scientific reports

Check for updates

OPEN

# Analysis of experiments with high frequency time series responses and the implications for power and sample size

Brian Rafor[1], Iris Ivy Gauran[2], Hernando Ombao[2], Joseph Ryan Lansangan[1] & Erniel Barrios[3]✉

**Given more accessible non-invasive measuring devices, experimental response can now be observed as high-dimensional and high-frequency time series. Amidst the complex dependence structure in the data analysis, sample size determination and power analysis remain to be the key thematic focus of statistical inference. The issue is confounded with the complexity of time lag structure and phase shift usually observed in a non-uniform but normal process typically present in medical imaging data. To address these issues in case-control studies, responses can be analyzed to obtain evidence of group differences through time series clustering based on dynamic time warping. The warping of multiple time series provides a flexible distance measure robust to time point concurrence. Time series clustering partitions experimental units into groups, enabling the computation of distances to measure effect size through sum of squares of pairwise distances in warped time series.Time series clustering provides an alternative to analysis of variance when experimental responses are high-frequency time series data. Kernel regression is formulated to link sample size, effect size, power of the test, and level of significance accounting for the structure of the data generating process of the time series responses. This provides a strategy for clinicians to optimize the power of the test that can be achieved with a minimal sample size for this experimental setup. Time series clustering method is able to differentiate case and control groups in the simulated data and in the ADHD-200 fMRI dataset. The distance measured between two or more groups of time series can be used to determine sample size for a target power.**

**Keywords** Multiple time series, Time series clustering, Dynamic time warping, Power analysis, Functional magnetic resonance imaging

Brain functionality is driven by a complex symphony of neural activity, with its spatially distributed, yet functionally interconnected regions continuously sharing information through efficient networks[1]. Understanding and measuring these brain dynamics has advanced significantly in recent decades, as the landscape of experimental psychology and neuroscience research has been transformed by modern high-frequency measurement devices. Various forms of advanced brain imaging technologies and wearable biosensors have become valuable instruments for collecting continuous physiological data. Within this technological array, functional magnetic resonance imaging (fMRI) has revolutionized our understanding of brain function by non-invasively mapping neural activity without contrast agents[2], serving multiple crucial objectives from localizing activated brain regions during specific tasks to determining distributed functional networks[3]. Among many fMRI experimental paradigms, resting state fMRI (rs-fMRI) has emerged a powerful tool for advancing precision medicine[4], offering unprecedented insights into disrupted brain networks underlying neuropsychiatric disorders without requiring patients to perform any experimental tasks[5]. Capitalizing on this potential, researchers have systematically employed rs-fMRI in case-control studies to compare brain function between patients and healthy controls across various psychiatric disorders classified in diagnostic manuals such as DSM-5 and ICD-11[6]. However, translating these advances into clinical practice requires careful consideration of the balance between high cost of imaging and the need for adequate sample sizes to ensure both precise results and cost-efficient data

[1]School of Statistics, University of the Philippines, Diliman, Quezon City 1101, Philippines. [2]Statistics Program, CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955, Kingdom of Saudi Arabia. [3]School of Business, Monash University Malaysia, Selangor 47500, Malaysia. ✉email: erniel.barrios@monash.edu

collection[7]. Addressing this critical methodological challenge, this study develops a novel approach for sample size calculation and power analysis for distance-based multivariate fMRI time series, focusing specifically on group-level time series characteristics in case-control rs-fMRI studies.

Despite these neuroimaging technological advances, recent analyses have revealed methodological problems that are particularly evident in studies of disorders such as Attention-Deficit/Hyperactivity Disorder (ADHD). Classification accuracy in these studies show alarming variability, ranging from chance-level to near-perfect predictions[8]. A comprehensive review of 69 ADHD neuroimaging studies by[9] revealed that this inconsistency stems from methodological issues. In particular, they found a negative correlation between accuracy and sample size, suggesting that smaller samples may introduce bias and artificially inflate reported accuracies. Their analysis concluded that the remarkably high classification accuracies reported in some ADHD neuroimaging studies likely result from both circular analysis and inadequate sample sizes. Thus, while neuroimaging techniques have grown increasingly sophisticated, the fundamental yet crucial aspect of determining appropriate sample sizes remains inadequately addressed.

On the other hand, power calculations are often underrepresented and underutilized, despite their critical importance in determining the appropriate sample size to detect hypothesized effects[10]. No standardized best practices for power analysis in fMRI studies have emerged[11]. The conventional method for calculating sample size, based on linear models, is still widely used despite its limitations in reproducibility and reusability. These methods are sensitive to the high dimensionality of neuroimaging data and are often limited to specific study designs[11]. As a result, these methods cannot easily be applied to related studies, even when there are minor changes in the phenotype or the subject inclusion criteria.

The convergence of these challenges makes it imperative to develop more robust methodologies for sample size determination and power analysis. The field urgently needs more sophisticated approaches that can properly account for both the temporal and spatial complexities of fMRI data while optimizing resource utilization. Without proper methods to account for these various sources of variance in sample size calculations, researchers risk either conducting underpowered studies that fail to detect true effects or overestimating required sample sizes, leading to wasteful resource allocation. This methodological crisis is further exacerbated by the increasing complexity of research designs and the growing demand for reliable biomarkers in clinical practice.

The contributions of our work are threefold. First, we introduce a novel framework for sample size calculation specifically designed for testing the distances (or discrepancies) between multivariate fMRI time series using dynamic time warping (DTW). This framework aims to equip clinicians and researchers with practical tools to determine appropriate sample sizes for experiments with multivariate time series responses, ensuring sufficient statistical power. Our proposed statistical test based on DTW distances effectively addresses challenges such as temporal misalignment, variable sequence lengths, and non-linear distortions in the data. This approach is particularly valuable in clinical research, where non-linear methods such as DTW can uncover functional connectivity patterns related to behavior and symptomatology[12]. Second, we explore the properties of DTW distance-based tests, evaluating their power, Type I error rate, and robustness against noise and variability in fMRI time series data. Our analysis includes a detailed examination of the assumptions and limitations of DTW in various time series structures. Third, we derive theoretical formulas for calculating sample sizes based on different levels of statistical power and significance thresholds. These formulas incorporate key factors such as expected DTW distances, variance, and dependence both within a time series and between components of the multivariate time series. Our method offers a flexible and generalizable approach to sample size calculation that can be adapted to similar studies, reducing implementation complexity and cost while improving efficiency and reusability. Ultimately, our goal is to provide a robust method for determining the required sample size in real-world applications.

Our work has significant implications for the neuroimaging community. The proposed framework enables more efficient resource allocation while maintaining statistical rigor, particularly valuable for clinical applications where accurate detection of group differences is crucial for diagnostic and treatment purposes. From a methodological perspective, our work establishes a foundation for future developments in the analysis of high-dimensional and high-frequency time series data, demonstrating how traditional statistical concepts can be extended to address contemporary challenges in complex data analysis. This approach may serve as a template for developing similar methods in other fields where high-dimensional time series data are prevalent. Ultimately, these advancements contribute to the broader goal of enhancing our understanding of brain function and dysfunction through more robust and well-designed neuroimaging research, bringing us closer to the realization of precision medicine in neurological and psychiatric care.

There are some limitations on the kind of high-frequency time series the clustering method will be useful. DTW assumes that some patterns (e.g., dependence structure like autoregression) are contained in the data. If the time series is a white noise or possesses a complex form of nonlinearity like those in threshold autoregression, it will be challenging for DTW to measure cluster distances. However, the distance measure proposed to quantify the treatment effect will still work even with conditional heteroskedasticity present. Furthermore, the method will not work if the time series is count since the measure of cluster distances might be at the boundaries, either it is very large or nearly zero. Thus, it is imperative to consider the real-world application and the features of the real dataset when deciding to apply the method. For instance, there are research findings that the clustering algorithm has less influence than the distance measure on some classification problems[13] and that when the complexity of the time series structure significantly distinguishes one from another, other distance methods such as the Complexity Invariance Distance (CID) maybe more effective[14] .

The rest of the paper is organized as follows. In section "Dynamic time warping (DTW)", we discuss existing methods in time series clustering. The details of the proposed sample size calculation are presented in section "Methodology". We evaluate the performance of the proposed method through simulation studies and real-world datasets, demonstrating the accuracy and practicality of the method in various fields in section "Results

and discussion". Finally, we summarize our contributions including our key findings, limitations, and future directions in section "Conclusions and future research".

## Existing methods in time series clustering

Although advanced high-frequency measurement equipment provides detailed insights into physiological and behavioral patterns, it also presents analytical challenges when working with multiple participants and measurement types. Time series clustering offers a powerful solution by grouping similar temporal patterns and identifying underlying structures that may reflect disorder-specific characteristics[15]. This approach not only helps to understand the synchrony and similarity across multiple time series but also uncovers underlying patterns that can be leveraged for pattern discovery, information retrieval, and outlier detection[16], thus providing a systematic approach to detect meaningful differences between cases and controls.

Outside the neuroimaging field, more complex forms of multiple high-frequency time series are considered. As an example, cyclostationary (processes that are not necessarily periodic but vary periodically over time)[17] is assumed in clustering spectral densities[18]. A fuzzy clustering method was also proposed[19], assuming that the time series is a fractional Brownian motion.

### Dynamic time warping (DTW)

Central to the effectiveness of time series clustering is the quantification of similarity between temporal patterns. The choice of similarity metric can significantly impact the performance of clustering algorithms. For two time series $\mathrm{x}, \mathrm{y}$ taking values in a feature space $\mathcal{F}$, comparing $\mathrm{x}, \mathrm{y} \in \mathcal{F}$ requires a local distance measure $D(\mathrm{x}, \mathrm{y}) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^+$. This measure $D(\mathrm{x}, \mathrm{y})$ assigns small values to similar sequence elements and large values to dissimilar ones. One of the most popular metrics is the Minkowski distance of order $p$[20], defined as

$$D_p(\mathrm{x}, \mathrm{y}) = \|\mathrm{x} - \mathrm{y}\|_p = \left( \sum_{i=1}^{N} |x_i - y_i|^p \right)^{1/p},$$

for input $\mathrm{x}, \mathrm{y} \in \mathcal{F} = \mathbb{R}^N$ and $p \in \mathbb{Z}$. In particular, when $p = 1$, it is called Manhattan distance, and when $p = 2$ , it is the Euclidean distance. Another popular measure is the Mahalanobis distance, which can be viewed as the scaled Euclidean distance, and is defined as $D(\mathrm{x}, \mathrm{y}) = ((\mathrm{x} - \mathrm{y})^\top \mathbf{\Sigma}^{-1} (\mathrm{x} - \mathrm{y}))^{1/2}$, where $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$ is the covariance matrix.

Although these geometrically intuitive distance measures provide a foundation for quantifying similarity between time series, they show significant limitations when utilized on multivariate time series that exhibit phase perturbations (such as phase shifts and time warps), thus necessitating more sophisticated approaches for temporal pattern analysis. This limitation is particularly evident in sequence data analysis, where both Minkowski and Mahalanobis distances fail to capture true temporal similarities between signals. Dynamic Time Warping (DTW)[21–23] resolves these challenges. Unlike conventional distance measures, DTW is invariant to temporal distortions, including time-axis shifting, scaling, and Doppler effects. This robustness to signal warping has established DTW as an optimal metric for pattern matching applications[24]. Consider, for example, two signals recorded at different sampling rates, where one signal essentially represents a compressed version of the other. In this scenario, traditional Euclidean distance calculations would indicate substantial dissimilarity, providing misleading results. In contrast, DTW recognizes the inherent relationship between these signals by accommodating temporal scaling, resulting in a very small distance between them. More importantly, a key feature of DTW is that it not only measures sequence similarity, but also maps out how patterns "align" in time, with this alignment information often proving more valuable than the distance measure itself[25]. Moreover, DTW serves as an effective feature extraction tool, enabling the identification of predefined patterns within temporal data. This capability has proven particularly valuable in classification tasks, where researchers have successfully leveraged DTW-based pattern recognition to categorize temporal data into meaningful groups[26].

To formally define the DTW framework, consider two time series with lengths $M, N \in \mathbb{N}$ denoted by $\mathrm{x} = (x_1, x_2, \ldots, x_N)$ and $\mathrm{y} = (y_1, y_2, \ldots, y_M)$. These sequences can represent either discrete values or vertices of continuous curves, with the fundamental requirement being uniform temporal sampling. Although non-uniform sampling scenarios do arise in practice, they can be effectively handled through appropriate resampling techniques. At its foundation, DTW employs a local cost matrix that captures the fundamental relationships between sequence elements. This matrix, denoted as $\mathrm{C} \in \mathbb{R}^{N \times M}$, comprises all pairwise distances between elements of the input sequences x and y. Each matrix element is computed as $c(x_i, y_j) = \|x_i - y_j\|_p$ for $i \in [N]$ and $j \in [M]$, where $[N] = \{1, 2, \ldots, N\}$, $[M] = \{1, 2, \ldots, M\}$, and the norm $\|x_i - y_j\|_p$ can be chosen as any $p$-norm (typically with $p = 1, 2$, or $\infty$) depending on the application requirements[27]. Using this matrix, the algorithm identifies an optimal alignment path that traverses through the low-cost regions, establishing correspondences between elements while preserving temporal relationships.

The alignment path in DTW, formally called the warping path, is defined as a sequence $\mathrm{w} = (w_1, w_2, \ldots, w_K)$ , where each element $w_k = (n_k, m_k)$ represents a mapping between points in the two sequences such that $w_k \in [N] \times [M]$ for all $k \in [K]$, $\max\{N, M\} \leq K < M + N - 1$. Also, let $\mathcal{N}_\mathrm{w} = \{n_k\}_{k=1}^K$ and $\mathcal{M}_\mathrm{w} = \{m_k\}_{k=1}^K$. This path must satisfy three fundamental constraints that ensure meaningful temporal alignment[28]. First, the boundary condition requires that $w_1 = (1, 1)$ and $w_K = (N, M)$, ensuring complete sequence coverage. Second, the monotonicity condition preserves temporal ordering through the requirements $1 = n_1 \leq n_2 \leq \cdots \leq n_K = N$ and $1 = m_1 \leq m_2 \leq \cdots \leq m_K = M$. Third, the continuity or step size condition prevents excessive temporal distortion by requiring that $w_{k+1} - w_k \in \{(1, 1), (1, 0), (0, 1)\}$. The quality of a warping path w is quantified through a cost function that aggregates the local costs along the path, defined as

$$\mathcal{C}_{\mathrm{w}}(\mathrm{x},\mathrm{y}) = \sum_{k=1}^{K} c(x_{n_k}, y_{m_k}) = \sum_{k=1}^{K} \|x_{n_k} - y_{m_k}\|_p.$$

Theoretically, obtaining the optimal warping path $\mathrm{w}^\star$ require testing all possible paths that minimizes the total warping cost, $\mathcal{C}_{\mathrm{w}^\star}(\mathrm{x},\mathrm{y}) = \min_{\mathrm{w}} \left\{ \mathcal{C}_{\mathrm{w}}(\mathrm{x},\mathrm{y}), \mathrm{w} \in \mathcal{P}^{N\times M} \right\}$ where $\mathcal{P}^{N\times M}$ represents the set of all valid warping paths satisfying the conditions in[28]. This could be computationally challenging due to the exponential growth of the number of optimal paths as $M$ and $N$ grow linearly[29]. Instead, DTW employs Dynamic Programming (DP) to efficiently compute the optimal alignment[29]. The algorithm constructs an accumulated cost matrix and the optimal warping path can be found by using the recursive formula given by

$$D_p(\mathrm{x},\mathrm{y}) = \mathcal{C}_{\mathrm{w}^\star}(\mathrm{x},\mathrm{y}) = \sum_{i\in\mathcal{N}_{\mathrm{w}^\star}} \sum_{j\in\mathcal{M}_{\mathrm{w}^\star}} d_p(x_i, y_j), \tag{1}$$

where $d_p(x_i, y_j) = \|x_i - y_j\|_p + \min\{d(x_{i-1}, y_{j-1}), d(x_i, y_{j-1}), d(x_{i-1}, y_j)\}$ is the cumulative distance with initialization conditions

$$d_p(x_1, y_j) = \sum_{k=1}^{j} \|x_1 - y_k\|_p \quad \text{and} \quad d_p(x_i, y_1) = \sum_{k=1}^{i} \|x_k - y_1\|_p$$

for the first row and first column, respectively. This approach reduces the computational complexity to $\mathcal{O}(MN)$. If $M = N$, this simplifies to $\mathcal{O}(N^2)$.

