



**HAL**  
open science

## **Semi-supervised clustering of quaternion time series: Application to gait analysis in multiple sclerosis using motion sensor data**

Pierre Drouin, Aymeric Stamm, Laurent Chevreuil, Vincent Graillet, Laetitia Barbin, Pierre-antoine Gourraud, David-axel Laplaud, Lise Bellanger

### ► To cite this version:

Pierre Drouin, Aymeric Stamm, Laurent Chevreuil, Vincent Graillet, Laetitia Barbin, et al.. Semi-supervised clustering of quaternion time series: Application to gait analysis in multiple sclerosis using motion sensor data. *Statistics in Medicine*, 2022, pp.433-456. <10.1002/sim.9625>. <hal-03895883>

**HAL Id: hal-03895883**

**<https://hal.science/hal-03895883v1>**

Submitted on 24 May 2024



**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

# Semi-supervised clustering of quaternion time series: Application to gait analysis in multiple sclerosis using motion sensor data

Pierre Drouin<sup>1,2</sup>  | Aymeric Stamm<sup>1</sup>  | Laurent Chevreuil<sup>2</sup> | Vincent Graillet<sup>2</sup> | Laetitia Barbin<sup>3</sup> | Pierre-Antoine Gourraud<sup>4</sup> | David-Axel Laplaud<sup>3</sup> | Lise Bellanger<sup>1</sup>

<sup>1</sup>Laboratoire de Mathématiques Jean Leray, Université de Nantes, Nantes, France

<sup>2</sup>UmanIT, Nantes, France

<sup>3</sup>CRTI-Inserm U1064, CIC, Service de Neurologie, CHU et Université de Nantes, Nantes, France

<sup>4</sup>Centre de Recherche en Transplantation et Immunologie, UMR 1064, ATIP-Avenir, Université de Nantes, CHU de Nantes, INSERM, Nantes, France

## Correspondence

Pierre Drouin, UmanIT, 13 place Sophier Trébuchet, 44000 Nantes, France.  
Email: [pdrouin@umanit.fr](mailto:pdrouin@umanit.fr)

## Funding information

Fondation pour l'Aide à la Recherche sur la Sclérose en Plaques

Recent approaches in gait analysis involve the use of wearable motion sensors to extract spatio-temporal parameters that characterize multiple aspects of an individual's gait. In particular, the medical community could largely benefit from this type of devices as they could provide the clinicians with a valuable tool for assessing gait impairment. Motion sensor data are however complex and there is an urgent unmet need to develop sound statistical methods for analyzing such data and extracting clinically relevant information. In this article, we measure gait by following the hip rotation over time and the resulting statistical unit is a time series of unit quaternions. We explore the possibility to form groups of patients with similar walking impairment by taking into account their walking data and their global disease severity with semi-supervised clustering. We generalize a compromise-based method named `hclustcomprom` to unit quaternion time series by combining it with the proper dissimilarity quaternion dynamic time warping. We apply this method on patients diagnosed with multiple sclerosis to form groups of patients with similar walking deficiencies while accounting for the clinical assessment of their overall disability. We also compare the compromise-based clustering approach with the method `mergeTrees` that falls into a sub-class of ensemble clustering named collaborative clustering. The results provide a first proof of both the interest of using wearable motion sensors for assessing gait impairment and the use of prior knowledge to guide the clustering process. It also demonstrates that compromise-based clustering is a more appropriate approach in this context.

## KEYWORDS

human gait analysis, quaternion time series, semi-supervised clustering, wearable sensors

## 1 | INTRODUCTION

Recent improvements of wearable sensing devices make it possible to monitor many aspects of the human gait.<sup>1</sup> Most of the time, such devices measure the variation of spatio-temporal parameters during time in the form of time series (TS). There has been a growing interest in studying the rotation and/or orientation in 3D space of body segments in particular for individual gait recognition<sup>2,3</sup> and for local stability.<sup>4-6</sup> These data come in the form of time series of unit quaternions (QTS). The hypothesis is that it might be a valuable tool for the quantitative assessment of symptomatic gait impairment. In the context of neurological disorders such as multiple sclerosis (MS), this would be key for improving follow-up and tailoring patient's management to prevent or slow down future degradation.<sup>7</sup>

A first step in this direction pertains to identifying groups of patients with similar gait impairment from the sensor data. This is known as unsupervised learning or clustering. A key ingredient for grouping individuals on the basis of their walking pattern is the definition of a dissimilarity between time series. Dynamic time warping (DTW) is a well-known approach for evaluating how far the shapes of two time series are from each other<sup>8</sup> which accounts for shape misalignment. Several studies that compared walking data achieved TS unsupervised learning via hierarchical agglomerative clustering (HAC) on a distance matrix obtained from DTW. Baghdadi et al<sup>9</sup> used this approach to study workers' fatigue monitored with wearable devices clipped to the ankle. Pullido-Valdeolivas et al<sup>10</sup> formed groups of patients with pediatric hereditary spastic paraplegia to discover gait phenotype with HAC on kinematic TS representing several gait features. Steinmetzer et al<sup>11</sup> clustered patients diagnosed with Parkinson disease comparing their gait data measured by insoles equipped with accelerometer. The resulting clusters tend to regroup patients with similar disease burden. Jablonski<sup>12</sup> extended DTW for QTS in a method called quaternion dynamic time warping (QDTW). These works therefore provide us with a suitable dissimilarity between QTS which we can use for clustering individuals based on their gait pattern. HAC is often more appealing than partition-based clustering (eg, *k*-means) for finding a suitable grouping structure with complex data because it provides a hierarchy of clustering structures in the form of a dendrogram which does not require running the algorithm for each possible number of clusters. In this article, since we are working with QTS, we therefore focus on HAC methods suitable for QTS.

