.

4 Real-World Applications

The method discussed in the current research is readily applicable to real data applications. To emphasize its general applicability, we present two clustering analyses using data from two clinical investigations.

4.1 The Systolic Blood Pressure Intervention Trial

The Systolic Blood Pressure Intervention Trial (SPRINT) is a randomized clinical trial aimed at reducing cardiovascular complications in people with hypertension by aggressively lowering systolic blood pressure (BP). Participants were randomly assigned to two arms: One is an intensive treatment arm where systolic BP goal was set to 120 mmHg, and the other is a standard treatment arm where the systolic BP goal remained at 140 mmHg. Therapeutic decisions on how to bring down BP were left to the treating physicians. The study tracked BP in study participants at three-month intervals approximately for up to five years. The study found the intervention had resulted in a significantly lower systolic and diastolic BP in patients received the SPRINT intervention [33]. The SPRINT data are publicly available from the National Heart, Lung, and Blood Institute, through its Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) (`https://biolincc.nhlbi.nih.gov/home/`) under signed Research Materials Distribution Agreements (RMDA).

In this research, we performed a clustering analysis of BP data in SPRINT participants. If the clustering method works, we should be able to identify two clusters, one for those with lower BP and one for those higher BP, and the cluster membership should roughly correspond to the original treatment group assignment because the trial has shown that the intervention was successful in lowering BP.

The SPRINT study had a total of $9,173$ participants, including $4,600$ in the intensive treatment arm and $4,573$ in the control arm. The participants generated $144,824$ BP measurements; each patient on average contributed 15.8 BP measures. We excluded those with only baseline BP because they did not contribute any discriminating information to the clustering. Systolic BP was used as the primary outcome so that all methods described in Section 3 could be applied. The pointwise mean longitudinal SBP patterns in the two treatment groups are shown in Figure 3 (A).

The clustering results were presented in Table 5. We used Cohen's Kappa ($\mathbf{k}$) coefficient [34] to assess the agreement between the cluster membership and the original treatment assignment. The proposed method identified two clusters ($\hat{K} = 2$), resulted in a high Kappa coefficient ($\kappa = 0.647$), and had a superior computational efficiency. The computing time on a MacBook equipped with a 2.3GHz 8-Core Intel i9 processor was 147.17 seconds to complete the clustering. The low and high systolic BP clusters respectively included $4,674$ and $4,499$ patients. In the low BP cluster, $3,828$

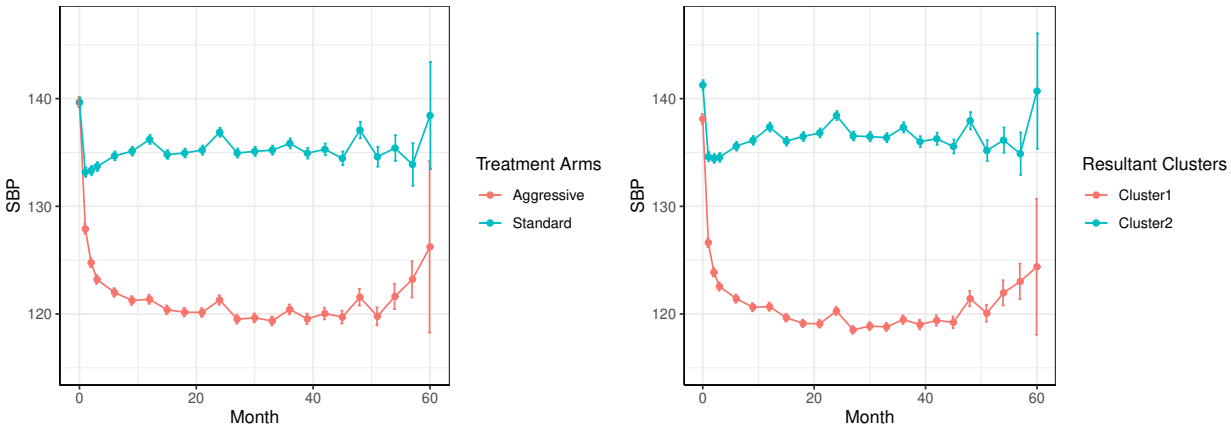

Figure 3: (a) The pointwise mean SBP trajectories by the original treatment groups. (b) The pointwise mean SBP trajectories by the clusters using the proposed algorithm. The pointwise vertical bar around the mean is the one standard deviation error bar for the mean.