The DTW distance measure satisfies the basic properties of a distance function defined in[30]. Given that $\mathrm{a}$ and $\mathrm{b}$ are vectors, the DTW distance measure conforms with non-negativity property ($D(\mathrm{a},\mathrm{b}) \geq 0$), definiteness property ($D(\mathrm{a},\mathrm{b}) = 0 \Leftrightarrow \mathrm{a} = \mathrm{b}$), symmetry property ($D(\mathrm{a},\mathrm{b}) = D(\mathrm{b},\mathrm{a})$), and triangle inequality property ($D(\mathrm{a},\mathrm{c}) \leq D(\mathrm{a},\mathrm{b}) + D(\mathrm{b},\mathrm{c})$).

The proposed method will use DTW which has been demonstrated to excel at calculating distances between time series data while handling phase shifts and time lag differences. There are some challenges that we will address. Its computational complexity poses significant challenges due to the sequential nature of dynamic programming which limits parallelization opportunities. Although researchers have explored various optimization approaches including lower bounds techniques[31], DP parallelization[32], and GPU acceleration[33], the complexity remains particularly challenging for higher-dimensional applications. Notably, two-dimensional DTW for image matching[34] can reach $\mathcal{O}(N^6)$ complexity, making it impractical for large datasets. To address these limitations, MiniDTW was proposed by[16] as an efficient alternative that summarizes time series data into natural-shaped seed clusters using norm distance, which are then merged through a Sparse Symmetric Non-negative Matrix Factorization (SSNMF) algorithm that factorizes the cluster centers' distance matrix. This approach demonstrated remarkable efficiency improvements, avoiding 97.90% of DTW computations and achieving a speed that is 10 times faster than the TADPole baseline method, which only reduced DTW utilization by 75.73%[16].

## Partition around medoids (PAM)

Our proposed framework utilizes the Partition Around Medoids (PAM) which offers distinct advantages for time series clustering compared to traditional centroid-based methods. Unlike algorithms that compute centroids as abstract points in multidimensional space, PAM selects actual data points (medoids) as cluster centers. This ensures that cluster centers are always valid time series from the dataset, making them interpretable and robust to outliers or data near the boundaries[35,36]. The PAM algorithm addresses the clustering optimization problem through an iterative process that minimizes within-cluster dissimilarity. This dissimilarity is quantified by calculating the sum of distances between each observation and its assigned medoid, utilizing DTW as the distance metric. Although PAM is considered as computationally more intensive than *K*-means, especially for large datasets[36], the optimization procedure follows the efficient implementation developed by[37], which significantly reduces computational complexity compared to traditional PAM implementations.

To initialize the algorithm, $M$ samples are randomly selected from the complete dataset $\mathrm{Z} = (\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_N)^\top \in \mathbb{R}^{N\times T}$. These $M$ samples serve as the initial medoids. Let $m \in \{1, 2, \ldots, M\} = [M]$ represent the medoid index. The set of medoids $\mathrm{Q}$ comprises individual medoids $\mathrm{q}_m \in \mathbb{R}^{1\times T}$, where each medoid is a time series vector of length $T$. The set $\mathrm{Q}$ is formally defined as $\mathrm{Q} = (\mathrm{q}_1, \mathrm{q}_2, \ldots, \mathrm{q}_M)^\top \in \mathbb{R}^{M\times T} \subset \mathrm{Z}$. This notation emphasizes that medoids are always actual time series sequences drawn from the original dataset $\mathrm{Z}$. Meanwhile, define $\mathrm{D}$ as the $N \times N$ DTW distance matrix between all pairs of time series in $\mathrm{Z}$. The iterative optimization procedure then proceeds as follows.

1. Using $\mathrm{D}$, associate the non-medoid samples from the set $(\mathrm{Z} - \mathrm{Q})$ into the closest medoid $\mathrm{q}_m \in \mathrm{Q}$ when $\min_{n,m} D(\mathrm{z}_n, \mathrm{q}_m), \mathrm{z}_n \neq \mathrm{q}_m$ is satisfied, $n \in [N], m \in [M]$.
2. Let $\mathrm{P}_m$ be the set containing the $m$th medoid and the associated samples to it i.e., $\mathrm{z}_n \neq \mathrm{q}_m, \mathrm{z}_n, \mathrm{q}_m \in \mathrm{P}_m$. Calculate the total distance cost

$$C = \sum_{m=1}^{M} \left( \sum_{\mathrm{z}_n, \mathrm{q}_m \in \mathrm{P}_m} D(\mathrm{z}_n, \mathrm{q}_m) \right).$$

3. Perform the swap as follows:

   (a) In each $\mathrm{P}_m$, randomly select a sample $\mathrm{z}_n \in \mathrm{P}_m$ to be the new temporary medoid $\mathrm{q}_m^*$. This temporarily updates the medoids and the set of medoids $\mathrm{q}_m^* \in \mathrm{Q}^*$.
   (b) Using D, associate the non-medoid samples from the set $(\mathrm{Z} - \mathrm{Q}^*)$ into the closest medoid $\mathrm{q}_m^* \in \mathrm{Q}^*$ when $\min_{n,m} D(\mathrm{z}_n, \mathrm{q}_m^*), \mathrm{z}_n \neq \mathrm{q}_m^*$ is satisfied.
   (c) Let $\mathrm{P}_m^*$ be the temporary set containing the new *m*th medoid and the associated samples to it i.e., $\mathrm{z}_n \neq \mathrm{q}_m^*, \mathrm{z}_n, \mathrm{q}_m^* \in \mathrm{P}_m^*$.
   (d) Calculate the temporary total distance cost as

$$C^* = \sum_{m=1}^{M} \left( \sum_{\mathrm{z}_n, \mathrm{q}_m^* \in \mathrm{P}_m^*} D(\mathrm{z}_n, \mathrm{q}_m^*) \right).$$

   (e) Compare the total distance cost *C* and the temporary cost $C^*$.

(i) If $C > C^*$, assign the following variables and repeat the swap steps:
(A) $\mathrm{q}_m := \mathrm{q}_m^* \; \forall \; m$.
(B)$\mathrm{Q} := \mathrm{Q}^*$.
(C) $\mathrm{P}_m := \mathrm{P}_m^* \; \forall \; m$.
(D) $C := C^*$.
(ii) Else, undo the swap i.e., the values of $\mathrm{q}_m, \mathrm{Q}, \mathrm{P}_m, C$ stay as is and repeat the swap steps. The algorithm converges when membership of $\mathrm{P}_m$ stabilizes or after an arbitrary maximum iteration specified by the analyst is reached.

While medoids and centroids both represent cluster centers, they differ fundamentally in their construction: medoids are *actual* observations selected from within the cluster, whereas centroids are abstract points computed as the geometric center of the cluster. The PAM algorithm iteratively optimizes cluster assignments by selecting *M* medoids and associating samples to them based on DTW distances. The optimization objective is to minimize the total distance *C*, defined as the sum of distances between each observation and its assigned medoid. This total distance *C* serves as a quantitative measure of clustering quality. During each iteration, the algorithm performs swap steps to refine both medoid selection and cluster memberships, aiming to reduce the total distance *C*. When a medoid $\mathrm{q}_m$ is replaced, the cluster memberships in $\mathrm{P}_m$ are automatically reevaluated since samples $\mathrm{z}_n$ may become closer (in terms of DTW distance) to different medoids, necessitating reassignment. The algorithm generates several essential outputs for the dataset Z: the metadata of Z, the final selected medoids $\mathrm{q}_m$, cluster assignments for each sample $\mathrm{z}_n$, and the complete DTW distance matrix D. These outputs collectively provide an extensive representation of the clustering solution and its underlying distance structure.

## Methodology

This study presents a methodological framework for analyzing time series responses in case-control experiments, the methods is then illustrated using the ADHD-200 fMRI data. In neuroimaging studies, multivarate time series (e.g., fMRI) may be recorded multiple times (across multiple trials). Given the high-dimensional nature of multiple time series data, we implement a pre-processing algorithm prior to clustering. The between-cluster time series distances are then computed to quantify effect sizes, which inform subsequent power analyses. While our primary focus is the analysis of multivariate time series responses in case-control experiments, we extend our framework to include power analysis for sample size (number of participants) determination. This is particularly valuable when cases are relatively rare or experimental conditions are difficult to control, making optimal sample size planning crucial for detecting meaningful differences (effect size) between groups.

### Analysis of experiments with time series responses

A typical experimental data often employ analysis of variance (ANOVA) to analyze treatment effects through comparison of group means and variances. However, when working with high-frequency, high-dimensional time series responses, we require a parallel analytical framework that preserves the temporal structure of the data while maintaining the fundamental goal of detecting and quantifying treatment effects. The responses in our framework consist of multiple high-frequency time series, where each observation represents a trajectory through time rather than a single point measurement.

To formulate this framework mathematically, suppose there are *G* groups, the dataset is represented as $N \times T$ matrix $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_G)^\top$, where $\mathrm{X}_g = (\mathrm{x}_{g1}, \mathrm{x}_{g2}, \ldots, \mathrm{x}_{g,N_g})^\top$ is the $N_g \times T$ matrix of observations in the *g*th group, and $\mathrm{x}_{gn} = (x_{gn}(t_1), x_{gn}(t_2), \ldots, x_{gn}(t_T))$ is the $1 \times T$ vector of time series measurements from the *n*th subject in the *g*th group, with $g \in \{1, 2, \ldots, G\} = [G]$ and $n \in \{1, 2, \ldots, N_g\} = [N_g]$. This structure indicates that the data X, from *N* subjects, can be partitioned into *G* hypothesized groups, each subject has *T* time points of measurements (response). This representation allows us to maintain parallel representation

with the group-wise comparisons in ANOVA while accommodating the temporal nature of our data. In a typical experimental setup, responses within each treatment group are assumed to come from similar distributions. For time series responses, subjects belonging to the same treatment groups are expected to exhibit comparable temporal dependencies, characterized by features such as autocorrelation functions, cross-correlations, periodograms, and coherence.

The analytical framework begins with data pre-processing, analyst needs to determine how to manage multiple time series per subject or group. The complexity of this decision is well illustrated in fMRI studies, where each subject generates at least $10^5$ Blood Oxygen Level Dependent (BOLD) time series measurements. These BOLD measurements, initially available at the voxel level, can be aggregated at various scales: region, network, hemisphere, or even at the whole-subject level. The selection of an appropriate aggregation level is driven by the objectives of the study, specifically the scale at which treatment effects would be most informative for the experiment. To manage complexity and enhance signal detection, each subject is typically represented by a single summarized vector of time series measurements. This aggregation step is crucial, as the absence of such dimensionality reduction could lead to high-dimensional time series responses that can potentially mask the treatment effects the experimenter aims to observe. Additionally, standard pre-processing steps such as outlier detection, handling missing values, spatial and temporal alignment are performed to ensure data quality and compatibility for subsequent analysis steps. Note further that too much aggregation could lead in missing out valuable signals present in the multiple time series.

Following pre-processing, the second step is distance computation and similarity assessment. In a typical experimental data, this reduces to computation of group means and variances. The $N \times T$ matrix $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_G)^\top$ can be expressed as $(\mathrm{x}_{11}, \ldots, \mathrm{x}_{1,N_1}, \ldots, \mathrm{x}_{G1}, \ldots, \mathrm{x}_{G,N_G})^\top = (\mathrm{z}_1, \ldots \mathrm{z}_{N_1}, \ldots, \mathrm{z}_{N_{G-1}+1}, \ldots, \mathrm{z}_N)^\top = \mathrm{Z}$, where Z represents the concatenation of the observations in $\mathrm{X}_1$, $\mathrm{X}_2$, until $\mathrm{X}_G$ and $N = N_1 + N_2 + \cdots + N_G$. Dynamic Time Warping (DTW) distances are then computed between all pairs of time series in Z, resulting in an $N \times N$ distance matrix D. By construction, D is symmetric with zero diagonal elements $D(\mathrm{z}_l, \mathrm{z}_l) = D_{ll} = 0$, i.e., distance of a time series with itself is zero. The non-negative elements of D are in fact Euclidean distances between data points in one-dimension, $D(\mathrm{z}_k, \mathrm{z}_l) = D_{kl} = \|\mathrm{z}_k - \mathrm{z}_l\|_2$, where $k, l \in [N]$. These pairwise distances quantify both between-group and within-group variations in temporal patterns, providing a foundation for subsequent analysis steps. Our method leverages on a key advantage of DTW which is its ability to handle time series of varying lengths, though its computational complexity necessitates optimized implementation as the size of Z increases further.

The third step implements time series clustering for group identification, applying Partition Around Medoids (PAM) clustering while leveraging on a priori knowledge of number of groups *G*. For the case where $G = 2$, we employ the previously computed DTW distance matrix D within a hypothesis testing framework based on the Welch *t*-statistic developed by[38]. This approach innovatively reformulates group comparisons in terms of aggregated pairwise differences across all time points of the series. The Welch *t*-statistic is given by

$$T_W^2 = \frac{\sum_{k=1}^{N}\sum_{k<l}^{N} D_{kl}^2 - \frac{N}{N_1}\sum_{k=1}^{N_1}\sum_{k<l}^{N_1} D_{kl}^2 - \frac{N}{N_2}\sum_{k=N_1+1}^{N_1+N_2}\sum_{k<l}^{N_1+N_2} D_{kl}^2}{\frac{N_1 N_2}{N_1^2(N_1-1)}\sum_{k=1}^{N_1}\sum_{k<l}^{N_1} D_{kl}^2 + \frac{N_1 N_2}{N_2^2(N_2-1)}\sum_{k=N_1+1}^{N_1+N_2}\sum_{k<l}^{N_1+N_2} D_{kl}^2}. \tag{2}$$

where $N_1$ and $N_2$ represent the sample sizes of each group, and $D_{kl}$ represents the DTW distance between time series from subjects *k* and *l*. The numerator captures the overall variability in the combined sample relative to within-group variations, while the denominator accounts for the uncertainty in these estimates, equivalent to the partial sum of squares ratio in ANOVA. While this is formulated for two groups, the framework naturally generalizes to multiple groups, $G > 2$, allowing for the analysis of more complex experimental designs parallel to the traditional ANOVA approaches. The DTW distances and the resulting Welch *t*-statistic serve as quantitative measures for evaluating clustering quality (i.e., similarity with and differences between clusters) and estimating effect sizes needed for power calculations. For example, in the fMRI dataset, these measures help assess the separation between control and ADHD groups while also providing the numerical basis for determining required sample sizes in future neuroimaging studies for similar subjects.

The fourth step is to optimize the output of clustering algorithm, focusing on the selection of optimal cluster configurations to minimize the effect of the choice of initial medoids. This step includes validation of cluster stability and robustness, as well as the assessment of cluster separation as a measure of treatment effect. The optimization process is important as it directly impacts the quality of subsequent power analysis. Finally, the fifth step is power analysis and sample size determination, where traditional power analysis is modified to adapt specifically for the time series data. Adaptation process utilizes cluster distances as effect size measures and develops simulation-based power estimation procedures, providing researchers with practical guidance for future study design.