In medical applications, additional information on the patients are often available thanks to the clinicians who can provide an expert assessment of the patient's condition. For instance, patients with neurodegenerative diseases (eg, MS) often suffer from progressive disabilities that are measured on specific scales. This represents an additional source of information that needs to be accounted for when clustering patients with similar gait impairment. In effect, it has been shown that using supplementary information when available improves the quality and/or interpretability of the clustering structure.<sup>13</sup> Methods that account for supplementary information are called semi-supervised clustering methods. They can mainly be divided into three categories (constraint-based methods, ensemble methods and compromise-based methods) depending on how the supplementary information is used within the clustering algorithm. Constraint-based clustering methods integrate supplementary information either by forcing a priori some observations to be grouped together<sup>14</sup> or by constraining final clusters to have a specific structure. These approaches are best suited for partition-based clustering methods<sup>13</sup> and hardly applicable to hierarchical methods<sup>15</sup> even though some attempts can be found in the literature.<sup>16-18</sup> A fairly strong prior knowledge of the clustering structure is necessary to determine the constraints, which is not always possible depending on the context of the study. Ensemble clustering methods are designed to generate clusters from multiple clustering structures obtained on the same individuals.<sup>19</sup> The input clustering structures may be achieved on the same data source by different clustering algorithms and/or on different data sources provided that they are measured on the same individuals. These methods are straightforwardly applicable to hierarchical methods and allow the use of a virtually unlimited number of data sources. However, by construction, the resulting clusters represent the best agreement between the initial multi-source clusters and does not weight their relative importance. Compromise-based clustering methods do not enforce any constraints but rather use the supplementary information to measure how far two observations are from each other on the basis of a compromise between the main and supplementary sources of information.

When assessing walking disability, the supplementary source of information is the score attributed by the clinician based on the overall disability of the patient. Such scores are often ordinal rating scales and are not based only on gait disability. Defining strict constraints from this information is then not possible as two patients with the same score may be differently affected in their gait. Constraint-based clustering is therefore not applicable in this context. We hypothesize that weighting the relative importance of each information sources with compromise-based clustering is the most appropriate approach when accounting for existing overall disability scores in the process of clustering individuals based on their gait patterns. The literature is thin when it comes to TS clustering with prior knowledge. Only constraint-based methods have been adapted for TS<sup>14,20</sup> and none are suitable for QTS clustering. We therefore aim at elaborating a compromise-based

clustering method for QTS data using QDTW as dissimilarity. Ma and Dhavala<sup>21</sup> proposed a compromise-based HAC method that leverages supplementary information that comes in the form of a dendrogram called ontological dendrogram. The dissimilarity between observations that they use to build the hierarchy of clustering structures is penalized by an ultrametric dissimilarity derived from the ontological dendrogram. This penalty is determined either by maximizing an internal cluster quality measure (such as Dunn or David-Bouldin index) or by cross-validation if external labels are available. This implies that the penalty is optimized for a given number of clusters. As a result, it boils down to a bi-dimensional optimization (on both the penalty coming from the external information and the number of clusters). In this way, the supplementary information plays an important role in the determination of the optimal number of clusters. Bellanger et al<sup>22</sup> proposed another compromise-based HAC method originally developed to answer chronological problems in archeology based on artifact data. It is designed for data sets in which observations are described by two potentially error-prone sources of information, corresponding to the main and supplementary sources respectively. Normalized dissimilarity matrices are first computed for each source. A HAC algorithm is then applied on a dissimilarity matrix computed as a convex combination of the two normalized dissimilarity matrices. As in the method proposed by Ma and Dhavala,<sup>21</sup> the weighting parameter in the convex combination determines the amount of supplementary information that is used to modify the final dissimilarity matrix. However, its determination does not involve maximizing cluster quality indices and is therefore independent from the choice of the number of clusters. This allows to leverage external information to obtain a compromise-based hierarchy of clustering structures (dendrogram). The method is also publicly available as the `hclustcompro` function in the SPARTAAS R package.<sup>23</sup>

Our contributions are two-fold: (i) to propose a generalization of the compromise-based method `hclustcompro` for accommodating QTS data and (ii) to compare it with a more classic unsupervised HAC approach which does not use supplementary information and with an ensemble-clustering approach named `mergeTrees`, using a case study of patients diagnosed with MS in which the main source is their gait data and the supplementary source is a clinical score of overall disability. We will subsequently describe and discuss the clustering structure established by the proposed compromise-based method in the light of patients' monitoring and follow-up. The article is outlined as follows. Section 2 introduces the semi-supervised clustering framework for QTS. Section 3 describes the collected data and the methods used in the application of the clustering approaches. Results are detailed in Section 4 followed by a discussion about the benefits of both using additional information and using a compromised-based approach in Section 5.

## 2 | A SEMI-SUPERVISED CLUSTERING FRAMEWORK FOR UNIT QUATERNION TIME SERIES

### 2.1 | General presentation of hierarchical clustering

Clustering relates to the process of building a partition of a set of  $n$  observations  $Q_1, \dots, Q_n$  such that similar observations ends up into the same cluster while observations too far from each other are assigned to separate clusters. Among the different approaches, hierarchical clustering (HC) presents the advantage to be directly applicable on the  $n \times n$  matrix  $D$  of pairwise dissimilarities between observations, that is,  $d_{ij} := d(Q_i, Q_j)$  for any  $i, j \in \{1, \dots, n\}$ . HC produces a binary tree called dendrogram (and denoted by  $\mathcal{T}$ ) whose leaves are the single observations. Each node of the tree represents a cluster that regroups all observations from the branches below it. Two main approaches can be adopted to generate the dendrogram. The divisive approach pertains to considering that all observations belong to a single cluster at the initial state. At each iteration, a cluster is chosen to be further splitted into two clusters. This iterative process stops when each observation ends up its own cluster. The agglomerative approach pertains to considering that each observation is in its own cluster at the initial state. At each iteration, the two closest clusters are merged together. This iterative process stops when all observations end up into the same cluster. Hierarchical divisive clustering requires more input choices than its agglomerative counterpart. In effect, in addition to the dissimilarity (which is the only ingredient that HAC requires), in a divisive strategy, one needs to define a criterion to select at each iteration which cluster will be further splitted. In the following, we will therefore focus on hierarchical agglomerative clustering (HAC). Let then set up the initial state in which each observation is its own cluster all clusters contains a single observation:  $C_i = \{Q_i\}, \forall i \in \{1, \dots, n\}$ . The first step for building the dendrogram is to determine the two closest clusters  $C_i$  and  $C_j$  of the set that will be merged into a single cluster  $C_{ij}$ . This is achieved by  $C_{ij} = C_i \cup C_j$ , with  $(C_i, C_j) = \operatorname{argmin}_{C_k, C_l} D(C_k, C_l)$ . The next step is to update the matrix  $D$  with the new dissimilarity value between the new cluster  $C_{ij}$  and every other cluster  $C_k, \forall k \in \{1, \dots, n\} \setminus \{i, j\}$ . Several linkage criteria have been described in the literature to determine how to compute the dissimilarity between clusters. They have been unified by the Lance and Williams recurrence formula:<sup>24</sup>