Table 5: Clustering results of SPRINT data from the proposed method, PAM with B-splines, GM with B-splines, and KmL package

	$\hat{K}$	Kappa ($\mathbf{k}$)	Time (sec)
Proposed	2	0.647	147.17
PAM (B-splines)	1	0.000	66046.21
GM (B-splines)	10	0.093	191.62
KmL	2	0.202	3377.42

(81.9%) were from the intensive treatment arm; in the high BP cluster, $3,729$ (82.8%) were from the standard treatment group. However, one should not anticipate a perfect agreement between cluster membership and the original treatment assignment because the SPRINT intervention did not work for every patient, and some control patients could achieve lower BP than peers in the standard treatment arm. We presented the pointwise mean systolic BP trajectories for the two clusters in Figure 3 (B); the mean trajectories clearly resembled the treatment group BP trajectories.

The competing methods did not fare well: PAM and GM failed to identify the correct number of clusters. KmL correctly identified two clusters, but the Kappa value ($\kappa = 0.202$) was disappointing, and the method consumed 20 times more computing time. We were not able to ascertain results from UReMix and NPG because they were not equipped to deal with a dataset of this size. We had to stop the program for the two latter methods because of memory overflow and excessive computing time.

4.2 The PREDICT-HD Study

In a second application, we used clustering analysis to investigate Huntington's disease (HD) progression phenotypes. HD is a neurodegenerative disease caused by the trinucleotide cytosine-adenine-guanine (CAG) in the first exon of the *Huntington (HTT)* gene [35]. The disease debilitates motor function in the afflicted, often accompanied by accelerated impairment cognitive functions [36, 37]. Diagnosis is typically made by using the motor subscale of the Unified HD Rating Scale (UHDRS) [38, 39]. The PREDICT-HD is a 12-year observational study conducted between 2002 and 2014

on individuals with the HTT gene and shown early signs of motor dysfunction but did not reach functional diagnostic criteria. The study was conducted in 33 sites across six countries (USA, Canada, Germany, Australia, Spain, and UK). Large volumes of data on neuroimaging, motor, cognitive, and psychiatric assessments were collected for predicting the onset of HD [40].

In this research, we studied phenotypes in HD progression in motor and cognitive impairment and how the progression phenotypes affected disease onset by using the PREDICT-HD data. The data are available publicly in the NIH Database for Genotypes and Phenotypes (dbGap) (ninds-dac@mail.nih.gov).

We selected five motor and cognitive measures that are commonly used for tracking HD progression: The total motor score (TMS) [39, 41], the Symbol Digit Modalities Test (SDMT) [42], and the three Stroop Color Word Tests, i.e., Stroop Color test (stroopco), Stroop Word Test (stroopwo), and Stroop Color-Word Inference Test (stroopin)[43, 44]. All five measures are on numerical scales, with a smaller value indicating more impairment, except for TMS, where a larger value corresponds to greater motor impairment. For $1,006$ participants in PREDICT-HD, assessments were made annually. The average number of observations per participant was 5.62, with a minimum of 2 and a maximum of 13 observations. The application represents a typical example of multiple-outcome longitudinal data with irregular and sparse observations, for which this method is developed.

We implemented the analysis as previously described. We fitted cubic B-spline curves using internal knots selected at the three quartiles of the total observation time. The Gap_b statistic indicated three clusters. For narrative convenience, we referred to them as Clusters 1, 2, and 3, respectively covering 317 (32.2%), 332 (33.7%), and 336 (34.1%) participants. We present the mean B-spline functional curves in Figure 4. The figure showed SDMT and the three Stroop tests were clearly separated at baseline, with subjects in Cluster 1 being more impaired throughout the observational window, especially near the end when impairment accelerated. For TMS, the analysis showed that cluster 1 had a much higher rate of increase, suggesting a more rapid deterioration. In contrast, Cluster 3 had progressed slowly in both motor and cognitive declines. A closer examination of the data showed that only 25 participants (7.4%) in Cluster 3 had HD diagnosis with a median time to diagnosis beyond 12 years after enrollment. The numbers of HD diagnoses for Clusters 2 and 1 were 64 (19.3%) and 149 (47.0%), with the median times to HD diagnosis of 11.1 and 5.6 years from study enrollment, respectively. The estimated survival functions were presented in Figure 5 (a), which confirmed that participants in the three clusters had significantly different time-to-HD-diagnosis distributions (p-value < 0.001) per log-rank test [45].