To demonstrate the practical implementation of this methodology, we utilize the ADHD-200 fMRI dataset[39]. This neuroimaging dataset exemplifies the challenges of analyzing high-dimensional, high-frequency time series in case-control studies. The pre-processing step addresses the substantial dimensionality of fMRI data, where the BOLD time series measurements from each subject must be appropriately aggregated, e.g., across different voxels. Following aggregation, the DTW distances are computed between the temporal patterns of brain activity between subjects, capturing the complex neurological dynamics that distinguish ADHD from control groups. The PAM clustering algorithm then leverages these DTW distances to identify distinct patterns in brain activity, while the Welch *t*-statistic framework quantifies the significance of these group differences. The methodology culminates in power calculations that translate observed effect sizes into practical guidelines for sample size determination in future neuroimaging studies.

This methodological framework extends traditional ANOVA concepts to accommodate the complex nature of modern experimental data, where responses are captured as high-dimensional (in space), high-frequency time series rather than single measurements. By carefully preserving temporal dependencies while maintaining statistical rigor, the approach enables robust detection of treatment effects in diverse applications, from neuroimaging studies to other fields generating rich temporal data. The ability of this framework to handle such complex data structures, while providing clear guidance for future experimental design through power analysis, makes it particularly valuable for contemporary case-control studies where traditional analysis methods may be insufficient.

## Selection of time series clustering outcome

The PAM clustering results based on the DTW distance matrix can vary depending on the initial selection of medoids. Since the iterative clustering process uses DTW to determine medoids, the final clusters are influenced by which time series are initially grouped together. To assess the stability of these clusters and the reliability of the cluster distances (which represent the treatment effect), we perform the clustering process multiple times on each time series dataset Z. Although the algorithm typically converges in fewer than 12 iterations, the computational time remains challenging. However, these multiple clustering replications help establish consistency in the resulting clusters, particularly in the composition of $\mathrm{P}_m$ across replicates. Selection of the optimal clustering replicate is crucial, as it enables more reliable sample size calculations in the subsequent power analysis since the effect of initialization in clustering has been mitigated.

The algorithm consists of two parts. The first part processes each replicate from the time series clustering of dataset Z, with the following steps:

1. Let $N_m$ be the count of samples $\mathrm{z}_n \neq \mathrm{q}_m$ and medoid $\mathrm{q}_m$ in the $m$th cluster $\mathrm{P}_m$.
2. Calculate the average DTW distance of the $m$th medoid to the associated samples $\mathrm{z}_n$ of the $m$th cluster $\mathrm{P}_m$ and the sum of average distances within clusters given by

$$\overline{D}_m = \frac{1}{N_m - 1} \sum_{\forall \mathrm{z}_n} \sum_{\mathrm{q}_m \in \mathrm{P}_m} D(\mathrm{z}_n, \mathrm{q}_m), \quad M\overline{D} = \sum_{m=1}^{M} \overline{D}_m. \tag{3}$$

3. Obtain the sum of distances between all pairs of medoids given by

$$\sum_{m=1}^{M} \sum_{m^*=1}^{M} D(\mathrm{q}_m, \mathrm{q}_{m^*}). \tag{4}$$

The second part of the algorithm focuses on analyzing the aggregated results to determine the optimal clustering outcome, proceeding with these steps:

1. Eliminate result of replicates that satisfies $\{\exists\ m \in [M] \mid \overline{D}_m = 0\}$.
2. Calculate the selection criteria value given by

$$\tau = a - \sum_{m=1}^{M} \sum_{m^*=1}^{M} D(\mathrm{q}_m, \mathrm{q}_{m^*}) + \sum_{m=1}^{M} \overline{D}_m. \tag{5}$$

3. Define the rank of the selection criteria value in ascending order as $\mathrm{Rank}(\tau)$.
4. Choose the result of replicate with the lowest value of $\tau_r$ i.e., the lowest $\mathrm{Rank}(\tau)$.

The tuning parameter $a \in \mathbb{R}$ was incorporated into Eq. (5) to establish a threshold for determining the optimal clustering replicate. In our empirical studies, we set $a = 1000$. Thus, the selection criteria value $\tau$ balances two equally important components. The first component measures the tightness of the clustered samples are around their representative centers (medoids) by summing the average distances within each cluster as shown in (3). The second component measures the separation between the clusters by calculating the distance between their representative centers in (4). Since both components use the same type of distance measure (DTW), they are naturally comparable without any need for scaling or rank-based combinations. Thus, one advantage of our proposed method is the direct comparability which allows for a straightforward assessment of clustering quality that considers both within-cluster compactness between-clusters separation. Moreover, in low-dimensional cases, this can be visualized as forming clusters where members are close to each other while maintaining maximum separation between units that belong to different clusters.

An important constraint in selecting the optimal clustering outcome is that no cluster should contain only its medoid. When $\overline{D}_m = 0$, it indicates that cluster $\mathrm{P}_m$ contains only one member, the medoid $\mathrm{q}_m$. While such singleton clusters might occasionally be justified by the data structure, they can artificially lower the sum of the average distances within clusters in (3), leading to potentially misleading optimization results. Therefore, each cluster $m$ must contain at least two samples. To illustrate this concept, consider the clustering result in Table 1 where we consider the case when $M = 2$. It shows the comparison of using outright ranking compared to that of the selection criteria value. This was implemented on simulated dataset with $N = 12$ and $T = 100$.

| Replicate | $\overline{D}_1 + \overline{D}_2$ | $R_1$ | $D(q_1, q_2)$ | $R_2$ | $R_1 + R_2$ | $\tau$ | (cut) R |
|---|---|---|---|---|---|---|---|
| 3 | 124.0 | 2 | 80.6 | 1 | 3 | 1043.4 | 2 |
| 11 | 112.0 | 1 | 79.2 | 2 | 3 | 1032.8 | 1 |

**Table 1**. Illustration of optimized time series cluster rankings: within-cluster and between-cluster distance metrics (lower ranks indicate better performance).

| Method | Assignment | Description |
|---|---|---|
| Method 1 | Random | The samples are allocated randomly and equally into *G* groups |
| Method 2 | Model-based | The samples are systematically allocated into *G* groups based on the model to which the sample was generated from |
| Method 3 | Clustering Based | The samples are allocated to the groups based on the result of the aggregated time series clustering using DTW |

**Table 2**. Comparison of subject allocation methods for statistical power determination.

When comparing replicates using simple rank-based methods ($R_1$ and $R_2$), two different clustering outcomes might receive the same overall ranking because each has distributed advantage in different criteria. However, this ranking method fails to capture the actual magnitude of their advantages. This limitation led to developing a more nuanced criteria value $\tau$ that better aligns with our definition of optimal clustering. The arbitrary value of $a \in \mathbb{R}$ was included in the equation so that both terms are on the same direction such that the lower combined value, the better. This supports the selection process by picking the result with the least value of $\tau$ because this is the optimal clustering replicate.

Furthermore, calculating $\tau$ across replicates may yield identical values, indicating that these replicates have produced exactly the same clusters, i.e., the same representative centers (medoids) and identical groupings of samples. This duplicate clustering outcome occurs more frequently in two cases: when analyzing smaller datasets and when working with data generated from heterogeneous processes, as observed in our empirical studies. In cases where multiple clustering results share the lowest $\tau$ value, any of these equivalent solutions can be chosen and as a convention, the first replicate from the list is selected.

## Sample size and power

Sample size determination represents a critical foundation of robust experimental design, requiring careful consideration of statistical power. Conventionally set at 80%, this power is influenced by a complex interplay of factors including sample size, absolute group difference magnitude, intrinsic measurement variability, and the nominal level of significance $\alpha$[40]. Our investigation focuses on high-dimensional time series data that extends beyond traditional fMRI BOLD measurements to include diverse high-frequency signals acquired through advanced neuroimaging techniques and wearable sensor technologies. We propose that the DTW clustering algorithm, when applied to these temporal response patterns, can effectively reveal meaningful group distinctions. This approach operates on the fundamental premise that observed pattern variations reflect actual treatment effects rather than random fluctuations — particularly in neurological contexts where signal patterns correspond to functional brain responses. The strength of between-group differences can be rigorously quantified by calculating distance metrics between time series clusters derived from different experimental conditions. This quantification provides an empirical foundation for subsequent statistical analyses and interpretation of results.

Our methodological framework for determining optimal sample size comprises two sequential phases: initially, participants are categorized into *G* distinct groups using one of several grouping approaches; subsequently, sample size requirements are calculated based on the measured magnitude of between and within group differences, while adhering to predetermined Type I ($\alpha$) and Type II ($\beta$) error thresholds. We evaluate three distinct grouping methodologies (detailed in Table 2), with particular emphasis on Dynamic Time Warping (DTW) for time series clustering. This technique offers valuable perspectives on how different participant grouping strategies influence power analysis outcomes. To establish methodological validity, we compare our DTW-based clustering approach against conventional random assignment and traditional model-based classification methods, which serve as comparative benchmarks.

In Method 1, participants were systematically allocated into *G* equal-sized groups using a pure randomization approach. This methodological procedure guarantees that each subject maintains an identical probability of assignment to any experimental group, effectively eliminating potential selection bias while establishing statistically independent groupings. The random allocation serves as a foundational baseline condition that preserves both group size equivalence and the inherent stochastic properties essential for robust statistical inference and comparison. This random assignment method was deliberately incorporated into our empirical evaluation framework because it provides critical insights into power analysis dynamics under conditions where no a priori grouping structure exists among the samples. This approach is particularly valuable in exploratory contexts where underlying grouping mechanisms remain undefined or when researchers need to establish a reference distribution against which other allocation strategies can be meaningfully compared.

On the other hand, the model-based assignment method employed a structured allocation approach, systematically taking the $\lfloor N/G \rfloor$ samples and associating them into $G-1$ distinct groups while allocating the remaining samples to Group *G*. Unlike random assignment, this method leverages a priori knowledge of the

underlying data generation models, thereby creating groups that reflect theoretical constructs or experimental conditions. This approach is particularly valuable when investigating phenomena with well-established taxonomies or when evaluating the sensitivity of statistical power calculations to structured group differences.

Lastly, the clustering-based assignment method represents a data-driven approach that dynamically forms groups based on inherent patterns within the time series data. Specifically, this method employs DTW as a distance metric to quantify similarities between temporal patterns following an aggregation procedure. By identifying natural clusters in the multidimensional time series space, this approach transcends predetermined categorizations and instead reveals emergent groupings based on functional or behavioral similarities. The DTW algorithm accommodates temporal shifts and varying progression rates, making it ideally suited for neurophysiological and behavioral time series that often exhibit complex alignment challenges. This method offers particular advantages when investigating heterogeneous populations where subgroup characteristics may not be apparent through conventional clustering approaches.

Starting with the $N \times T$ matrix $\mathrm{Z} = (\mathrm{z}_1, \ldots, \mathrm{z}_N)^\top$, we compute DTW distances between all pairs of time series in Z, yielding an $N \times N$ distance matrix D with elements:

$$D(\mathrm{z}_k, \mathrm{z}_l) = D_{kl} = \begin{cases} \|\mathrm{z}_k - \mathrm{z}_l\|_2, & k \neq l; k, l \in [N] \\ 0, & k = l. \end{cases}$$

These pairwise distances capture both between-group and within-group variations in temporal patterns, establishing the foundation for subsequent analysis. Each sample $\mathrm{z}_n \in \mathrm{Z}, n \in [N]$ is assigned to its respective group $g \in [G]$ according to the specified sample grouping method in Table 2. In the clustering-based method, the $m$th cluster $\mathrm{P}_m$ corresponds to the $g$th group. Using any of the three methods, the dataset Z is partitioned into $G$ distinct groups, with each group $g$ containing $\widetilde{N}_g$ samples and $\widetilde{N}_1 + \widetilde{N}_2 + \cdots \widetilde{N}_G = N$.

After the assignment of samples to $G$ groups, the DTW matrix D is rearranged to form distinct group partitions given by

$$\underset{N\times N}{\mathrm{D}} = \begin{pmatrix} \underset{\widetilde{N}_1\times\widetilde{N}_1}{\mathrm{D}_{11}} & \underset{\widetilde{N}_1\times\widetilde{N}_2}{\mathrm{D}_{12}} & \cdots & \underset{\widetilde{N}_1\times\widetilde{N}_G}{\mathrm{D}_{1G}} \\ \underset{\widetilde{N}_2\times\widetilde{N}_1}{\mathrm{D}_{21}} & \underset{\widetilde{N}_2\times\widetilde{N}_2}{\mathrm{D}_{22}} & \cdots & \underset{\widetilde{N}_2\times\widetilde{N}_G}{\mathrm{D}_{2G}} \\ \vdots & \vdots & \ddots & \vdots \\ \underset{\widetilde{N}_G\times\widetilde{N}_1}{\mathrm{D}_{G1}} & \underset{\widetilde{N}_G\times\widetilde{N}_2}{\mathrm{D}_{G2}} & \cdots & \underset{\widetilde{N}_G\times\widetilde{N}_G}{\mathrm{D}_{GG}} \end{pmatrix} \tag{6}$$

where $\mathrm{D}_{gh}$ represent the between-group distances across groups $g$ and $h$ when $g \neq h$ while it corresponds to within-group distances when $g = h$. Each submatrix $\mathrm{D}_{gh}$ contains the pairwise distances

$$\underset{\widetilde{N}_g\times\widetilde{N}_h}{\mathbf{D}_{gh}} = \begin{array}{c} \\ \mathbf{x}_{g1} \\ \mathbf{x}_{g2} \\ \vdots \\ \mathbf{x}_{g,\widetilde{N}_g} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{x}_{h1} & \mathbf{x}_{h2} & \cdots & \mathbf{x}_{h,\widetilde{N}_h} \end{array} \\ \begin{bmatrix} D(\mathbf{x}_{g1}, \mathbf{x}_{h1}) & D(\mathbf{x}_{g1}, \mathbf{x}_{h2}) & \cdots & D(\mathbf{x}_{g1}, \mathbf{x}_{h,\widetilde{N}_h}) \\ D(\mathbf{x}_{g2}, \mathbf{x}_{h1}) & D(\mathbf{x}_{g2}, \mathbf{x}_{h2}) & \cdots & D(\mathbf{x}_{g2}, \mathbf{x}_{h,\widetilde{N}_h}) \\ \vdots & \vdots & \ddots & \vdots \\ D(\mathbf{x}_{g,\widetilde{N}_g}, \mathbf{x}_{h1}) & D(\mathbf{x}_{g,\widetilde{N}_g}, \mathbf{x}_{h2}) & \cdots & D(\mathbf{x}_{g,\widetilde{N}_g}, \mathbf{x}_{h,\widetilde{N}_h}) \end{bmatrix} \end{array}. \tag{7}$$

The following quantities can be computed using the DTW matrix:

$$\mathcal{D}_{gh} = \sum_{k=1}^{\widetilde{N}_g} \sum_{l>k}^{\widetilde{N}_h} D^2(\mathrm{x}_{gk}, \mathrm{x}_{hl}) \tag{8}$$

$$\mathcal{D}_{g} = \sum_{k=1}^{\widetilde{N}_g} \sum_{l>k}^{\widetilde{N}_h} D^2(\mathrm{x}_{gk}, \mathrm{x}_{gl}) \tag{9}$$

$$S_g^2 = \frac{\mathcal{D}_g}{N_g\,(N_g - 1)} \tag{10}$$

The sum of squares specified in (8) pertains to the sum of squares of the pairwise distances of samples between groups $g$ and $h$, $g \neq h$. These corresponds to the sum of squares of the upper triangular elements $\{D(\mathrm{x}_{gk}, \mathrm{x}_{hl})\}_{k<l}$ of the partition $\mathrm{D}_{gh}$. Likewise, (9) is the sum of squares of the pairwise distances of samples within group $g$ which is the sum of squares of the upper triangular elements $\{D(\mathrm{x}_{gk}, \mathrm{x}_{gl})\}_{k<l}$ of the partition $\mathrm{D}_{gg}$.