$$D(C_{ij}, C_k) = \alpha_1 D(C_i, C_k) + \alpha_2 D(C_j, C_k) + \beta D(C_i, C_j) + \gamma \left| D(C_i, C_k) - D(C_j, C_k) \right|, \quad (1)$$

where  $\alpha_1$ ,  $\alpha_2$ ,  $\beta$ , and  $\gamma$  are real scalar numbers. Linkage criteria can be divided into two categories. Geometric criteria (eg, Ward, centroid, median) assume that observations belong to a Euclidean space and the Euclidean distance is therefore implicitly assumed for measuring proximity of observations. Graph criteria compute dissimilarity between two clusters only from the dissimilarities between observations included in these clusters.<sup>25</sup> This property makes them the only candidates suitable for accommodating non-Euclidean dissimilarity matrices. The three most common graph linkage criteria are single linkage, complete linkage and average linkage. Coefficients in the Lance and Williams formula in Equation (1) to achieve each one of them are given in Table 1, where  $|C_i|$  denotes the number of observations in cluster  $C_i$ .

When observations are QTS, a proper dissimilarity measure needs to be used to compute  $D$ . Section 2.2 will now provide a brief introduction to the unit quaternion algebra which will set up the necessary tools to describe, in Section 2.3, a DTW algorithm adapted to compute dissimilarity between two QTS.

## 2.2 | Unit quaternion algebra

Quaternions are hypercomplex numbers of rank 4 and are denoted as follows:

$$\mathbf{q} = (w, x, y, z)^\top = w + ix + jy + kz \in \mathbb{R}^4, \quad (2)$$

where  $i$ ,  $j$ , and  $k$  generalize the single imaginary number  $i$  using the following rule  $i^2 = j^2 = k^2 = ijk = -1$ . This rule implies in particular that the product of two quaternions, called the Hamilton product, is not commutative. The quaternion algebra  $\mathbb{H}$  is isomorphic to  $\mathbb{R}^4$ . Similarly to complex numbers, we can define a real and an imaginary part for a quaternion. In details, we have:

$$\text{Re}(\mathbf{q}) = w \in \mathbb{R} \quad \text{and} \quad \text{Im}(\mathbf{q}) = (x, y, z)^\top \in \mathbb{R}^3.$$

1. *Representing 3D rotations.* The set of unit quaternions  $\mathbb{H}_u := \{\mathbf{q} \in \mathbb{H}; \|\mathbf{q}\| = 1\}$  forms a Lie group which is isomorphic to the special unitary group  $SU(2)$  which is exactly twice as big as the rotation group  $SO(3)$  of 3-dimensional rotation matrices. In effect, unit quaternions represent rotations in 3-dimensional space and both  $\mathbf{q}$  and  $-\mathbf{q}$  encode the same rotation. The unit quaternion algebra can therefore be seen as a particular group structure on the 3-sphere<sup>12</sup>  $S^3 \subset \mathbb{R}^4$ .

A somehow more natural representation of a rotation in 3-dimensional space is provided by its rotation angle  $\theta \in \mathbb{R}$  and its axis of rotation  $\mathbf{u} \in S^2$ , where  $S^2$  is the 2-sphere. There is a relationship between the angle-axis and quaternion representations of a rotation. In details, the unit quaternion  $\mathbf{q}$  representing the direct rotation of angle  $\theta$  around the unit vector  $\mathbf{u}$  is given by:

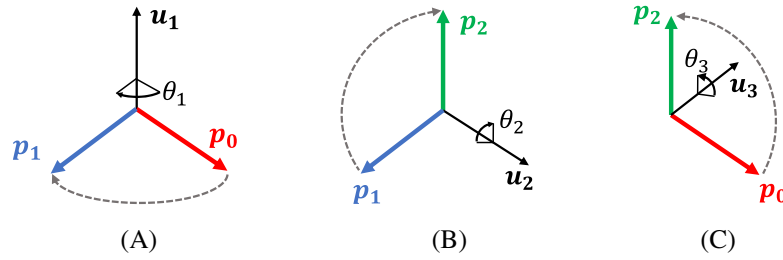
$$\mathbf{q} = \left( \cos \frac{\theta}{2}, u_x \sin \frac{\theta}{2}, u_y \sin \frac{\theta}{2}, u_z \sin \frac{\theta}{2} \right)^\top. \quad (3)$$

2. *Defining a proper distance for quaternions.* Just like with complex numbers, we can define the conjugation of a quaternion. The conjugate quaternion  $\bar{\mathbf{q}}$  of a quaternion  $\mathbf{q} = (w, x, y, z)^\top$  is the same quaternion with opposite imaginary part:  $\bar{\mathbf{q}} = (w, -x, -y, -z)^\top$ . When dealing with unit quaternions, the conjugate quaternion equals the inverse quaternion which reads  $\mathbf{q}^{-1} = (w, -x, -y, -z)^\top$ . It can then be easily proven by resorting to the angle-axis representation that  $\mathbf{q}^{-1}$  rotates around the same axis as  $\mathbf{q}$  but with an opposite angle of  $-\theta$ .

We can also endow  $\mathbb{H}_u$  with a proper metric. While we could use the Euclidean distance in  $\mathbb{R}^4$ , this would define a metric space that is not closed with respect to the quaternion algebra. We instead introduce the geodesic distance

TABLE 1 Lance and Williams coefficients for graph linkage criteria

Linkage criterion	$\alpha_1$	$\alpha_2$	$\beta$	$\gamma$
Single	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete	$\frac{1}{2}$	$\frac{1}{2}$	0	$+\frac{1}{2}$
Average	$\frac{ C_i }{ C_i + C_j }$	$\frac{ C_j }{ C_i + C_j }$	0	0



**FIGURE 1** Composition of rotations by unit quaternion multiplication. (A) Rotation of  $\mathbf{p}_0$  by  $\mathbf{q}_1$ . (B) Rotation of  $\mathbf{p}_1$  by  $\mathbf{q}_2$ . (C) Rotation of  $\mathbf{p}_0$  by  $\mathbf{q}_3 = \mathbf{q}_2\mathbf{q}_1$

between two rotations  $\mathbf{q}_1$  and  $\mathbf{q}_2$ , which reads:<sup>26</sup>

$$d(\mathbf{q}_1, \mathbf{q}_2) = 2 \arccos \operatorname{Re}(\mathbf{q}_1^{-1}\mathbf{q}_2). \quad (4)$$

It corresponds to the minimum length of a geodesic line connecting the two quaternions on the 3-sphere.<sup>12</sup> It can also be interpreted as the angle of the necessary rotation to obtain  $\mathbf{q}_2$  from  $\mathbf{q}_1$ .