Previously, [46] explored using the CAG trinucleotide to quantify the HD genetic burden. They proposed a CAG-Age Product scale, hereby referred to as the CAP score $\text{CAP} = \text{AGE} \times (\text{CAG} - 33.7)$, which was found to strongly predict HD onset and has since been used to characterize prodromal HD risk and used for disease screening [47]. In the current analysis, CAP was predictive of disease progression – the three clusters had significantly different CAP values (p-value < 0.001). But as a static metric assessed at baseline, CAP does not fully capture the disease progress. A large overlap in CAP values was observed among the three clusters, as shown in Figure 5 (b). We performed an additional Cox

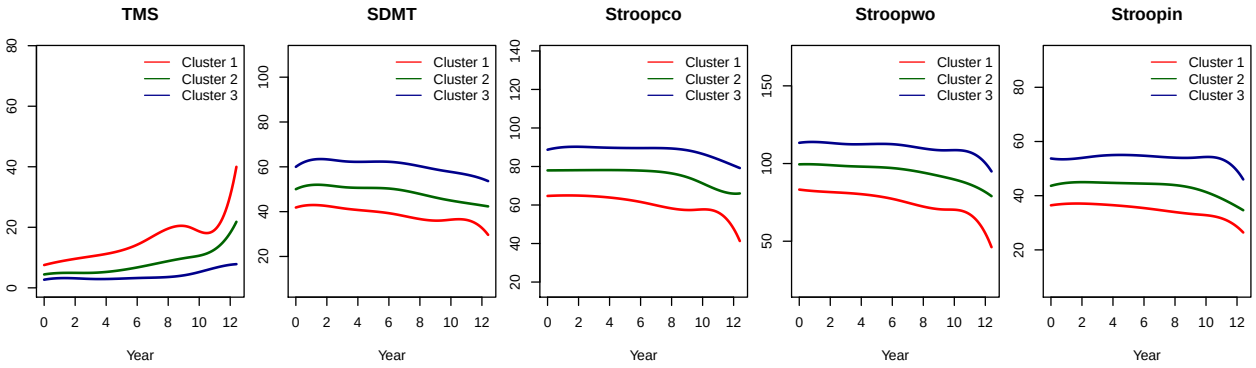

Figure 4: The estimated cluster-specific mean trajectories.