When there are two groups ($G = 2$), such as in case-control studies, researchers compare individuals with specific neurological or psychiatric conditions to healthy controls by examining differences in their time series patterns. For this comparison to be meaningful, researchers must establish that time series patterns are similar within each group (cases similar to other cases, controls similar to other controls) while showing clear differences between the groups. After establishing these pattern distinctions, we can use (8) and (9) to calculate the distance-based effect size $\delta_G$ from[38] given by

$$\delta_G = \sqrt{0\frac{\frac{\sum_{g=1}^{G}\sum_{h>g}^{G}\mathcal{D}_{gh}}{\prod_{g=1}^{G}N_g} - \left(\frac{\sum_{g=1}^{G}N_g}{\prod_{g=1}^{G}N_g}\right)\sum_{g=1}^{G}\frac{\mathcal{D}_g}{N_g}}{\frac{\sum_{g=1}^{G}\frac{\mathcal{D}_g}{N_g}}{\sum_{g=1}^{G}(N_g-1)}}}. \tag{11}$$

The process of sample size determination relies on the power function relationship defined as $1-(\mathrm{F}_T(t^\star)-\mathrm{F}_T(-t^\star))=1-\beta$, where $\mathrm{F}_T(\cdot)$ represents the cumulative distribution function of the test statistic. In this formulation, $t^\star = t_{\alpha/2,\nu_G}(\lambda)$ denotes the critical value from the non-central $t$-distribution with non-centrality parameter $\lambda = \delta_G\sqrt{N_u}/2$, $N_u$ is the required sample size, and $\nu_G$ is the degrees of freedom associated with the linear combination of the sample variances in (10) approximated by the Welch-Satterthwaite equation as follows

$$\nu_G \approx \frac{\left(\sum_{g=1}^{G}\frac{S_g^2}{N_g}\right)^2}{\sum_{g=1}^{G}\frac{S_g^4}{N_g^2(N_g-1)}}. \tag{12}$$

This power function has no closed-form solution for the sample size $N_u$. Consequently, when designing a study with pre-specified values of significance level $\alpha$ and power $1-\beta$, we must solve for $N_u$ numerically. The procedure involves systematically evaluating the power function for different values of $N_u$ until we identify the minimum sample size that achieves the desired power for a given effect size $\delta_G$. This numerical approach generates a collection of discrete sample size determinations across various effect sizes. To transform these discrete solutions into a continuous and practically applicable tool, we apply kernel regression smoothing. The expected sample size given an effect size is modeled as $\mathbb{E}(N_u|\delta_G) = m(\delta_G)$, where the function $m(\delta)$ is estimated using a Nadaraya-Watson kernel estimator defined as

$$\widehat{m}_b(\delta_G) = \frac{\sum_{u=1}^{U}K_b(\delta_G-\delta_u)N_u}{\sum_{u=1}^{U}K_b(\delta_G-\delta_u)}.$$

In this estimator, $K_b(\cdot)$ represents the kernel function with bandwidth parameter $b$ that controls the smoothness of the regression curve, and $U$ denotes the number of data points (pairs of effect sizes $\delta_u$ and corresponding sample sizes $N_u$) used in the estimation. The kernel function assigns weights to observations based on the proximity of their effect sizes to the target effect size, with greater weight given to closer observations. This approach enables us to generate a continuous function that maps any effect size within the estimation range to an appropriate sample size recommendation. By applying this methodology across various combinations of significance levels, power values, and other relevant study parameters, we create a comprehensive sample size planning framework. The resulting kernel function $\widehat{m}_b(\delta_G)$ serves as a statistical tool that provides immediate sample size recommendations for researchers, eliminating the need for repeated numerical calculations when planning studies across different effect size scenarios.

## Simulation studies

Empirical experiments were conducted to assess both our proposed clustering algorithms and power analysis methodology for sample size determination. To establish realistic simulation parameters, we analyzed the ADHD-200 fMRI dataset's characteristics, which informed our Data Generating Process (DGP) design, particularly regarding time point quantities ($T$) and source model specifications. In addition, the simulation framework was designed to examine hypothesis testing in a case-control context following the two-sample setting of the ADHD 200 data.

Our simulated datasets fall into two primary categories based on their source model derivation: single group datasets where $G=1$, and multigroup datasets where $G>1$. Table 3 presents a comprehensive overview of the various DGP scenarios across both dataset classifications. While some sample size values (100 and 1000) were selected arbitrarily, others (12 and 50) reflect the median sample sizes reported in the literature for fMRI single

| Dataset parameter | Parameter values for single group dataset | Parameter values for multigroup dataset |
|---|---|---|
| Dataset replicates | 100 | 100 |
| Sample size | $N$ = 12, 50, 100, 1000 | $N$ = 12, 50, 100, 1000 |
| Time points | $T$ = 100, 150, 200 | $T$ = 100, 150, 200 |
| Number of groups | $G$ = 1 | $G$ = 2, 3, 4, 8 A multigroup dataset can only have one of the above options. The number of hypothesized groups applied on specific dataset are presented in the succeeding tables |
| Proportion of samples per group $g$ | Not applicable | Equal proportions with some small adjustments done to match target total sample size |

**Table 3**. Data generating process scenarios for single group and multigroup dataset.

group and structural MRI studies, respectively[41]. The time point values $T$ were derived from the most frequently occurring counts in valid BOLD time series from the ADHD-200 dataset, rounded to the nearest 50.

The simulated data generated through our DGP framework also enables a extensive evaluation of time series clustering methods employing DTW, with detailed implementation parameters presented in Table 4. For each dataset, we executed 100 clustering replicates to ensure statistical robustness and reliability of our findings. An important methodological consideration is that while DTW consistently produces identical distance matrices when calculating pairwise distances between samples within a given dataset, the final clustering solutions exhibit natural variation across replicates. This variability stems from inherent stochastic elements in the clustering algorithm's implementation — specifically, the random initialization of medoids and subsequent centroid calculations. These randomized starting conditions, despite operating on identical underlying distance measures, introduce controlled variability that allows for a more thorough assessment of the clustering method's stability and performance characteristics.

Our DGP settings incorporated additional scenarios to reflect various group characteristics that might influence clustering performance. The time series models with their respective parameters used in generating both single-group and multi-group datasets are documented in Appendix C. These model specifications were not arbitrarily chosen but were based on empirical findings from the ADHD-200 dataset analysis, where subject-level BOLD time series were best modeled by AR(5) processes in 33% of cases and AR(1) processes in 21% of cases. To ensure coverage of potential temporal dynamics, we also included MA and ARMA models in our simulation framework, which effectively account for the higher-order autoregressive structures observed in neuroimaging data. All the time series model settings among the DGP scenarios have fixed mean $\mu = 0$ and error variance $\sigma_\varepsilon^2 = 1$. This standardization is justified by our preliminary exploratory simulations, which demonstrated that variations in these particular parameters exert only marginal effects on clustering outcomes. The mean parameter $\mu$ merely induces a vertical shift in the location of the time series without altering its fundamental pattern or temporal dynamics. Similarly, the error variance $\sigma_\varepsilon^2$ simply modifies the scale of the time series through contraction or expansion, preserving the underlying temporal structure that is critical for clustering performance. By holding these parameters constant, we effectively isolate and evaluate the impact of other model characteristics that more significantly influence cluster formation and hypothesis testing results.

For single group datasets, all generated samples adhere consistently to their specified model parameters. In contrast, the multigroup datasets implement a structured allocation strategy wherein each model specification initially receives an equal distribution of samples, with any remaining samples (when $N$ is not perfectly divisible by the number of models) being systematically allocated to the first model(s) in the sequence presented in the reference tables. Critically, each generated sample retains metadata identifying its originating model — a design choice essential for our research objective of examining how clustering outcomes are influenced by the underlying time series generating mechanisms.

Our simulation study incorporates 4 distinct sample size levels ($N$), 3 different time point dimensions ($T$), 34 diverse model specifications, and 100 replicate datasets for each unique parameter combination. This extensive design framework has yielded a total of 40,800 unique datasets, providing a robust foundation for our investigation into time series clustering sensitivity and performance. This large-scale simulation approach ensures that our findings regarding the effectiveness of DTW-based clustering are generalizable across a wide spectrum of data characteristics commonly encountered in neuroimaging research.

## Results and discussion

In this section, estimates of the treatment effect following time series clustering procedures are presented for both our simulated datasets and the ADHD-200 fMRI data. These estimates provide crucial insights into the performance and efficacy of our clustering methodology across different data structures. Additionally, this section includes a detailed analysis of the sample size-power relationship after the application of time series clustering. This analysis highlights the statistical implications of our clustering approach, quantifying how the clustering process affects the statistical power of subsequent analyses at various sample sizes. Through this dual presentation of treatment effect estimates and sample size-power analysis, we establish a complete statistical framework for researchers applying time series clustering to neuroimaging data.

### Empirical studies

*Estimates of treatment effect*

The time series clustering algorithm successfully converged across all simulation study scenarios, demonstrating its computational reliability. The processing time required for implementing time series clustering based on DTW exhibits a proportional relationship to the product of sample size and time points ($NT$), as illustrated in Fig. 1. Generally, the processing time increases linearly with sample size $N$. This relationship becomes particularly

| Simulation parameter | Parameter values |
|---|---|
| No. of time series clustering replicates | 100 |
| No. of clusters to be formed | $M = 2$ |
| Distance calculation method | DTW |
| Clustering method | Partitional |
| Centroid calculation | Partition around medoids (PAM) |

**Table 4**. Simulation settings for empirical evaluation of proposed method.

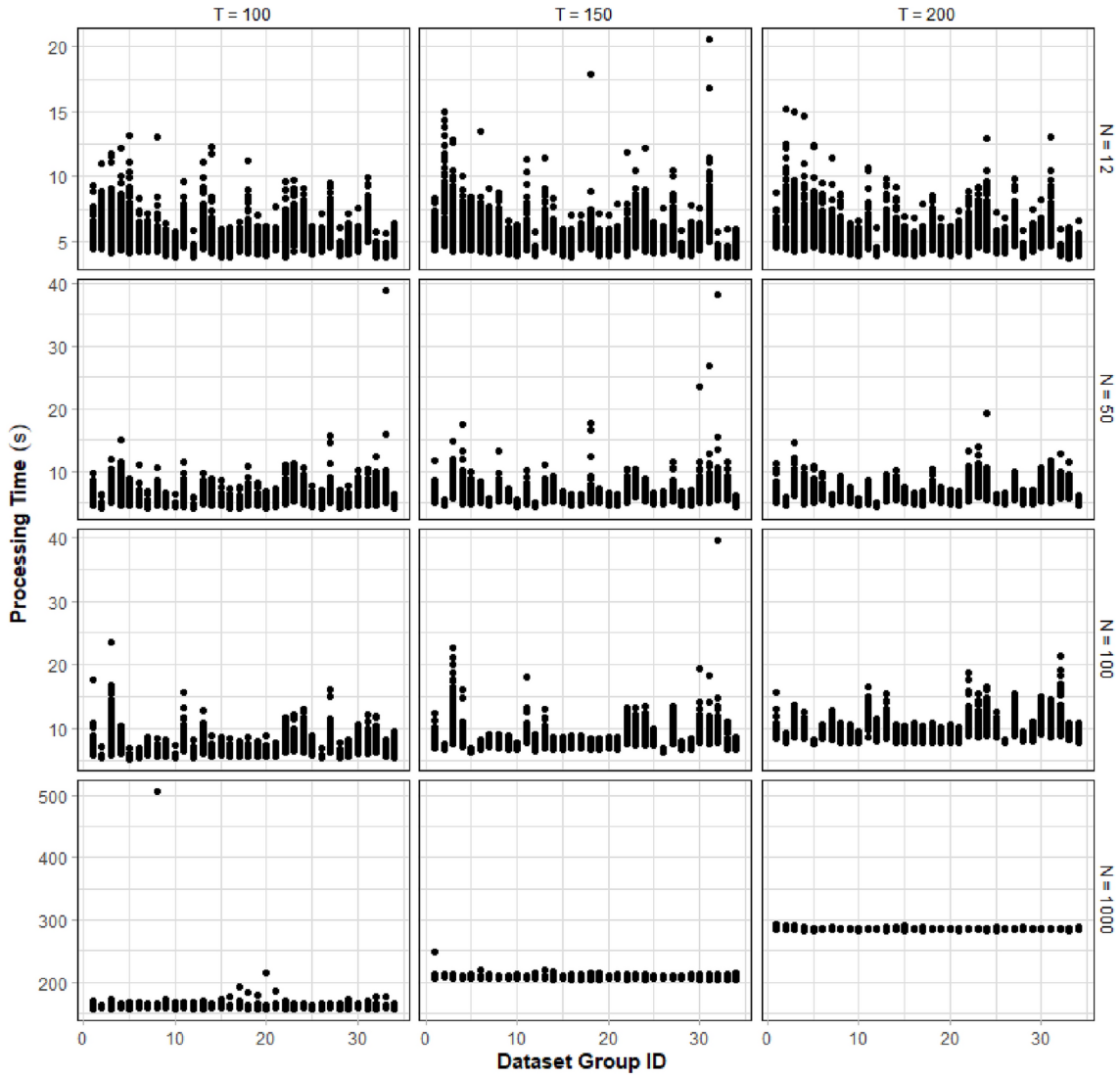


**Fig. 1**. Visual representation of computational processing times across different dataset configurations. The plot demonstrates how processing time scales with both sample size $N$ and time points $T$. Higher $N$ and $T$ generally result in increased processing times, with particularly notable increases for $N = 1000$.

evident in simulations with larger sample sizes, such as $N = 1000$, where the impact of varying time point dimensions $T$ on processing time becomes more pronounced. Specifically, larger values of $T$ consistently result in notably higher processing times, reflecting the increased computational demands associated with comparing longer time series sequences during the DTW calculation process.

Moreover, datasets with smaller sample sizes $N$, such as 12, demonstrate a distinctive multi-modal distribution in the occurrence of cluster sizes. In these cases, the clustering algorithm behaves in a manner that produces nearly all possible combinations of cluster sizes with relatively equal frequency. This pattern is consistently observed across all dataset groups, suggesting that with limited samples, the clustering solution space is more uniformly explored. Meanwhile, the first-order autocorrelation coefficient $\rho_1$ exerts a notable influence on cluster membership distribution. As time series approach non-stationarity, they tend to drift collectively in the same direction, resulting in reduced distances between series after DTW processing. This phenomenon becomes particularly pronounced in single group datasets generated from the same model family, such as AR(1), as visualized in Fig. 2. Multigroup datasets exhibit a distinct clustering tendency when both groups are characterized by high autocorrelation values. As shown in Fig. 3, these datasets frequently produce results where the majority of samples are consolidated into a single dominant cluster. This imbalanced allocation occurs because high autocorrelation in multiple groups creates similar drift patterns across different underlying models, effectively reducing the discriminatory power of DTW despite theoretical differences in the generating processes. The time series from different groups with high autocorrelation values develop comparable temporal trajectories that the clustering algorithm interprets as belonging to the same underlying pattern, despite their different generative origins.