3. *Composing rotations with quaternions.* Let  $\mathbf{p}_0 \in \mathbb{R}^3$  be a 3-dimensional point. The rotation encoded by  $\mathbf{q}_1$  sends  $\mathbf{p}_0$  into a new point  $\mathbf{p}_1$  given by:

$$\begin{pmatrix} 0 \\ \mathbf{p}_1 \end{pmatrix} = \mathbf{q}_1 \begin{pmatrix} 0 \\ \mathbf{p}_0 \end{pmatrix} \mathbf{q}_1^{-1}. \quad (5)$$

The rotation encoded by  $\mathbf{q}_1$  is illustrated in Figure 1A using its angle-axis representation. Similarly, Figure 1B illustrates the effect of applying another rotation  $\mathbf{q}_2$  to  $\mathbf{p}_1$  which sends it into a new point  $\mathbf{p}_2$ . Following Equation (5), we can write:

$$\begin{pmatrix} 0 \\ \mathbf{p}_2 \end{pmatrix} = \mathbf{q}_2 \begin{pmatrix} 0 \\ \mathbf{p}_1 \end{pmatrix} \mathbf{q}_2^{-1} = \mathbf{q}_2 \left( \mathbf{q}_1 \begin{pmatrix} 0 \\ \mathbf{p}_0 \end{pmatrix} \mathbf{q}_1^{-1} \right) \mathbf{q}_2^{-1} = (\mathbf{q}_2\mathbf{q}_1) \begin{pmatrix} 0 \\ \mathbf{p}_0 \end{pmatrix} (\mathbf{q}_2\mathbf{q}_1)^{-1}. \quad (6)$$

Equation (6) shows that the Hamilton product of two unit quaternions  $\mathbf{q}_3 = \mathbf{q}_2\mathbf{q}_1$  is equivalent to applying the rotation  $\mathbf{q}_1$  followed by the rotation  $\mathbf{q}_2$ . This is illustrated in Figure 1C using the angle-axis representation of  $\mathbf{q}_3$ .

In this article, we analyze unit quaternion time series (QTS). This means that the statistical unit is an ordered set denoted  $i$  of unit quaternions representing consecutive 3D rotations on a time grid  $t_{i,1}, \dots, t_{i,N_i}$ . We will use the following notation for a unit QTS:

$$Q_i = (\mathbf{q}_{i,1}, \dots, \mathbf{q}_{i,N_i}).$$

### 2.3 | Dissimilarity measure for QTS: QDTW

For a given representation of the data, we might have multiple choices for the dissimilarity measure. This effectively defines a metric space from which we can perform statistical analyses. The choice of the dissimilarity, together with the choice of the data representation, will obviously have a large impact on the results emanating from these statistical analyses. In this article, the data are 3D rotations over time which we represent as QTS and we choose to focus on shape dissimilarity only.

When comparing shapes between time series, a common issue is time shifting<sup>27</sup> which refers to these situations in which two similar events are observed but at a different time in different TS. This can arise in our case for events such as the landing of the foot on the ground or the take-off of the foot from the ground which has led to fluctuations in the reported duration of gait cycles (GC) in the literature.<sup>28</sup> Time shifting is classically addressed by integrating an alignment

in time as part of the dissimilarity measure. These metrics are known as elastic metrics. The most common one in shape analysis of time series is the DTW dissimilarity.<sup>8</sup> It has been widely used in the field of clinical gait analysis to compare the shape of walking data measured by optical device<sup>10,12,29</sup> and wearable sensors.<sup>9,11,30,31</sup>

DTW determine the optimal nonlinear alignment between the elements of two time series  $Q_i = (\mathbf{q}_{i,1}, \dots, \mathbf{q}_{i,N_i})$  of size  $N_i$  and  $Q_j = (\mathbf{q}_{j,1}, \dots, \mathbf{q}_{j,N_j})$  of size  $N_j$  using the following the following formula:<sup>32</sup>

$$\text{DTW}(Q_i, Q_j) = \begin{cases} 0, & \text{if } N_i = N_j = 0, \\ \infty, & \text{if } N_i = 0 \text{ or } N_j = 0, \\ c(\mathbf{q}_{i,N_i}, \mathbf{q}_{j,N_j}) + \min \begin{cases} \text{DTW}(Q_{i,1:(N_i-1)}, Q_{j,1:(N_j-1)}), \\ \text{DTW}(Q_{i,1:N_i}, Q_{j,1:(N_j-1)}), \\ \text{DTW}(Q_{i,1:(N_i-1)}, Q_{j,1:N_j}), \end{cases} \end{cases} \quad (7)$$

with  $c(\cdot, \cdot)$  a local cost function representing the distance between two elements and  $Q_{i,1:\ell}$  the subsequence  $(\mathbf{q}_{i,1}, \dots, \mathbf{q}_{i,\ell})$ , for  $1 \leq \ell \leq N_i$ .

Jablonski<sup>12</sup> generalizes the method for QTS using the geodesic distance defined in Equation (4) as the local cost  $c(\cdot, \cdot)$  in Equation (7) and coins the resulting DTW dissimilarity the ‘‘QDTW.’’ The author presents clustering results both on simulated data and on the MoCap database HDM05,<sup>33</sup> obtained by applying a HAC method using QDTW as the dissimilarity metric. The MoCap database HDM05 stores Human motion data already distributed into known groups. The simulated data was as well generated with a ground truth clustering structure. They could then in both situations assess the performance of the QDTW-based HAC by evaluating how well it was able to retrieve the true grouping structures. In both cases, the method proved very efficient with slightly better results for the simulated data. To the best of our knowledge, QDTW is the only measure suitable for assessing the dissimilarity between QTS that have been described in the literature. This leads us to consider QDTW as a good candidate dissimilarity for clustering QTS.