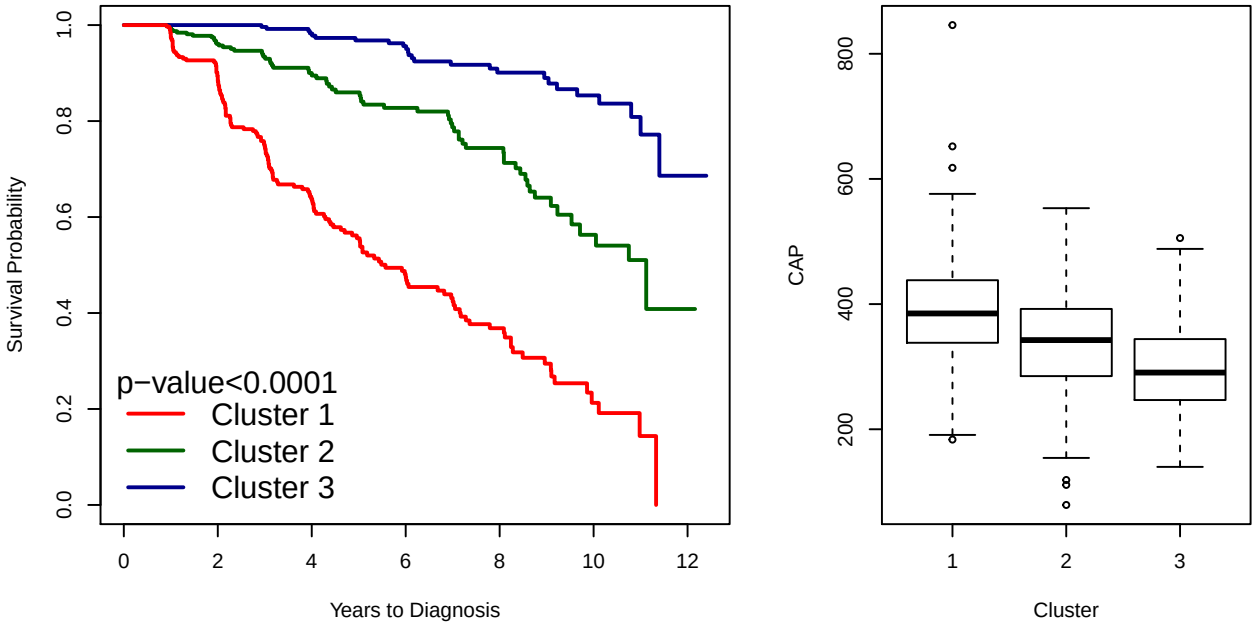

Figure 5: (a) Kaplan-Meier curves for time to HD diagnosis, according to three detected clusters. (b) Box plot for CAP by cluster

proportional hazards analysis with both CAP and cluster labels as covariates. The result showed that after adjusting for CAP, HD progression clusters remain significant (p-value < 0.001), and thus confirming the added value of clustering analysis. Clustering analysis has made targeted interventions possible – by identifying individuals at higher risk for HD progression, valuable resources, and early medical interventions can be focused on individuals with heightened susceptibility. The analysis may also aid the discovery of new biomarkers by linking them directly to HD progression [48].

5 Discussion

This paper presents a new algorithm-based clustering method for longitudinal data, including those with irregular and sparse assessments. The method incorporates functional curve fitting to alleviate the influences of noisy and sparse data observation. The algorithm takes an agglomerative hierarchical approach that merges subjects and subclusters in

a bottom-up fashion. At the core of the algorithm is a metric that quantifies the cost of each merging. The metric is calculated by comparing the sum of squared residuals of the separate and combined models; thus, it can be viewed as a metric of dissimilarity between two subclusters. Since the calculation is based on the sum of squared residuals, it can be used in a broad class of parametric and nonparametric models, including the B-spline regression models we chose in this research. With traditional linear models, the metric has a well-grounded foundation in classical statistical theory. An essential advantage of the method is that the readily calculated sums of squared residuals have significantly reduced the computational burden. The least-square-based cost metric has drastically simplified the problem of comparing longitudinal models. We were surprised that such an approach had not been previously explored because it clearly can be adapted for many different models, thus representing a more flexible and extendable strategy for clustering longitudinally observed data.

The strongest support for the method perhaps comes from the numerical studies, highlighting the many advantages that set this method apart from the existing ones. Major appeals include the accurate identification of cluster numbers, clustering accuracy, and computational efficiency, in addition to its ability to handle sparse and irregular data, as well as multiple outcomes. Notably, the advantages appear to grow with sample size, as demonstrated in the simulation studies. In many practical analyses, the sample size is a double-edged sword. While larger data sets lend more information to the analysis, very large datasets tend to present greater challenges to data processing. The computational burden of running iterative and computationally heavy algorithms in larger datasets has frustrated many analysts. For a dataset of size N, agglomerative clustering algorithms typically have a time complexity of $\mathcal{O}(N^2)$ and require $\Omega(N^2)$ memory; these could make them too slow for medium size analyses. Parallel computing could help alleviate the burden by splitting the original data into multiple subsets, each of size n, and then applying the algorithm to each of them in a parallel fashion. The algorithm stops when d subclusters are resulted instead of continuing the process until only one cluster remains. As long as it is larger than the actual number of clusters, no specific requirement for this d is needed. The resultant subclusters from each subset are then combined. This split-and-pool procedure can be continuously applied multiple times until the number of remaining subclusters reaches a manageable scale. Finally, the algorithm is applied to keep merging subclusters until only one cluster is left. This algorithm helps reduce the time complexity to $\mathcal{O}((n+d)N)$. In general, larger n and d improve the clustering performance but slow the algorithm. Our analysis of the SPRINT data has confirmed the practical feasibility in moderate-to-large datasets.

From a practical data analytical perspective, the proposed method poses few restrictions. Time trajectories are assumed to be smooth functions and can be adequately depicted by the spline models that the analyst chooses. Additionally, the Gap_b statistic is not defined for $K = 1$, thus precluding the possibility of testing whether heterogeneity exists in longitudinally observed data. We considered the data heterogeneity to be *a priori*. When in doubt, this could be tested with goodness-of-fit statistics. Finally, the algorithm as it currently stands does not work for discrete data where new similarity metrics are needed. Notwithstanding these limitations, we put forward a robust clustering algorithm based on a newly developed merging cost metric designed to work for continuous outcomes.