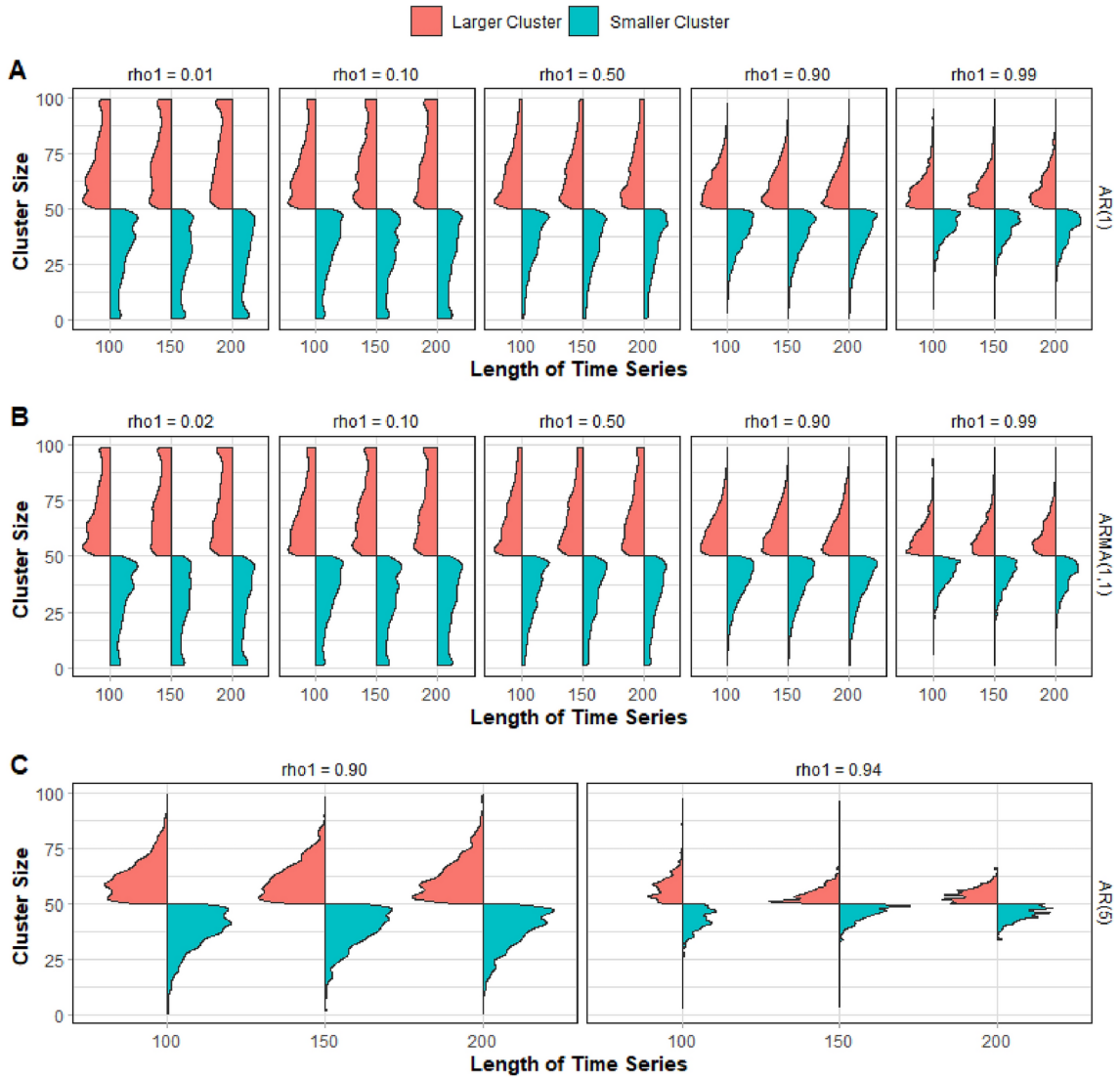


**Fig. 2**. Distribution of cluster sizes in single group dataset: violin plots showing the distribution of cluster sizes across different time series models, lengths of time series $T$, and fixed $N = 100$. Panels A-C compare cluster size distributions for AR(1), ARMA(1,1) and AR(5) models with varying autocorrelation values, respectively. This displays how autocorrelation strength influences cluster formation patterns.

As the number of time points $T$ increases, we observe that the average distance between clusters tends to increase as well. This suggests that distance is more strongly influenced by $T$ than by the sample size $N$. Figure 4 illustrates this relationship, showing a consistent upward trend in the distance between cluster centroids as time progresses. This pattern is not limited to between-cluster distances; a similar increasing trend appears in the within-cluster average distances, regardless of cluster size. Both smaller clusters (Fig. 5) and larger clusters (Fig. 6) demonstrate this same time-dependent behavior in their internal distance metrics.

It is also important to note that time series data generated from higher-order AR models, such as AR(5), consistently exhibit greater average distance spreads compared to other models. This finding is intuitive, as models with more parameters can capture subtle differences between time series (and consequently between clusters) that simpler models might miss. Upon closer examination of each model family, we find that the first-order autocorrelation coefficient $\rho_1$ emerges as the key factor influencing cluster distances. Specifically, higher autocorrelation values correspond to larger average distances in both smaller and larger clusters. The data reveals a particularly notable pattern: time series generated from AR(1) and ARMA(1,1) models display significantly larger average cluster distances compared to those generated from MA(1) models.

The primary goal of aggregating time series clustering results across replicates is to identify the optimal clustering solution for each simulation scenario, as measured by our chosen metric $\tau$. This aggregation process ensures that we select only those replicates that most effectively differentiate between samples and optimize cluster assignments. Figure 7 compares the clustering results before and after aggregation, specifically examining the average distances within both smaller and larger clusters. The figure demonstrates that our aggregation

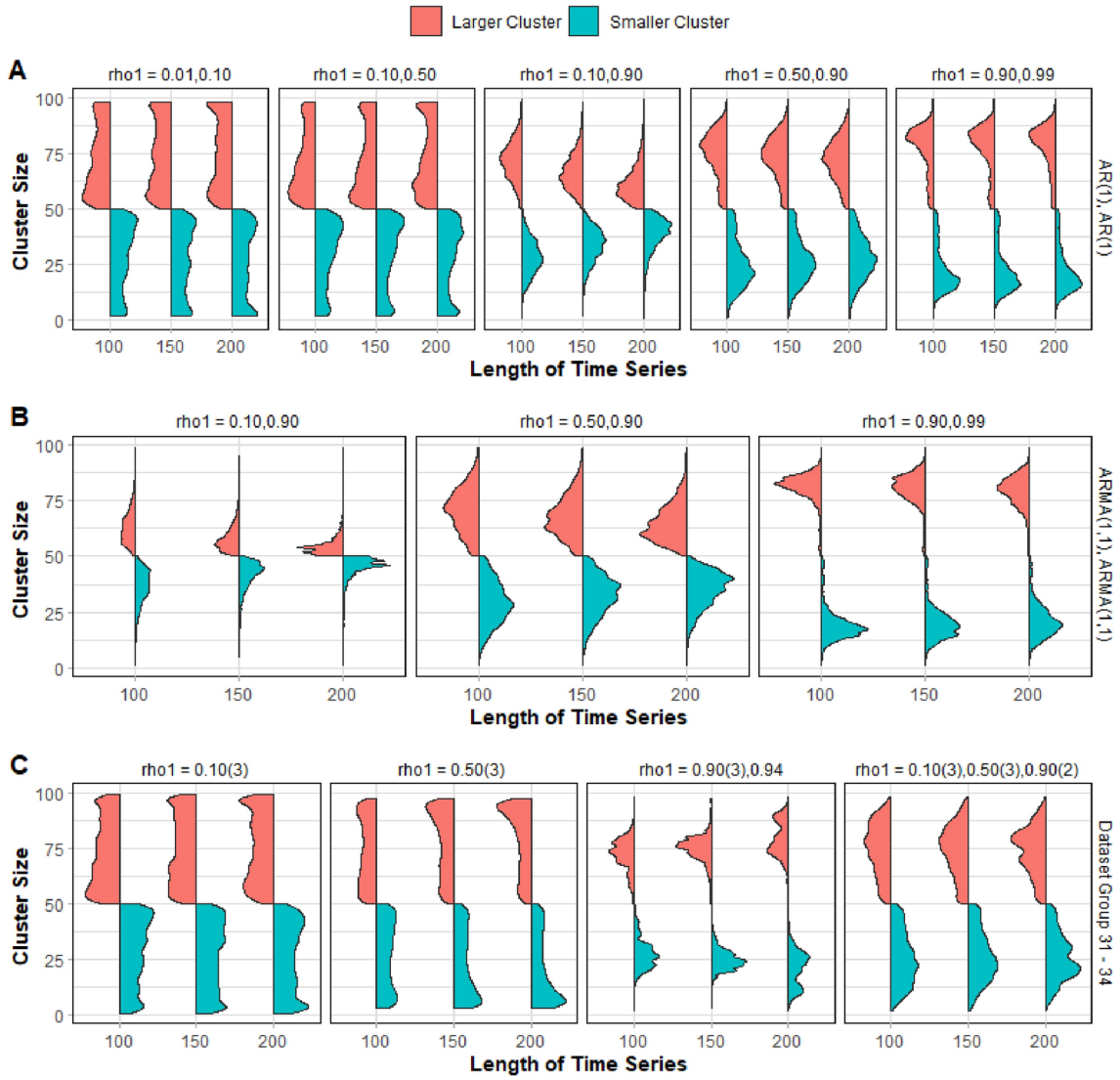


**Fig. 3**. Distribution of cluster sizes in multigroup dataset: violin plots showing how cluster sizes are distributed across different time series models when analyzing multi-group time series data for fixed $N = 100$. The plot reveals patterns in how samples are partitioned between clusters across different model specifications and lengths of time series. Higher autocorrelation in both groups tends to produce more balanced cluster sizes.

procedure successfully selects replicates that minimize within-cluster distances while maximizing the distances between cluster centroids. This dual optimization precisely embodies the definition of our selection criterion $\tau$.

In addition, Fig. 8 presents the post-aggregation results, revealing distinct patterns across different time series models. For datasets generated from AR(1) and ARMA(1,1) models with higher autocorrelation, the distribution of samples across clusters maintained its original shape, with a tendency toward equal-sized clusters. In contrast, datasets from MA(1) models and those from AR(1) and ARMA(1,1) models with lower autocorrelation showed a different pattern after aggregation. In these cases, the retained replicates favored an uneven distribution where the larger cluster contained the majority of samples, despite this not being characteristic of the original distribution patterns.

*Power and sample size*

Distances between clusters are taken as treatment effect and subsequently used in the computation of the sample size. The range of values of effects size $\delta$ are also computed for each scenario (for all dataset replicates). Assuming the level of significance $\alpha$ and power $1 - \beta$, we compute the sample size $N$ from a power function constructed by kernel regression. A more flexible power function is necessary because the inputs are based on simulated data. However, given an effect size, one can easily use the nonparametric power function to compute the sample size.

The sample sizes computed from the method aligns with the statistical convention on power analysis and sample size calculation, i.e., given the probability of Type I error at $\alpha$ , power $1 - \beta$, and degrees of freedom, the sample size requirement increases for dataset scenarios with lower values of effect size $\delta$ and increases with

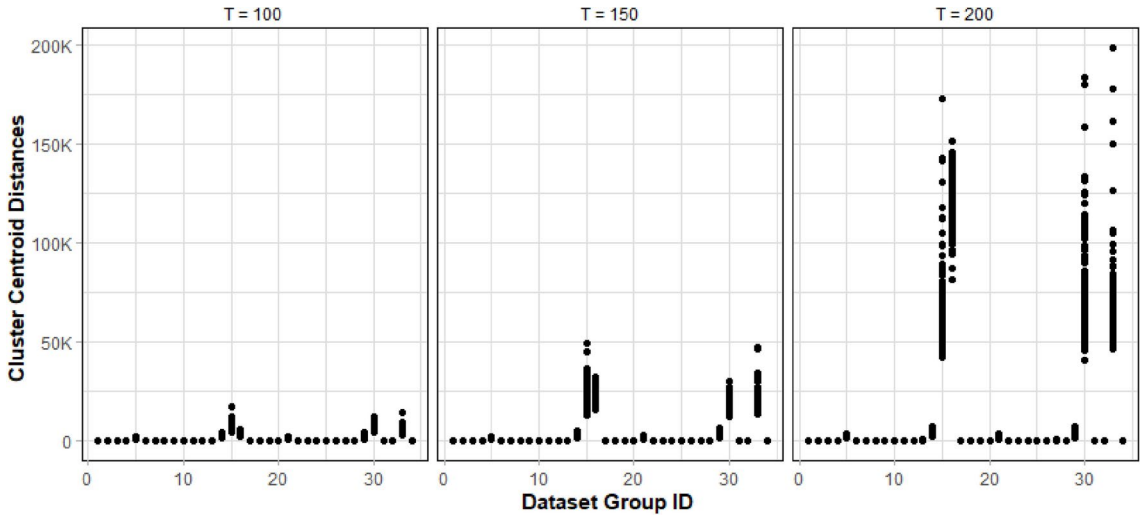


**Fig. 4**. Comparison of DTW distances between cluster centroids across different dataset configurations and temporal resolutions under fixed $N = 100$. The plot presents an increasing trend of centroid distances with larger $T$ values, demonstrating how temporal resolution affects cluster separation. This relationship is particularly pronounced in datasets with high autocorrelation, where longer time series lead to more distinct cluster formations. The systematic increase in distances across time points suggests that longer temporal sequences provide better discrimination between cluster centers.

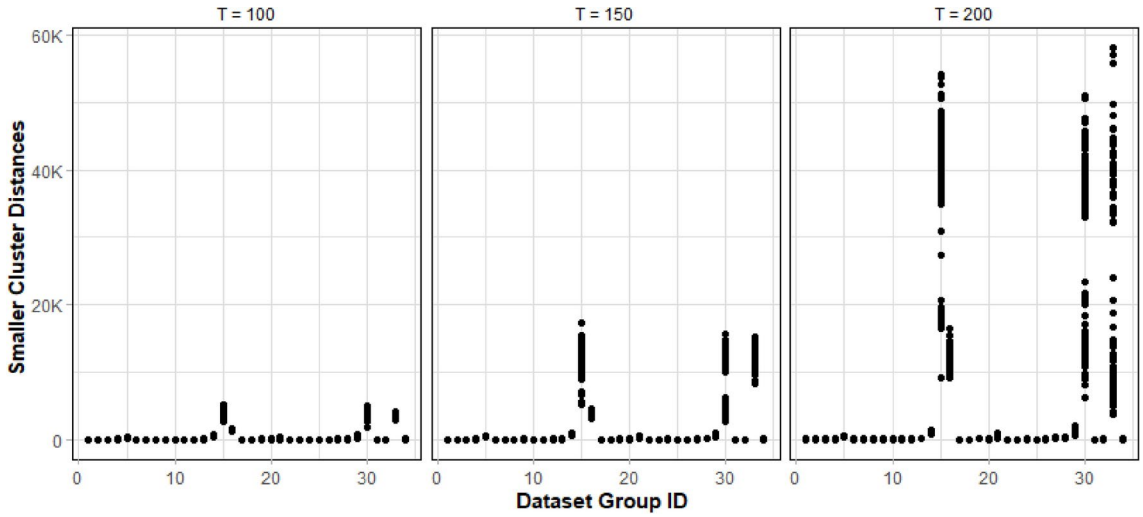


**Fig. 5**. Comparison of the average DTW distances within smaller clusters across dataset groups and temporal scales under fixed $N = 100$. The visualization demonstrates how cohesion within smaller clusters varies with both data characteristics and temporal resolution. Higher $T$ values consistently result in larger within-cluster distances, indicating that longer time series tend to form more dispersed clusters. This pattern provides insights into cluster stability and the impact of temporal resolution on cluster formation in smaller groupings.

larger effect size. An example is shown in Fig. A1 in the Appendix for the dataset simulated from AR(1) with high autocorrelation value $\rho = 0.99$ and an input sample size of 100.

Upon closer examination, we noted that for time series data generated from AR(1) model, larger values of $\rho_1$ requires lower sample size to achieve the desired power. Larger autocorrelation (but still stationary) implies more prominent (similar) time series patterns, closer to each other, resulting in larger effect size when compared to another group. It is obvious from Fig. A2 in the Appendix that the effect size computed from the data with high $\rho_1$ is higher than that of the lower $\rho_1$. The MA(1) generated data shows no significant difference on sample size requirement even for varying $\rho_1$ as shown in Fig. A3 in the Appendix. Recall however, that the maximum value for $\rho_1$ (autocorrelation at lag 1) in MA(1) is approximately 0.50 which is far from non-invertibility of the time series.