*Remark 1.* Jablonski<sup>12</sup> proposed a more complex dissimilarity as well, named QDTWFull. This measure takes into account not only the geodesic distance between two quaternions but also between their first and second derivatives with respect to time. However, their conclusion is that this dissimilarity did not clearly outperform the simpler QDTW on real data sets. Hence, we decided to use QDTW in the present study.

## 2.4 | Semi-supervised clustering for unit QTS with `hclustcompro`

The QDTW dissimilarity leads to the computation of a matrix of pairwise dissimilarities between observations in a sample of QTS. This means that we can perform HAC provided that we use graph linkage criteria only. In this section, we will describe a compromise-based approach to integrate supplementary information about the observations when available.

The semi-supervised hierarchical approach named `hclustcompro`, first described in Bellanger et al,<sup>22</sup> is a clustering method adapted to cases where  $n$  observations are represented by a main and a supplementary source of information whose data type may differ. Let  $D_1$  and  $D_0$  be the two  $n \times n$  matrices storing the normalized dissimilarities between observations in the main feature space and in the supplementary feature space respectively. The principle of `hclustcompro` is to apply a HAC method to a dissimilarity matrix obtained by the following convex combination:<sup>22</sup>

$$D_\alpha = \alpha D_1 + (1 - \alpha) D_0, \quad (8)$$

where  $\alpha \in [0, 1]$  is a fixed parameter weighting the contribution of each data source to the dissimilarity matrix  $D_\alpha$ . The HAC performed on  $D_\alpha$  outputs a dendrogram  $\mathcal{T}_\alpha$  which is strongly dependent upon the choice of the weighting parameter  $\alpha$ . The determination of  $\alpha$  is carried out by maximizing a criterion which is inspired by the cophenetic correlation proposed by Sokal and Rohlf.<sup>34</sup> This is achieved by the following steps:

1. Compute the cophenetic matrix  $D_\alpha^{\text{coph}}$  from the dendrogram  $\mathcal{T}_\alpha$ . This is an  $n \times n$  matrix in which each element stores the height at which the two corresponding observations has become members of the same cluster in  $\mathcal{T}_\alpha$ .

2. Compute the correlations  $\text{Cor}(D_\alpha^{\text{coph}}, D_1)$  and  $\text{Cor}(D_\alpha^{\text{coph}}, D_0)$  between each initial dissimilarity matrix ( $D_1$  and  $D_0$ ) and the cophenetic matrix  $D_\alpha^{\text{coph}}$ . These two correlations measure how faithfully the pairwise dissimilarities between the observations in each data source are preserved in the final dendrogram.
3. Minimize over  $\alpha$  the following suitability index:

$$\text{CorCrit}(\alpha) := \left| \text{Cor}(D_\alpha^{\text{coph}}, D_1) - \text{Cor}(D_\alpha^{\text{coph}}, D_0) \right|. \quad (9)$$

This criterion represents the difference in absolute value between two correlations that inform about the extent to which the hierarchical structure of  $\mathcal{T}_\alpha$  well represents  $D_1$  and  $D_0$ .

In order to balance the weight of  $D_1$  and  $D_0$  in the final clustering, the value of the weighting coefficient  $\alpha$  is therefore determined by minimizing the following objective function:

$$\hat{\alpha} = \underset{\alpha}{\text{arg min}} \text{CorCrit}(\alpha). \quad (10)$$

A final HAC can be achieved on  $D_{\hat{\alpha}}$  with one of the graph linkage criteria listed in Table 1. When the main data source is a sample of QTS, the matrix  $D_1$  can be obtained using QDTW to compute the pairwise dissimilarities, which effectively generalizes `hclustcompro` to QTS-valued data. The choice of the dissimilarity measure to compute  $D_0$  depends upon the data type of the supplementary data source. In compromise-based clustering, the supplementary data source is therefore used to define and estimate the final dissimilarities between observation.

In the following section, we will compare traditional HAC with no supplementary data source, `mergeTrees` and `hclustcompro` in the context of gait analysis in MS.

### 3 | CLUSTERING GAIT DATA MEASURED WITH A MOTION SENSOR IN MULTIPLE SCLEROSIS

Assessing gait impairment is critical in MS because it has been reported by the patients to be the criterion that most negatively impacts their quality of life.<sup>35</sup> Several scores are already used by clinicians to evaluate the overall disability of patients, including walking disabilities. The use of wearable motion sensors to complete these scores with quantitative measurements of gait is a dynamic and promising field of research.<sup>7</sup> In this study, gait is represented as the 3D rotation of the hip over time measured by a motion sensor clipped on the belt. We compare the three clustering methods, namely HAC, `mergeTrees`, and `hclustcompro`. The main data source is a sample of QTS representing the hip rotation over time. The supplementary data source (used in `mergeTrees` and `hclustcompro`) is a sample of overall disability scores for the same patients.

The data acquisition and their pretreatment process are presented in the next section following by the description of the experimental design.

#### 3.1 | Data acquisition and pretreatment

Two sources of information were collected on the patients:

- *Walking data* in the form of time series of unit quaternions was measured during the Timed 25 Foot Walk (T25FW) test using a wearable motion sensor (Section 3.1.1);
- *Clinical data* as a score on the Expanded Disability Status Scale (EDSS) which was determined by a neurologist to assess the overall disability of the patient (Section 3.1.2).

##### 3.1.1 | Walking data

During the regular clinical evaluation of MS patients, there is an existing test dedicated to assessing gait impairment. It is called the T25FW in which the instruction is to perform a 25-foot walk as fast as possible without running. Patients

are asked to perform twice this walking distance under these instructions.<sup>36</sup> The average time the patient takes to walk 25 feet is then computed. The T25FW test is considered as an ideal primary endpoint for assessing walking disabilities in MS.<sup>7</sup> We aim at providing the neurologist with a richer information about the current walking abilities of a patient.