Appendix

In this section, we list the propositions used in section 2.

Proposition 1. *For any* $\mathcal{C}_u, \mathcal{C}_v$, $\mathcal{D}(\mathcal{C}_u, \mathcal{C}_v) \geq 0$.

Proof: To proof Proposition 1, we only need to justify the nonnegativity of the numerator since the denominator is always positive.

$$\begin{aligned}
&\text{SSR}(\mathcal{C}_{u,v}) - \text{SSR}(\mathcal{C}_u) - \text{SSR}(\mathcal{C}_v) \\
&\quad = \sum_{i \in \{c_u, c_v\}} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2 - \sum_{i \in c_u} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij,c_u})^2 - \sum_{i \in c_v} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij,c_v})^2 \\
&\quad \geq \sum_{i \in \{c_u, c_v\}} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2 - \sum_{i \in c_u} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2 - \sum_{i \in c_v} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2 \\
&\quad = 0,
\end{aligned}$$

where $\hat{Y}_{ij,c_u}$ and $\hat{Y}_{ij,c_v}$ are the cluster-specific ordinary least-square (OLS) estimator. The inequality naturally comes from the fact that OLS estimator yields the smallest SSR. So replacing the cluster-specific OLS estimators $\hat{Y}_{ij,c_u}$ and $\hat{Y}_{ij,c_v}$ by overall OLS estimator $\hat{Y}_{ij}$ will increase the sum of residuals, i.e. $\sum_{i \in c_k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij,c_k})^2 \leq \sum_{i \in c_k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2, k = u, v$. The equality holds when $\mathcal{C}_u = \mathcal{C}_v$.

Proposition 2. *Suppose the longitudinal outcome* Y_{ij} *is governed by (7). For two subclusters* $\mathcal{C}_1$ *and* $\mathcal{C}_2$*, where* $\boldsymbol{\beta}_i = \boldsymbol{\beta}_{c_1}$ *for* $i \in c_1$*, and* $\boldsymbol{\beta}_j = \boldsymbol{\beta}_{c_2}$ *for* $j \in c_2$*, it can be shown that*

$$E[\text{SSR}(\mathcal{C}_{1,2}) - \text{SSR}(\mathcal{C}_1) - \text{SSR}(\mathcal{C}_2) \mid \boldsymbol{\beta}_{c_1} = \boldsymbol{\beta}_{c_2}] = p\sigma^2.$$

Proof: Denote the data in $\mathcal{C}_1$ and $\mathcal{C}_2$ are $(\mathbf{X}_{c_1}, \mathbf{Y}_{c_1})$ and $(\mathbf{X}_{c_2}, \mathbf{Y}_{c_2})$, respectively. For simplicity, denote $\boldsymbol{\beta}_{c_1} = \boldsymbol{\beta}_{c_2} = \boldsymbol{\beta}$. Then for the combined data $\mathcal{C}_{1,2}$, $\boldsymbol{\beta}_{c_{1,2}} = \boldsymbol{\beta}$. Let $\mathbf{Y}_{c_{1,2}} = \begin{bmatrix} \mathbf{Y}_{c_1} \\ \mathbf{Y}_{c_2} \end{bmatrix}$ and $\mathbf{X}_{c_{1,2}} = \begin{bmatrix} \mathbf{X}_{c_1} \\ \mathbf{X}_{c_2} \end{bmatrix}$, we have

$$\begin{aligned}
&E[\mathbf{Y}_{c_1}] = \boldsymbol{\mu}_{c_1} = \mathbf{X}_{c_1}\boldsymbol{\beta}, \quad Var[\mathbf{Y}_{c_1}] = \sigma^2 \mathbf{I}_{|c_1|} \\
&E[\mathbf{Y}_{c_2}] = \boldsymbol{\mu}_{c_2} = \mathbf{X}_{c_2}\boldsymbol{\beta}, \quad Var[\mathbf{Y}_{c_2}] = \sigma^2 \mathbf{I}_{|c_2|}; \\
&E[\mathbf{Y}_{c_{1,2}}] = \boldsymbol{\mu}_{c_{1,2}} = \mathbf{X}_{c_{1,2}}\boldsymbol{\beta}, \quad Var[\mathbf{Y}_{c_{1,2}}] = \sigma^2 \mathbf{I}_{|c_{1,2}|}.
\end{aligned}$$