The ARMA(1,1) datasets also exhibits similar behavior to AR(1) datasets, i.e., higher $\rho_1$ results in lower sample size requirement as shown in Fig. A4 in the Appendix. This observation also extends to higher order AR

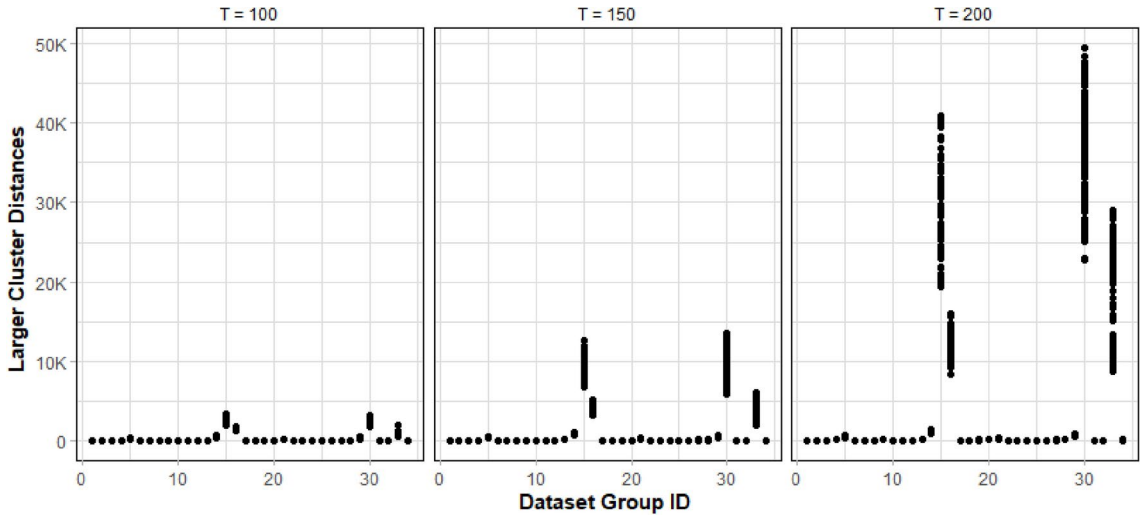


**Fig. 6**. Comparison of the average DTW distances within larger clusters across dataset groups and temporal scales under fixed $N = 100$. The plot reveals generally lower and more stable within-cluster distances compared to smaller clusters, suggesting stronger cohesion in larger groupings. The relationship between temporal resolution and within-cluster distance remains evident but shows less variation than in smaller clusters, indicating more robust cluster formation in larger groups. This pattern holds particular significance for understanding the stability of cluster assignments in time series analysis.

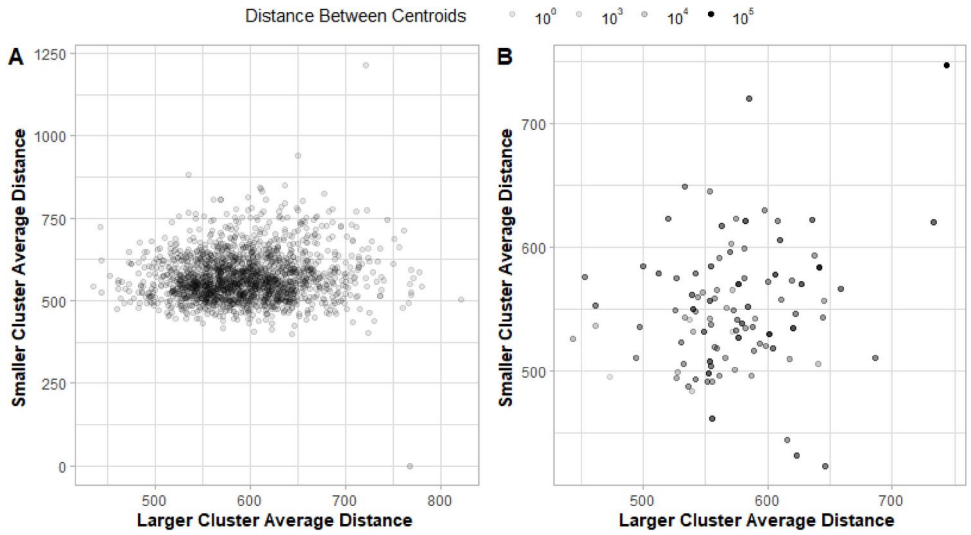


**Fig. 7**. Comparative scatter plot visualization demonstrating the difference between raw clustering results and post-aggregation outcomes for average cluster distances in a selected dataset sourced from AR(1) model with high autocorrelation and a fixed $N = 100$. Panel A displays the raw clustering distribution while Panel B shows the results after optimization through the $\tau$ metric. The reduced scatter and more focused distribution in Panel B demonstrates the effectiveness of the aggregation procedure in selecting optimal clustering outcomes that minimize within-cluster distances while maximizing between-cluster separation. This visualization provides clear evidence that the aggregation process successfully identifies clustering solutions that better align with the theoretical goals of cluster analysis, showing how the $\tau$ metric effectively filters out suboptimal clustering results while preserving those that achieve both strong cluster cohesion and separation.

model data as shown in Fig. A5 in the Appendix. Autocorrelation is the more influential feature of the time series to the sample size requirement and not the order of model.

Other methods for calculating the sample size from Table 4 are also explored as a reference. In Method 1 (Random Assignment Method), samples are assigned randomly and equally to both groups, while Method 2 (Model-based Assignment Method) assigns samples into groups according to the model where they were generated from, applicable for multigroup datasets. Sample sizes computed from the Random Assignment Method are presented in Fig. A6 in the Appendix while Model-based Assignment Method are in Fig. A7 in the

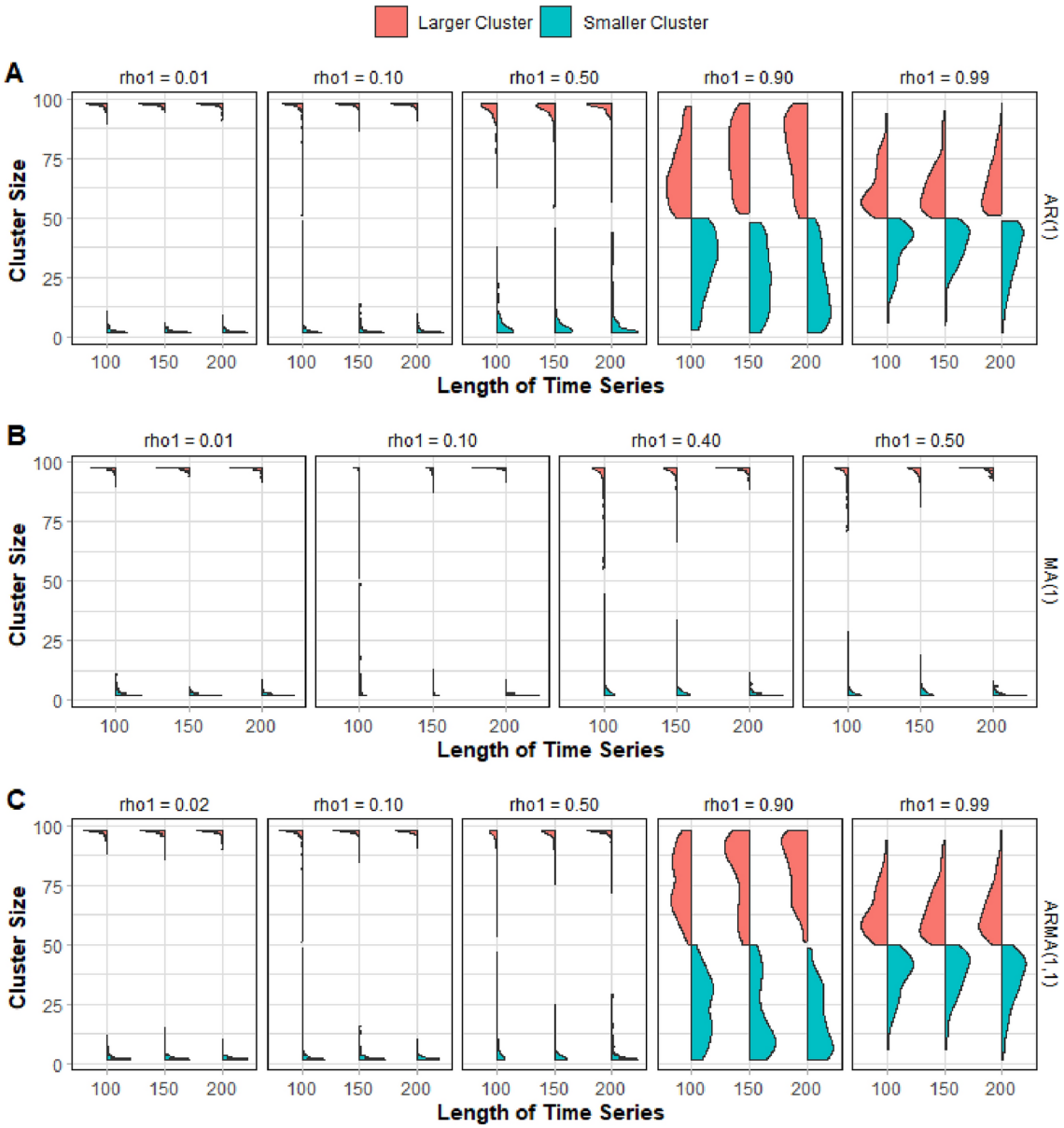


**Fig. 8**. Violin plots showing the results after aggregating multiple clustering replicates for $N = 100$. The visualization demonstrates how cluster size distributions stabilize after optimization using the $\tau$ metric. The plot reveals that models with higher autocorrelation in AR(1) and ARMA(1,1) maintain more balanced cluster size distributions even after aggregation.

Appendix. It is clear that Method 3 (Clustering-based Assignment Method) is more efficient in distinguishing similarities and differences among samples from different groups, resulting to sample size requirement that is more reasonable (conservative) compared to the other methods.

The sample size computed from various scenarios in the empirical studies are generalized into a nonparametric OC curve estimated through kernel regression. Figure A8 in the Appendix illustrates the sample size calculated over various values of the bandwidth of the OC curve. For the second column (derived from AR(1) with high autocorrelation), the sample size requirement was calculated for all scenarios for the bandwidth that was selected for the model. Just like in any other power analysis cases, the effect size needs to be approximated from existing data, whether this is a past study or based on pilot studies. The data used to calculate the effect size can also be used in computing the optimal bandwidth through the cross-validation method.

## Application to ADHD-200

*ADHD-200 dataset*

The ADHD-200 dataset[39] is a product of the collaborative effort across eight international neuroimaging sites, and provides an open-access resource designed to facilitate research into Attention-Deficit/Hyperactivity Disorder (ADHD). This dataset is composed of various neuroimaging modalities, output from standardized

preprocessing pipelines, and detailed phenotypic information, offering a large-scale] and reproducible research for studying brain structure and function. The contributing sites include Bradley Hospital/Brown University (BU), Kennedy Krieger Institute (KKI), NeuroIMAGE, New York University Child Study Center (NYU), Oregon Health and Science University (OHSU), Peking University (PU), University of Pittsburgh (UP), and Washington University in St. Louis (WU)[42].

This study focused specifically on one of the imaging modalities available in the dataset: the resting-state functional MRI (rs-fMRI). The dataset includes both fMRI scans and phenotype information from 973 subjects, comprising 585 typically developing children (controls) and 362 children diagnosed with ADHD, while 26 subjects had an unknown diagnostic status[43].

The raw fMRI scans from the cohort of 947 participants were processed through the Athena Neuroimaging Preprocessing Pipeline[44], utilizing tools such as the Analysis of Functional NeuroImages (AFNI) and FMRIB's Software Library (FSL). This preprocessing was performed on Virginia Tech's High Performance Computing (HPC) cluster[39]. The preprocessing pipeline of the fMRI data followed a series of well-established steps to ensure data quality and consistency. First, the initial four Echo Planar Imaging (EPI) volumes were removed to allow magnetization stabilization. When a scan begins, the MRI scanner takes a few seconds to reach a steady-state signal due to initial fluctuations in the magnetization field. These early volumes can contain artifacts or inconsistent signal intensity, which could negatively impact the quality of the data. By discarding these volumes, we ensure that the data used in the analysis reflects a stable and consistent signal, improving the overall reliability and accuracy of the subsequent analysis steps. Next, slice timing correction was then applied to account for temporal differences in slice acquisition. The dataset was then deobliqued and reoriented into the Right-Posterior-Inferior (RPI) orientation, ensuring the right hemisphere of the brain is on the right side of the image.

To correct for motion artifacts, the EPI volumes were aligned to the first image of the time series, followed by brain masking to exclude non-brain regions. A mean image was generated by averaging the volumes and subsequently co-registered with the corresponding anatomical image. Both the fMRI data and the mean image were then written into template space with a resolution of 4 mm $\times$ 4 mm $\times$ 4 mm. Next, white matter (WM) and cerebrospinal fluid (CSF) masks, derived from anatomical preprocessing, were downsampled to match the EPI resolution. The time courses for WM and CSF were extracted from the EPI volumes using the respective masks. To further refine the data, WM, CSF, and motion time courses derived from the motion correction step, along with a low-order polynomial for detrending, were regressed from the EPI data. Finally, a band-pass filter ($0.009\ \text{Hz} < f < 0.08\ \text{Hz}$) was applied to the voxel time courses to exclude frequencies that are widely believed to be unrelated to resting-state functional connectivity. Both the filtered and unfiltered data were then spatially smoothed using a 6-mm Full Width at Half Maximum (FWHM) Gaussian filter. The detailed description of each step is accessible on the NITRC website[39]. Certain filtering steps were applied further to fit the goal of the analysis. The number of qualifying subjects are also shown on Table 5 as the selected subjects.

The fMRI images produced from the Athena pipeline were standardized in the Montreal Neurological Institute (MNI) space, with a voxel resolution of 4 $\text{mm}^3$. Across all subjects, the three spatial dimensions remained consistent, with uniform measurements ($d_x = 49$ units, $d_y = 58$ units, $d_z = 47$ units), yielding an overall image size of 196 mm $\times$ 232 mm $\times$ 188 mm. Figure 9 presents the processing and analysis pipeline for the ADHD-200 dataset. In the illustration, the green box indicates the data source, while the yellow circles highlight the different data types. The icons, primarily sourced from flaticon.com, along with the arrows, depict the sequential steps involved.

To extract individual images of regions of interest (ROIs), a brain atlas is used to assign each voxel to a specific region, network, and hemisphere. In this study, the 2018 Schaefer 17-network, 100-region parcellation atlas[45] was used. This atlas is derived from and aligned with Thomas Yeo's 2011 17-network parcellation of the cerebral cortex[46], providing a fine-grained mapping of the brain's functional networks. Figure 10 illustrates the inflated brain image based on the 2018 Schaefer 17-network, 100-region atlas[47]. The atlas, initially available at a voxel resolution of 1 $\text{mm}^3$, was resampled using AFNI to a resolution of 4 $\text{mm}^3$ to match the voxel size of the subjects' fMRI images. This resampling ensured that the atlas and fMRI data shared identical spatial dimensions. Subsequently, each of the 100 regions was extracted and saved as an individual image file. These region-specific masks were then used to generate BOLD time series for each voxel, across all regions and subjects, facilitating region-wise analysis of the fMRI data.