Ambulatory characteristics of the patient can be assessed during an instrumental examination called clinical gait analysis. This method pertains to quantifying several aspects of the movements performed during a walking pattern coined a GC.<sup>37</sup> A GC is defined as the sequence of movements performed by the body during a time period delimited by two consecutive contacts of a given foot with the ground. Gait analysis can be divided into the following steps:

1. Data representing the movements of the patients during walk are measured using devices designed for Human motion analysis.<sup>38</sup>
2. Identify important gait events such as the beginning and the end of the contact of the foot with the floor in the data.<sup>39</sup> The segmentation of the signal according to these time points produces a set of subsequences representing the movements of the patient during each GC.
3. The kinematic of the patient's gait, for example, the angle of the lower limb joints during the GC, can also be analyzed.<sup>37,39</sup> For this purpose, a single sequence representing the patient's gait is obtained by averaging the set of his/her GCs.<sup>28</sup> However, it is known that the semi-periodicity of walking leads to variability in the duration of walking cycles and in the occurrence of events such as the contact of the foot with the ground or its detachment. Thus, averaging the GCs require their temporal alignment to handle these within and between-GC temporal differences.<sup>40</sup>
4. Finally, most gait analysis methods described in the literature represent GCs in percentage of their total duration to facilitate the comparison between individuals with different velocity.<sup>28,40-44</sup>

We hereby describe the gait of an individual as the average rotation of the hip during a typical gait cycle. Hip mobility has been shown to be representative of the gait<sup>28</sup> and correlated with MS severity.<sup>37</sup> It also presents the benefit that we can measure this motion by clipping a sensor on the belt which is far less invasive with respect to optoelectronics measurement systems that required dedicated laboratory<sup>38</sup> and to monitoring lower parts of the body with motion sensors striped on the patient limbs.<sup>39</sup> Specifically, we opted for a 9-axis inertial measurement unit (IMU) called MetaMotionR (MMR) from Mbientlab to measure walking data. This device can determine its orientation in space over time by fusing data collected from a 3-axis accelerometer, a 3-axis gyroscope and a 3-axis magnetometer using the sensor fusion algorithm developed by Bosch Sensortec. In details, the sensor defines its orientation at a given point in time as the rotation that brings a fixed frame (aligned with the Earth's coordinate system)  $\mathfrak{R}_f = (\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)$  onto the sensor frame  $\mathfrak{R}_s = (\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3)$ . This rotation is described in Figure 2A by two parameters: its axis of rotation  $\mathbf{u} = (u_1, u_2, u_3)^\top$  and its angle of rotation  $\theta$ . The sensor logs its orientation in the form of a unit quaternion computed from these two parameters via Equation (3). The IMU is worn on the belt (see Figure 2B). The sensor logs its orientation at a frequency of 100 Hertz which effectively boils down to a time series of unit quaternions, each quaternion encoding the rotation between the fixed frame  $\mathfrak{R}_f$  and the sensor frame  $\mathfrak{R}_s$  at a given time  $t$ . After the identification of the GCs delimited by two consecutive contacts of the right foot with the floor in the IMU data, we temporally align the obtained segments. In order to do that, we provide an extension of the  $k$ -means alignment method,<sup>45</sup> originally designed for  $\mathbb{R}^p$ -valued functional data, to quaternion-valued functional data. The result is an average gait cycle, which we called individual gait pattern (IGP). We express each time point as a percentage of the

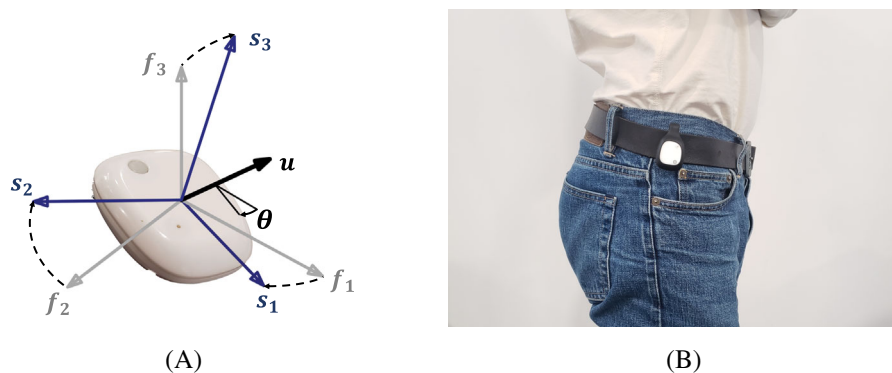


FIGURE 2 The motion sensor in its environment. (A) IMU and spatial orientation. (B) IMU clipped on the belt

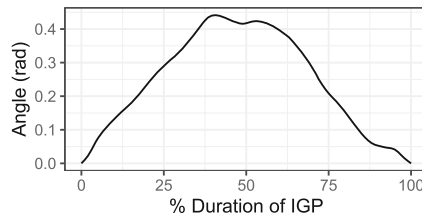


FIGURE 3 Example of the hip angle of an IGP

total IGP duration, ranging from 0% to 100% with steps of 1%. Finally, the IGP is straightened out so that the last rotation matches the first rotation. The final IGP for subject  $i$  is a 101-dimensional unit QTS  $Q_i := (\mathbf{q}_{i,0} \cdots \mathbf{q}_{i,100}) \in \mathbb{H}_u^{101}$ , in which (i) each element is a unit quaternion representing the hip orientation, that is, the rotation between the orientation of the sensor observed at a given percent of the IGP and its orientation observed at the first time of the IGP, and (ii) such that the first and last element always contain the identity rotation  $(1, 0, 0, 0)^T$ . We hypothesize that the shape of the IGP of an individual is defined by his/her walking abilities and thus is impacted by any gait-related impairment. Plotting raw QTS can give a poorly readable graphical representation, especially if there are several QTS to be drawn, as it consists in the evolution of the 4 dimensions  $w$ ,  $x$ ,  $y$  and  $z$  over time. We chose to plot a more intuitive, interpretable and easy-to-read representation of the IGP as the time series of the angles between the first and current orientation of the hip given by:

$$\theta_i = (2 \arccos \operatorname{Re}(\mathbf{q}_{i,0}), \dots, 2 \arccos \operatorname{Re}(\mathbf{q}_{i,100})). \quad (11)$$

Figure 3 displays an example of a series  $\theta_i$  computed from an IGP with Equation (11).