Use $\mathcal{C}_k$ to denote either $\mathcal{C}_{1,2}$, $\mathcal{C}_1$, or $\mathcal{C}_2$, it can be shown by Lemma 1 that the expectation of $\text{SSR}(\mathcal{C}_k)$ is

$$
\begin{aligned}
E[\text{SSR}(\mathcal{C}_k)] =& E[\mathbf{Y}_{c_k}^T(\mathbf{I}-\mathbf{H}_{c_k})\mathbf{Y}_{c_k}] \\
=& \boldsymbol{\mu}_{c_k}^T(\mathbf{I}-\mathbf{H}_{c_k})\boldsymbol{\mu}_{c_k} + \text{tr}((\mathbf{I}-\mathbf{H}_{c_k})\sigma^2) \\
=& \boldsymbol{\beta}^T\mathbf{X}_{\mathcal{C}_k}^T\mathbf{X}_{\mathcal{C}_k}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}_{\mathcal{C}_k}^T\mathbf{X}_{\mathcal{C}_k}(\mathbf{X}_{\mathcal{C}_k}^T\mathbf{X}_{\mathcal{C}_k})^{-1}\mathbf{X}_{\mathcal{C}_k}^T\mathbf{X}_{\mathcal{C}_k}\boldsymbol{\beta} + \sigma^2\text{tr}(\mathbf{I}_{|c_k|}) - \sigma^2\text{tr}(\mathbf{H}_{c_k}) \\
=& \sigma^2|c_k| - \sigma^2 p, \qquad (13)
\end{aligned}
$$

where $\mathbf{H}_{c_k} = \mathbf{X}_{c_k}(\mathbf{X}_{c_k}^T\mathbf{X}_{c_k})^{-1}\mathbf{X}_{c_k}^T$. In the last equality, we use the fact that the trace of a hat matrix–an idempotent matrix–is its rank, p. Hence, it immediately follows that

$$
\begin{aligned}
E[\mathcal{D}(\mathcal{C}_1,\mathcal{C}_2)] =& E[\text{SSR}(\mathcal{C}_{1,2}) - \text{SSR}(\mathcal{C}_1) - \text{SSR}(\mathcal{C}_2)] \\
=& \sigma^2|c_{1,2}| - \sigma^2 p - (\sigma^2|c_1| - \sigma^2 p) - (\sigma^2|c_2| - \sigma^2 p) \\
=& \sigma^2 p.
\end{aligned}
$$

Lemma 1 (Hogg, McKean, and Craig 2005). *Let* $\mathbf{Y}$ *be a n-dimensional random vector and* $\mathbf{A}$ *be a constant* $n \times n$ *symmetric matrix. If* $E(\mathbf{Y}) = \boldsymbol{\mu}$ *and Cov*$(\mathbf{Y}) = \boldsymbol{\Sigma}$*, then*

$$E(\mathbf{Y^T A Y}) = \text{tr}(\mathbf{A\Sigma}) + \boldsymbol{\mu^T A \mu}$$

where tr() denotes the trace of a matrix. It comes from the expectation of a quadratic form.

References

[1] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1*, pages 281–297. University of California Press, Berkeley, CA, USA, 1967.

[2] Robin Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973.

[3] Daniel Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

[5] Christophe Genolini and Bruno Falissard. Kml: k-means for longitudinal data. *Computational Statistics*, 25(2):317–328, 2010.

[6] Christophe Abraham, Pierre-André Cornillon, ERIC Matzner-Løber, and Nicolas Molinari. Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics*, 30(3):581–595, 2003.

[7] Luis Angel Garcia-Escudero and Alfonso Gordaliza. A proposal for robust curve clustering. *Journal of classification*, 22(2):185–201, 2005.

[8] Nicoleta Serban and Larry Wasserman. Cats: clustering after transformation and smoothing. *Journal of the American Statistical Association*, 100(471):990–999, 2005.

[9] Bobby L Jones, Daniel S Nagin, and Kathryn Roeder. A sas procedure based on mixture models for estimating developmental trajectories. *Sociological methods & research*, 29(3):374–393, 2001.

[10] Faicel Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 86(12):2308–2334, 2016.

[11] Faicel Chamroukhi and Hien D. Nguyen. Model-based clustering and classification of functional data. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1298, 2019.

[12] Gareth M James and Catherine A Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408, 2003.

[13] Yihui Luan and Hongzhe Li. Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19(4):474–482, 2003.