Finally, only the fMRI BOLD time series data from the most recent valid scan of each subject were used, ensuring that each subject contributed only one scan dataset. Additionally, instead of using all 100 regions in the

| Imaging site | No. of subjects downloaded | Dimension measure (units) | Unit length (s) | No. of selected subjects |
|---|---|---|---|---|
| KKI | 83 | 148 | 2.50 | 78 |
| NeuroIMAGE | 48 | 257 | 1.96 | – |
| NYU | 222 | 172 | 2.00 | 126 |
| OHSU | 79 | 74 | 2.50 | 63 |
| PU | 194 | 232 | 2.00 | 191 |
| UP | 89 | 192 | 1.50 | – |
| WU | 61 | 129 | 2.50 | – |
| Total | 776 | | | 458 |

**Table 5**. No. of subjects downloaded and selected.

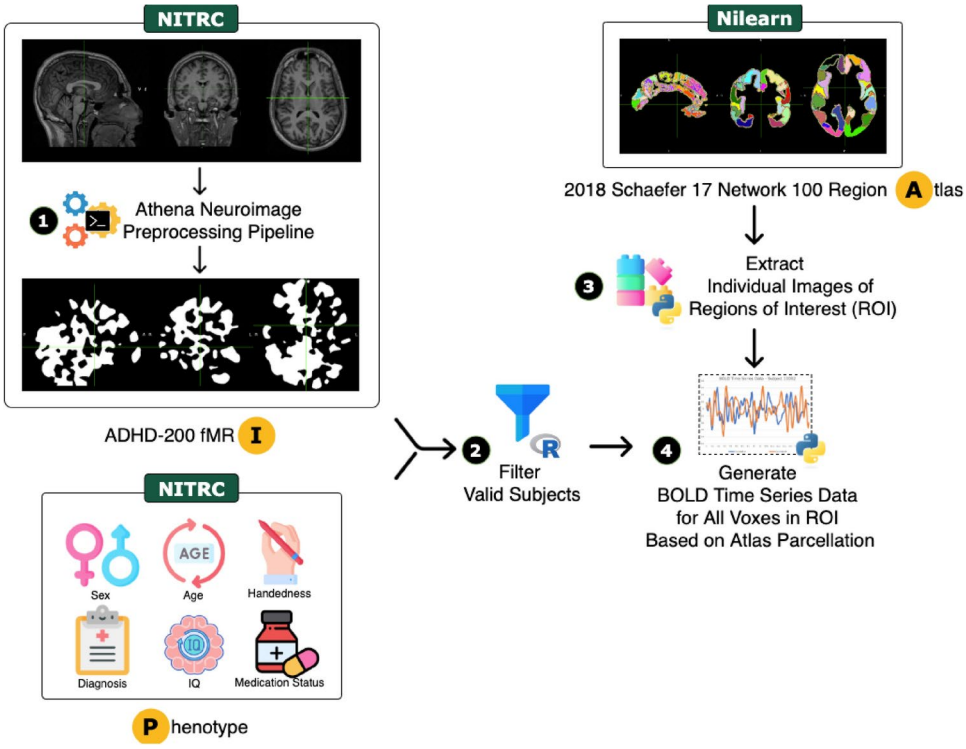


**Fig. 9**. Comprehensive flowchart illustrating the complete data processing workflow, from raw fMRI data acquisition through preprocessing steps to final analysis. Key components include the Athena Neuroimaging Preprocessing Pipeline, phenotype data integration, and ROI extraction using the Schaefer Atlas. Icons and arrows clearly indicate the sequential flow of data processing steps.

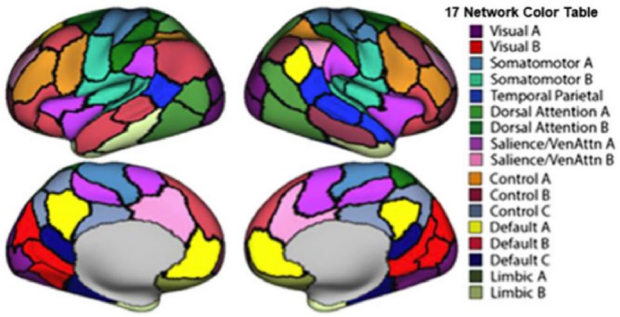


**Fig. 10**. Brain surface visualization showing the 17-network, 100-region parcellation scheme used for ROI definition. Different colors represent distinct functional networks, demonstrating how the brain is divided into functionally-related regions. This parcellation provides the anatomical framework for subsequent time series analysis.

atlas, only a subset of 82 regions that have a corresponding counterpart in the opposite hemisphere were selected for analysis. As a result, each of the 458 participants had a total of 13,658 time series curves measured across the 82 regions, with each curve containing up to 232 time points.

*Estimates of treatment effect*

We also applied our proposed method to analyze the ADHD-200 dataset. This dataset contains BOLD measurements from subjects across various imaging sites, with each subject's time series having a different length. To maintain consistency with our simulation setup presented in Table 3, we standardized the analysis by truncating each subject's time series to lengths of $T = 100, 150$, and 200 time points. For visual representation of the data, Fig. 11 displays the BOLD signals grouped by diagnosis, while Fig. 12 provides a more detailed visualization that categorizes the signals by both imaging site and diagnosis.

It is important to note that the Dynamic Time Warping (DTW) procedure described in section "Methodology" offers considerable flexibility, as it can be applied to time series pairs of different lengths. This adaptability is a key

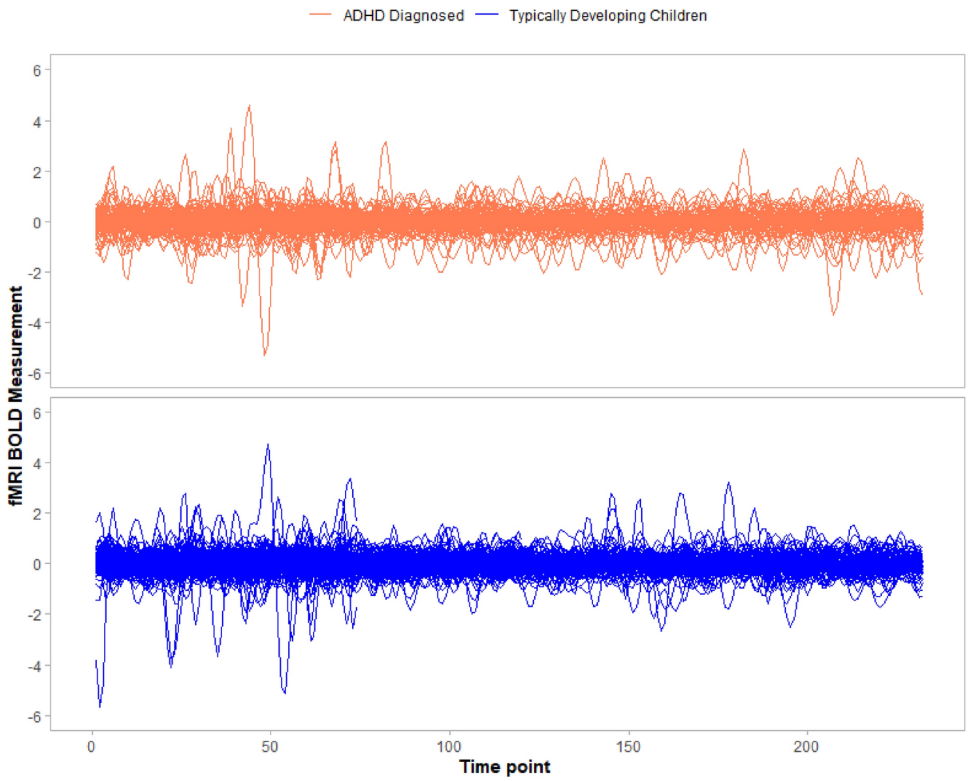


**Fig. 11**. BOLD signal fluctuations across different diagnostic groups demonstrating temporal patterns and variations in brain activity.

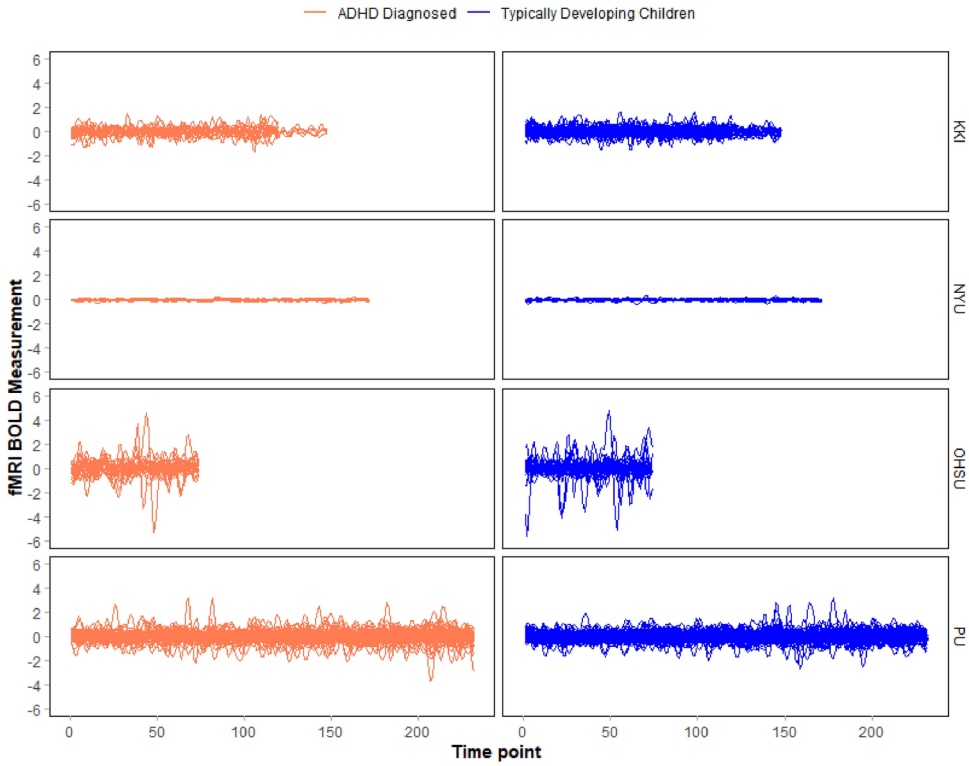


**Fig. 12**. BOLD signal fluctuations across different diagnostic groups and imaging sites. The plots demonstrate temporal variations, with potential diagnostic group differences visible in signal characteristics.

factor contributing to the robustness of our proposed method. To gain additional insights from our empirical experiments, we estimated statistical models for each subject's BOLD data to identify the most appropriate model structure. Figure 13 summarizes these findings, revealing that most subjects' data are best characterized by AR(1), AR(4), or AR(5) models. This distribution of model types closely aligns with the settings used in our simulation study. The results from our time series clustering analysis match the characteristics observed in the simulated datasets generated from AR(1) and AR(5) models, which typically produce clusters of roughly equal size. Figures 14 and 15 illustrate this pattern, showing the frequency of different cluster sizes and their distribution.

*Power and sample size experiments*

Given the unique characteristics of the ADHD-200 dataset, we chose to include all time series clustering results in our analysis rather than aggregating them. This approach allowed us to use the complete set of distinct clustering outputs for sample size calculations, applying kernel regression to estimate sample sizes from these varied results. The sample size calculation results are presented in Figures B9 and B10 in the Appendix, showing patterns that closely align with our findings from simulated data. Among the methods tested, the Clustering-based Assignment Method, which employs DTW-based time series clustering, most effectively characterizes the logical sample size requirements. This method's strength lies in its ability to establish a robust distance-based approximation of effect size while providing clear differentiation between sampling distributions. In contrast, the model-based assignment method, which uses the source model of the data as the grouping factor, consistently produced the highest sample size estimates among all approaches-even exceeding those from the random assignment method. This suggests that the source model of a sample does not necessarily serve as an effective distance-based discriminating factor when comparing samples from multiple model sources. Consequently, this translates to higher calculated sample size requirements when considering treatment effects. These patterns remained consistent across all significance levels and power values examined.

Our sample size calculation results also confirm classical statistical principles: higher desired power corresponds to larger required sample sizes, and lower significance levels necessitate larger sample sizes. As expected, we observed that larger effect sizes (calculated from the distance between sample groups) correspond to smaller required sample sizes. Among the three methods evaluated, the clustering-based assignment method generally yielded the highest effect sizes, indicating that it achieves the most effective grouping through its clustering approach.

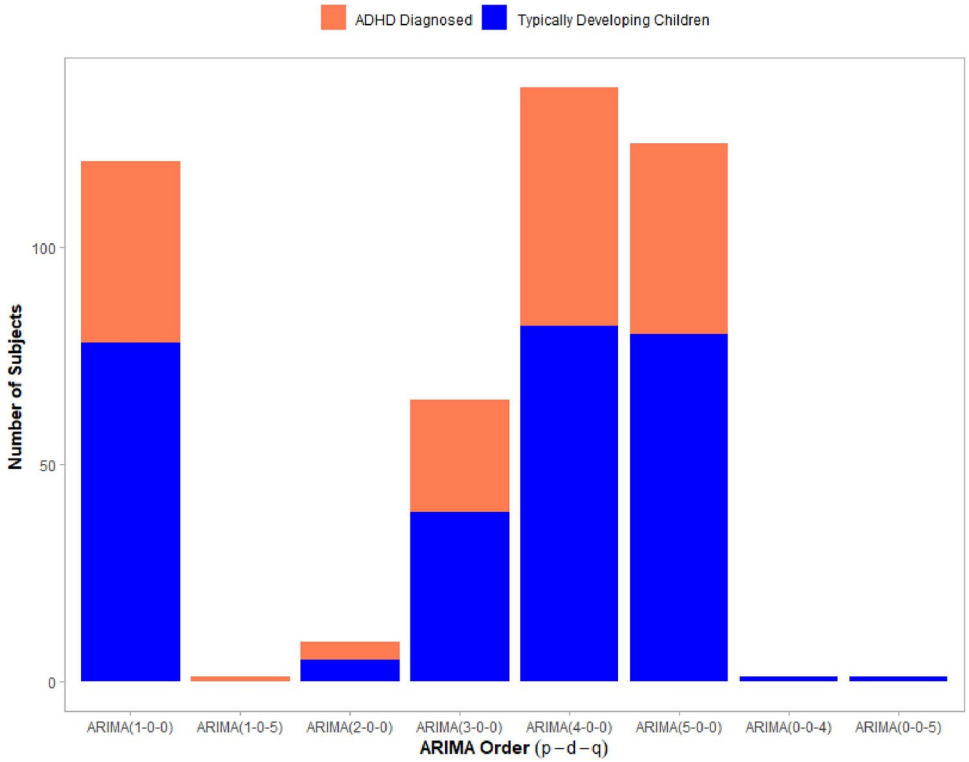


**Fig. 13**. Stacked bar plot showing the distribution of best-fitting ARIMA models across diagnostic groups. This visualization reveals that the majority of subjects' BOLD signals are best characterized by specific ARIMA models, particularly AR(1), AR(4), and AR(5), providing crucial insights for subsequent analysis approaches.

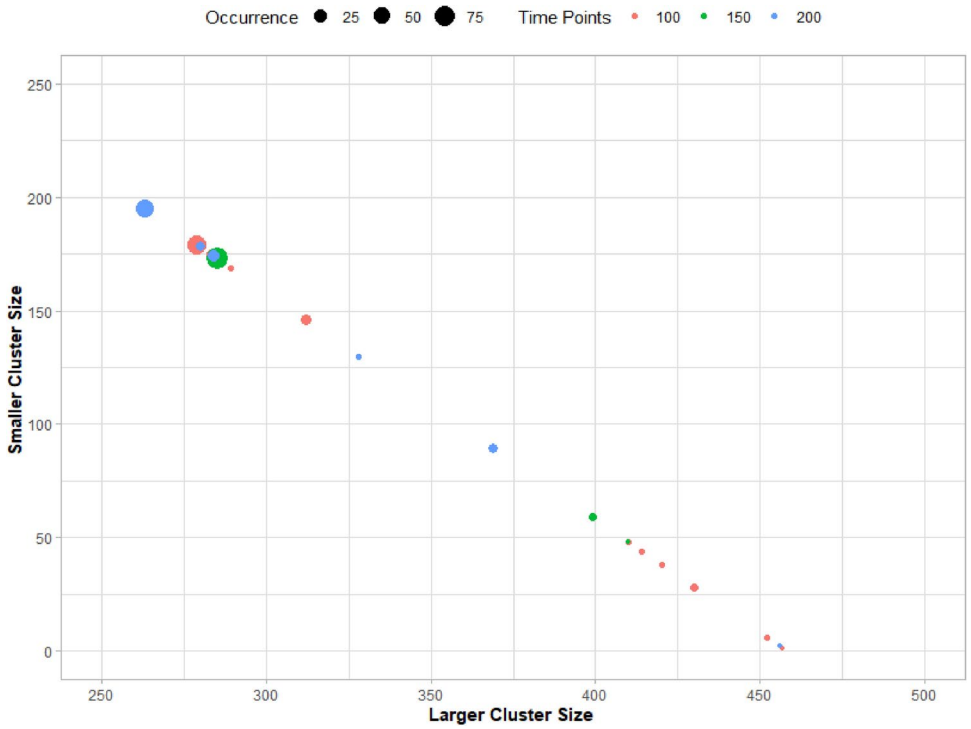


**Fig. 14**. ADHD-200 time series clustering results on cluster size: scatter plot displaying cluster size patterns from the ADHD-200 dataset analysis, with point sizes indicating occurrence frequency. Different colors represent different lengths of time series $T$, demonstrating how temporal resolution affects clustering outcomes in real neuroimaging data.