### 3.1.2 | Clinical data

The EDSS<sup>46</sup> is the most widely used scale to assess the extent of overall disability of patients diagnosed with MS.<sup>47</sup> The score on the EDSS is attributed by a neurologist on the basis of the amount of functional impairment in the central nervous system, which includes any type of walking deficiencies. The EDSS is described as an ordinal rating system ranging from 0 (normal neurological status) to 10 (death due to MS).<sup>47</sup> For our purposes, we can distinguish low EDSS scores (below 4) that are attributed to patients with mild gait impairment, moderate EDSS scores (from 4 to 6.5) that are attributed to patients with severe gait impairment and high EDSS scores (above 7) that are attributed to patients who cannot walk anymore without bilateral assistance (canes, crutches, or braces). The EDSS is criticized for its lack of linearity (ie, clinically unequal incremental scores).<sup>48</sup> In addition, gait impairment is only one information taken into account when assigning an EDSS score, which aims at providing a general assessment of the overall disability of the patient. Two patients with identical EDSS score might therefore still have different individual gait patterns. Nevertheless, the EDSS is currently the gold standard for assessing disease progress in MS<sup>48</sup> and, as such, needs to be taken into account when analyzing gait impairment.

## 3.2 | Methodology for computing and comparing clustering structures of walking data

We aim at comparing the compromise-based clustering strategy `hclustcompro` with its unsupervised version `HAC` and with a method that is representative of the ensemble clustering strategy. Ensemble clustering pertains to merging clustering results obtained from different methods applied independently on multiple data sources collected on the same observations. We choose the method `mergeTrees`<sup>49</sup> which builds up a consensus tree from a set of dendrograms that share the same  $n$  leaves (see Appendix A for a more detailed description). It is implemented in the `mergeTrees` R package. We will therefore summarize the commonalities and differences between these approaches in Section 3.2.1. Then, Section 3.2.2 will be dedicated to explaining how we selected the optimal number of clusters. Finally, we will define the criteria that we used to perform cluster validation in Section 3.2.3.

### 3.2.1 | Clustering methods

First, two dissimilarity matrices are computed from each of the two data sources presented in Section 3.1:

- The matrix  $D_1$  contains the pairwise QDTW dissimilarities between IGPs computed following Equation (7), using  $c(\cdot, \cdot)$  as the geodesic distance defined in Equation (4), and properly normalized by dividing all dissimilarities by the largest one.
- The matrix  $D_0$  contains the pairwise Gower dissimilarities between EDSS scores. Its advantages are three-fold: (i) it is by definition already normalized, (ii) it has been designed for ordinal data<sup>50</sup> and (iii) an implementation is provided through the function `daisy()` from the R package `cluster`.<sup>51</sup>

All three compared clustering methods subsequently use one or both dissimilarity matrices to propose a hierarchy of possible clustering structures: (i) HAC uses  $D_1$  only to provide a dendrogram  $\mathcal{T}_1$ , (ii) `mergeTrees` combines  $\mathcal{T}_1$  with another dendrogram  $\mathcal{T}_0$  obtained from  $D_0$  to output a consensus tree  $\mathcal{C}(\mathcal{T}_1, \mathcal{T}_0)$ , and (iii) `hclustcompro` uses both  $D_1$  and  $D_0$  to compute a weighted dissimilarity matrix  $D_\alpha$  according to Equation (8) from which a dendrogram  $\mathcal{T}_\alpha$  is obtained. We applied all three methods with complete linkage criteria.

### 3.2.2 | Selection of the number of clusters

HAC outputs a dendrogram which represents a hierarchy of possible grouping structures of the data points. A critical point is to elucidate the optimal number of clusters  $K$  from this dendrogram. There are many criteria in the literature that provide assistance for this task.<sup>52</sup> The elbow method is one of the most used in practice. It pertains to visualizing the within-cluster sum of squared (WSS) distances to a cluster prototype as a function of the number of clusters. A cluster prototype is a (possibly non-observed) individual that is the most central w.r.t. the individuals in the cluster. The WSS is a monotonically decreasing function of the number of clusters and it reaches 0 when the number of clusters matches the number of individuals in the sample. The WSS usually decays faster at the beginning and slows down at some point. A good candidate for the optimal number of clusters is to find that transition spot which resembles an elbow, hence the name of the method. We chose to represent the cluster prototype as the medoid,<sup>27</sup> which is the most central individual among the individuals that compose a given cluster. It therefore depends upon the chosen representation and dissimilarity. We chose to use the IGP representation and the QDTW dissimilarity. Mathematically, we can therefore define the WSS that we used as follows. Let  $Q_1, \dots, Q_n$  be a sample of  $n$  QTS grouped into  $K$  clusters  $C_1, \dots, C_K$ . In addition, let  $V_k := \{\ell \in [1, n] : Q_\ell \in C_k\}$ . The WSS is given by:

$$\text{WSS}(K) = \sum_{k=1}^K \sum_{i \in V_k} \text{QDTW}^2(Q_i, \tilde{Q}_k), \quad (12)$$

where  $\tilde{Q}_k$  is the medoid of the cluster  $C_k$  computed as:

$$\tilde{Q}_k = Q_{i_k}, \quad \text{where } i_k = \arg \min_{i \in V_k} \sum_{j \in V_k, j \neq i} \text{QDTW}^2(Q_i, Q_j).$$

For each compared method (HAC, `mergeTrees`, and `hclustcompro`), we evaluated the partitions generated by  $K$  ranging from 2 to 10. We made an informed decision about the optimal number of clusters for each method based on the elbow method and the expertise of a neurologist who gave an assessment of clinical relevance for various preselected partitions. We also discarded partitions that generated singleton clusters.

### 3.2.3 | Cluster validation

We used both internal and external cluster validation criteria.

1. *Internal criteria.* We can define the within-cluster inertia  $I_W^{(k)}$  of each cluster  $C_k$  as:

$$I_W^{(k)} := \frac{1}{|V_k|} \sum_{i \in V_k} \text{QDTW}^2(Q_i, \tilde{Q}_k).$$

This value should be small in comparison to the between-cluster inertia  $I_B^{(k)}$  which is classically defined as:

$$I_B^{(k)} = \text{QDTW}^2(\tilde{Q}_k, \tilde{Q}),$$

where  $\tilde{Q}$  is the overall medoid of the data set computed as:

$$\tilde{Q} = Q_i, \quad \text{where } i = \arg \min_{i \in \{1, \dots, n\}} \sum_{j \in \{1, \dots, n\}, j \neq i} \text{QDTW}^2(Q_i, Q_j).$$

We can express this condition through the proportion of within-cluster inertia defined as:

$$p_W^{(k)} := \frac{I_W^{(k)}}{I_W^{(k)} + I_B^{(k)}}. \quad (13)$$

This proportion can be computed for each cluster and quantifies how far the observations are from the cluster medoid w.r.t. how far the cluster is from the overall medoid of the data. This criterion should be as small as possible which is achieved when the observations are highly concentrated in their cluster and the cluster is well separated from the others.