[14] Madison Giacofci, Sophie Lambert-Lacroix, Guillemette Marot, and Franck Picard. Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, 69(1):31–40, 2013.

[15] Norma Coffey, John Hinde, and Emma Holian. Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics & Data Analysis*, 71:14–29, 2014.

[16] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

[17] Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

[18] Gideon Schwarz et al. Estimating the dimension of a model. *Annals of statistics*, 6(2):461–464, 1978.

[19] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, 2012.

[20] Xiaolu Zhu, Annie Qu, et al. Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12(1):171–193, 2018.

[21] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[22] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[23] Gregory C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

[24] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[25] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[26] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[27] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[28] Christophe Genolini and Bruno Falissard. Kml: A package to cluster longitudinal data. *Computer methods and programs in biomedicine*, 104(3):e112–e121, 2011.

[29] Christophe Genolini, Xavier Alacoque, Mariane Sentenac, Catherine Arnaud, et al. kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, 65(4):1–34, 2015.

[30] Isaac Jacob Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. part b. on the problem of osculatory interpolation. a second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(2):112–141, 1946.

[31] Larry Schumaker. *Spline Functions: Basic Theory*. Cambridge Mathematical Library. Cambridge University Press, 3 edition, 2007.

[32] David Ruppert. Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, 11(4):735–757, 2002.

[33] The-SPRINT-Research-Group. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine*, 373(22):2103–2116, 2015. PMID: 26551272.

[34] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[35] Marcy E MacDonald, Christine M Ambrose, Mabel P Duyao, Richard H Myers, Carol Lin, Lakshmi Srinidhi, Glenn Barnes, Sherryl A Taylor, Marianne James, Nicolet Groot, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell*, 72(6):971–983, 1993.

[36] Jeffrey D Long, Jane S Paulsen, PREDICT-HD Investigators, and Coordinators of the Huntington Study Group. Multivariate prediction of motor diagnosis in huntington's disease: 12 years of predict-hd. *Movement Disorders*, 30(12):1664–1672, 2015.

[37] Francis O Walker. Huntington's disease. *The Lancet*, 369(9557):218–228, 2007.

[38] Ira Shoulson and Stanley Fahn. Huntington disease: clinical care and evaluation. *Neurology*, 29(1):1–1, 1979.

[39] Karl Kieburtz, John B Penney, Peter Corno, Neal Ranen, Ira Shoulson, Andrew Feigin, Davi Abwender, J Timothy Greenarnyre, Donald Higgins, Frederick J Marshall, et al. Unified huntington's disease rating scale: reliability and consistency. *Neurology*, 11(2):136–142, 2001.

[40] Jane S Paulsen, Jeffrey D Long, Christopher A Ross, Deborah L Harrington, Cheryl J Erwin, Janet K Williams, Holly James Westervelt, Hans J Johnson, Elizabeth H Aylward, Ying Zhang, et al. Prediction of manifest huntington's disease with clinical and imaging measures: a prospective observational study. *The Lancet Neurology*, 13(12):1193–1201, 2014.

[41] Penelope Hogarth, Elise Kayson, Karl Kieburtz, Karen Marder, David Oakes, Diana Rosas, Ira Shoulson, Nancy S Wexler, Anne B Young, Hongwei Zhao, et al. Interrater agreement in the assessment of motor manifestations of huntington's disease. *Movement disorders*, 20(3):293–297, 2005.

[42] Aaron Smith. *Symbol digit modalities test*. Western Psychological Services Los Angeles, 1973.

[43] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

[44] Charles J Golden. *Stroop color and word test: cat. no. 30150M; a manual for clinical and experimental uses*. Stoelting, 1978.

[45] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, 135(2):185–198, 1972.

[46] Ying Zhang, Jeffrey D Long, James A Mills, John H Warner, Wenjing Lu, Jane S Paulsen, PREDICT-HD Investigators, and Coordinators of the Huntington Study Group. Indexing disease progression at study entry

with individuals at-risk for huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 156(7):751–763, 2011.

[47] Filipe B Rodrigues, Lori Quinn, and Edward J Wild. Huntington's disease clinical trials corner: January 2019. *Journal of Huntington's disease*, 8(1):115–125, 2019.

[48] Jane S Paulsen, Spencer Lourens, Karl Kieburtz, and Ying Zhang. Sample enrichment for clinical trials to show delay of onset in huntington disease. *Movement Disorders*, 34(2):274–280, 2019.