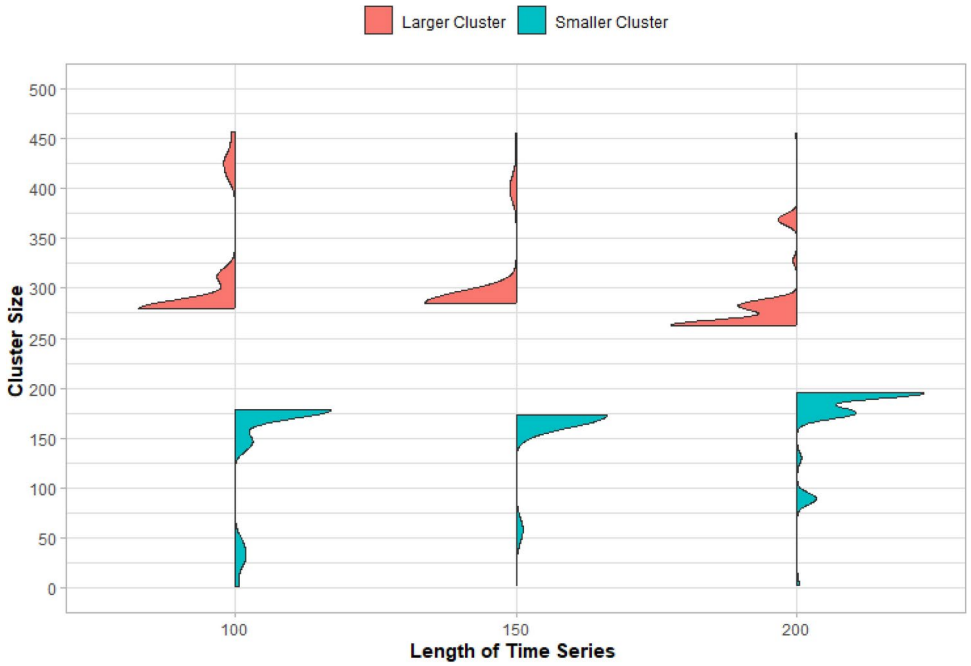


**Fig. 15**. Violin plot representation of cluster size distributions across different time points in the ADHD-200 Dataset. This visualization helps understand the stability and balance of cluster assignments across different temporal resolutions.

## Conclusions and future research

We proposed sample size determination methods based on time series clustering that compute distances between two groups of experimental units with high-dimensional and high-frequency time series responses. The efficiency of our algorithms in both time series clustering and sample size computation is demonstrated using simulated data and the ADHD-200 fMRI dataset.

By employing DTW distance in our time series clustering approach, we successfully partition samples into control and treatment groups. This partitioning allows us to compute a measure of treatment effect based on the ratio of quantities related to the differences in sum of squares of pairwise DTW distances between and within groups, along with their respective variances. This methodology effectively approximates the treatment effect that distinguishes the sampling distributions of responses between control and treatment groups.

Our simulations demonstrate that the method produces consistent outcomes that directly relate to the underlying data generating processes. Notably, we identified that time series autocorrelation is a critical characteristic influencing both clustering results and subsequent sample size calculations. High autocorrelation leads to a more balanced distribution of input samples and greater average distances between formed clusters. This translates to enhanced discrimination between groups, highlighting treatment effects more clearly and ultimately requiring smaller sample sizes to achieve desired statistical power.

The method has proven particularly robust when applied to real-world data. Using DTW to establish distance matrices among samples enables effective time series clustering even when time series responses vary in length. The warping capability of DTW resolves challenges associated with non-uniform time intervals in the data. This adaptability is especially valuable when working with real datasets such as BOLD measurements, which frequently exhibit these characteristics due to variations in measurement protocols and equipment settings across imaging sites-a concept analogous to covariates that can influence clinical trials. Additionally, our approach successfully addresses common challenges in high-frequency time series data, including response spikes and shifts that might otherwise compromise analysis.

Future studies can build upon our work by leveraging our estimated sample size requirements based on prior knowledge of data characteristics-specifically, by matching the model fit of experimental samples to our simulated datasets. However, these results should be interpreted with caution, as they are specifically applicable to datasets that conform to the models specified in our simulation design. Any deviation from these dataset definitions will naturally yield different outcomes in sample size calculation and power analysis. For cases where prior datasets do not resemble our simulated data, researchers can adapt our methodology by using their study's existing data from previous experiments or pre-collected samples. They can perform multiple iterations of time series clustering using DTW as the distance measure following our specified algorithm. Once these clustering results are available, researchers should optimize the selection of results for sample size calculation using our proposed methodologies, and then apply kernel regression to estimate the sample size requirements.

While our research has addressed a broad spectrum of analytical challenges related to high-dimensional and high-frequency time series clustering, several related problems remain to be resolved. Future work in this area will further enhance and extend the contributions presented in this study, potentially expanding the application of these methods to other domains and data types.

1. ***Vertical aggregation of time series data.*** One of the assumptions of the methodology is that each prior sample, or subject, has a single time series data to be used as input for the steps beginning from the time series clustering. The preprocessing of ADHD-200 data underwent vertical aggregation prior to time series clustering. The BOLD time series measure is extracted on a $V \times T$ voxel level information. To present this at the subject level, the BOLD measurements were averaged. The use of average may not be the most optimal way to perform vertical aggregation, especially for high-dimensional setup. A more robust method that preserves the inherent characteristics of the data and minimal loss of features will be needed.
2. ***Effects of varying time points $T$ of the time series data of observations.*** This study has made discussions and remarks about the robustness of the proposed method even for cases of varying $T$ in calculating sample size. One interesting problem to look for is the effects of the varying $T$ and its scales on the analysis of multiple time series. One potential practical application is on functional connectivity analysis.
3. ***Potential use of a subset of the time series data of observations.*** This pertains to the use of a non-uniform $T_i^* \subset T_i$ of each sample for analysis. This recommendation roots from the hypothesis that in clustering samples with time series data, the use of $T_i^*$ is a better discriminant feature than using full $T_i$ which may have noise or that $T_i - T_i^*$ shows inherent similarities across phenotype grouping. This has been demonstrated in this study where some subjects within the same group has non-harmonious BOLD measure.
4. ***Implementation of the proposed method for problems where hypothesized groups*** $G > 2$. This scenario is yet to be explored, but equally interesting. Although the proposed method can theoretically be extended into multigroup problem using similar univariate techniques, further evaluation is needed to understand the implementation behavior of the method.

## Data availability

The datasets (ADHD-200 fMRI) that stimulates the research problem of the current study are available in the NITRC: Neuroimaging Tools and Resources Collaboratory, https://www.nitrc.org.

Received: 10 December 2024; Accepted: 29 April 2025
Published online: 14 May 2025

## References


1. Van Den Heuvel, M. P. & Pol, H. Exploring the brain network: A review on resting-state fmri functional connectivity. *Eur. Neuropsychopharmacol.* **20**(8), 519–534 (2010).
2. Ogawa, S. et al. Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci.* **89**(13), 5951–5955 (1992).
3. Lindquist, M. A. The statistical analysis of fMRI data. *Stat. Sci.* **6**, 66 (2008).
4. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**(9), 793–795 (2015).
5. Biswal, B., Zerrin Yetkin, F., Haughton, V. M. & Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* **34**(4), 537–541 (1995).
6. Parkes, L., Satterthwaite, T. D. & Bassett, D. S. Towards precise resting-state fmri biomarkers in psychiatry: Synthesizing developments in transdiagnostic research, dimensional models of psychopathology, and normative neurodevelopment. *Curr. Opin. Neurobiol.* **65**, 120–128 (2020).
7. Mumford, J. A. & Nichols, T. E. Power calculation for group fmri studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage* **39**(1), 261–268 (2008).
8. Kassraian-Fard, P., Matthis, C., Balsters, J. H., Maathuis, M. H. & Wenderoth, N. Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Front. Psych.* **7**, 177 (2016).
9. Pulini, A. A., Kerr, W. T., Loo, S. K. & Lenartowicz, A. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**(2), 108–120 (2019).
10. Suckling, J. et al. Power calculations for multicenter imaging studies controlled by the false discovery rate. *Hum. Brain Mapp.* **31**(8), 1183–1195 (2010).
11. Mumford, J. A. A power calculation guide for fmri studies. *Soc. Cogn. Aff. Neurosci.* **7**(6), 738–742 (2012).
12. Linke, A. C. et al. Dynamic time warping outperforms Pearson correlation in detecting atypical functional connectivity in autism spectrum disorders. *Neuroimage* **223**, 117383 (2020).
13. Batista, G., Keogh, E., Tataw, O. & Souza, V. Cid: An efficient complexity-invariant distance for time series. *Data Min. Knowl. Discov.* https://doi.org/10.1007/s10618-013-0312-3 (2013).
14. Modak, S., Chattopadhyay, T. & Chattopadhyay, A. K. Unsupervised classification of eclipsing binary light curves through k-medoids clustering. *J. Appl. Stat.* **47**(2), 376–392. https://doi.org/10.1080/02664763.2019.1635574 (2020).
15. Aghabozorgi, S., Shirkhorshidi, A. S. & Wah, T. Y. Time-series clustering—A decade review. *Inf. Syst.* **53**, 16–38 (2015).
16. Cai, B., Huang, G., Samadiani, N., Li, G. & Chi, C.-H. Efficient time series clustering by minimizing dynamic time warping utilization. *IEEE Access* **9**, 46589–46599 (2021).
17. Gardner, W. A., Napolitano, A. & Paura, L. Cyclostationarity: Half a century of research. *Signal Process.* **86**(4), 639–697. https://doi.org/10.1016/j.sigpro.2005.06.016 (2006).
18. Mahmoudi, M. R., Maleki, M., Borodin, K., Pho, K.-H. & Baleanu, D. On comparing and clustering the spectral densities of several almost cyclostationary processes. *Alex. Eng. J.* **59**(4), 2555–2565. https://doi.org/10.1016/j.aej.2020.03.043 (2020).
19. Mahmoudi, M. R., Baleanu, D., Qasem, S. N., Mosavi, A. S. & Band, S. Fuzzy clustering to classify several time series models with fractional Brownian motion errors. *Alex. Eng. J.* **60**(1), 1137–1145. https://doi.org/10.1016/j.aej.2020.10.037 (2021).
20. Şuhubi, E. Metric spaces. *Funct. Anal.* **66**, 261–356 (2003).
21. Myers, C., Rabiner, L. & Rosenberg, A. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **28**(6), 623–635 (1980).
22. Itakura, F. Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust. Speech Signal Process.* **23**(1), 67–72 (1975).
23. Sakoe, H. Dynamic-programming approach to continuous speech recognition. In: *1971 Proceedings of the International Congress of Acoustics, Budapest* (1971)
24. Berndt, D. J. & Clifford, J. Using dynamic time warping to find patterns in time series. In: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* 359–370 (1994)
25. Cai, X., Xu, T., Yi, J., Huang, J. & Rajasekaran, S. DTWNet: A dynamic time warping network. *Adv. Neural Inf. Process. Syst.* **32**, 66 (2019).
26. Kate, R. J. Using dynamic time warping distances as features for improved time series classification. *Data Min. Knowl. Disc.* **30**, 283–312 (2016).
27. Jeong, Y.-S., Jeong, M. K. & Omitaomu, O. A. Weighted dynamic time warping for time series classification. *Pattern Recogn.* **44**(9), 2231–2240 (2011).
28. Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26**(1), 43–49 (1978).
29. Kuusi, E., et al. *Application of Dynamic Time Warping in Functional Magnetic Resonance Imaging*. Master's thesis, N/A (2016)
30. Basic, B. D. In: Lovric, M. (ed.) *Distance Measures* 397–398 (Springer, 2011). https://doi.org/10.1007/978-3-642-04898-2_626.
31. Keogh, E. & Ratanamahatana, C. A. Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* **7**, 358–386 (2005).
32. Wang, F.-Y., Zhang, J., Wei, Q., Zheng, X. & Li, L. PDP: Parallel dynamic programming. *IEEE/CAA J. Autom. Sin.* **4**(1), 1–5 (2017).
33. Steffen, P., Giegerich, R. & Giraud, M. GPU parallelization of algebraic dynamic programming. In: *Parallel Processing and Applied Mathematics: 8th International Conference, PPAM 2009, Wroclaw, Poland, September 13–16, 2009, Revised Selected Papers, Part II 8* 290–299 (Springer, 2010).
34. Lei, H. & Govindaraju, V. Direct image matching by dynamic warping. In: *2004 Conference on Computer Vision and Pattern Recognition Workshop* 76–76 (IEEE, 2004).
35. Modak, S. Book review: Finding groups in data: An introduction to cluster analysis. *J. Appl. Stat.* **51**(8), 1618–1620. https://doi.org/10.1080/02664763.2023.2220087 (2024).
36. Kaufman, L. & Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, 2005)
37. Schubert, E. & Rousseeuw, P. J. Fast and eager k-medoids clustering: O (k) runtime improvement of the pam, Clara, and Clarans algorithms. *Inf. Syst.* **101**, 101804 (2021).
38. Alekseyenko, A. V. Multivariate welch t-test on distances. *Bioinformatics* **32**(23), 3552–3558 (2016).
39. Craddock, C. *NITRC Neuro Bureau Athena Pipeline*. https://www.nitrc.org/plugins/mwiki/index.php?title=neurobureau:AthenaPipeline. Neuro Bureau Wiki—Athena Pipeline (2011).
40. Desmond, J. E. & Glover, G. H. Estimating sample size in functional mri (fmri) neuroimaging studies: Statistical power analyses. *J. Neurosci. Methods* **118**(2), 115–128 (2002).
41. Szucs, D. & Ioannidis, J. P. Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *Neuroimage* **221**, 117164 (2020).
42. ADHD-200, C. The ADHD-200 Consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Front. Syst. Neurosci.* **6**, 62 (2012).
43. Bellec, P. et al. The neuro bureau adhd-200 preprocessed repository. *Neuroimage* **144**, 275–286 (2017).
44. Craddock, C. *Scripts That Implement The Athena Pipeline for the ADHD-200 Preprocessed initiative*. https://github.com/preprocessed-connectomes-project/adhd200_athena_scripts. Github Repository for preprocessed-connectomes-project-adhd200-athena-scripts (2016).

45. Schaefer, A. et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cereb. Cortex* **28**(9), 3095–3114 (2018).
46. Yeo, B. T. et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **6**, 66 (2011).
47. Rieck, J. R., Baracchini, G., Nichol, D., Abdi, H. & Grady, C. L. Dataset of functional connectivity during cognitive control for an adult lifespan sample. *Data Brief* **39**, 107573 (2021).

## Author contributions

BR, EB, IG contributed in formulating the research problem; BR, EB, IG contributed in the empirical studies; BR, IG accessed ADHD data; BR prepared the initial manuscript; EB and IG revised the manuscript; HO and JL edited the manuscript; EB finalized the manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00554-w.

**Correspondence** and requests for materials should be addressed to E.B.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.