We can compute the same proportion at the level of the entire partition by defining:

$$I_W := \frac{\sum_{k=1}^K |V_k| I_W^{(k)}}{\sum_{k=1}^K |V_k|}, \quad I_B := \frac{\sum_{k=1}^K |V_k| I_B^{(k)}}{\sum_{k=1}^K |V_k|}, \quad \text{and} \quad p_W = \frac{I_W}{I_W + I_B}. \quad (14)$$

We also use the Dunn index to assess how compact and well-separated the clusters are. It is defined as:

$$DI := \frac{\min_{k, \ell \in \{1, \dots, K\}^2, k \neq \ell} \delta(C_k, C_\ell)}{\max_{k \in \{1, \dots, K\}} \Delta(C_k)}, \quad (15)$$

with

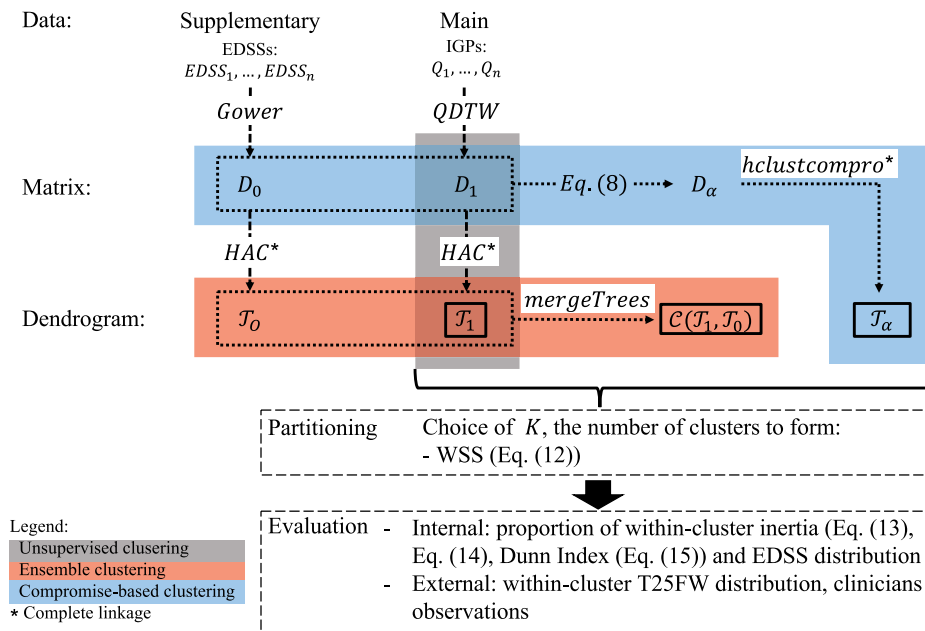
$$\delta(C_k, C_\ell) = \min_{i \in V_k, j \in V_\ell} \text{QDTW}(Q_i, Q_j) \quad (\text{separation}) \quad \text{and} \quad \Delta(C_k) = \max_{i, j \in V_k} \text{QDTW}(Q_i, Q_j) \quad (\text{cohesion}).$$

This index is expected to be large when clusters are compact and well-separated.<sup>53</sup>

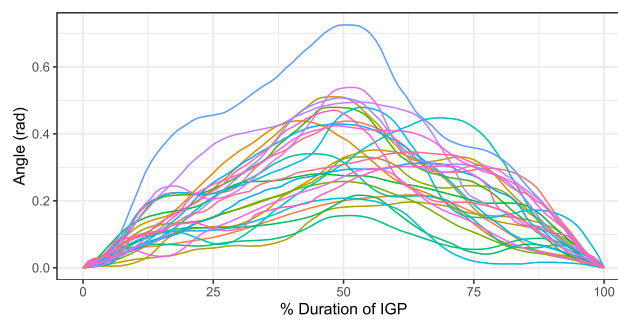
The above-mentioned criteria provide internal assessment of a clustering structure in terms of the IGP data. We will also include an internal assessment in terms of the EDSS data. This will be achieved by looking at the within-cluster distribution of EDSS scores which represents the extent to which a cluster groups patients with similar overall clinically-observable disability.

2. *External criteria.* We used the following two external criteria:

- The within-cluster distribution of the time achieved at the T25FW test (time to walk a distance of 25 feet). This obviously depends on the patient walking velocity and thus on his/her gait impairment.
- An external clinical assessment and ranking of the three proposed clustering structures performed by five neurologists who are all experts on MS. This was achieved by first providing them with a brief presentation of the context of the study and the semi-supervised methods. Then, we asked them to blindly evaluate the clustering structures obtained from the three candidate methods using both their expert knowledge and additional clinical information about the patients.



**FIGURE 4** Overall methodological pipeline. It summarizes both the computation of the clustering structures according to each candidate clustering method and the way they were evaluated and compared



**FIGURE 5** Hip angle of the IGP of the 27 patients

**TABLE 2** Distribution of EDSS scores in the sample

EDSS	0	1	1.5	2	2.5	3	4	5.5	6	Total
Size	7	3	1	4	5	2	3	1	1	27

The complete methodology for computing and comparing clustering structures from the three candidate clustering methods is summarized in Figure 4.

## 4 | RESULTS

The data were collected at the University Hospital of Nantes (France) on 27 patients who met the inclusion criteria of the OFSEP-HD cohort from the *Observatoire Français de la Sclérose En Plaques* (OFSEP) and are thus regularly seen at the hospital. Inclusions began in September 2019 and ended in May 2020. Data were collected during a routine examination by either a neurologist or a staff member of the neurology department. Figure 5 provides a visual representation of the hip rotation during the IGP of the 27 patients. Table 2 shows the distribution of EDSS scores among the 27 patients.

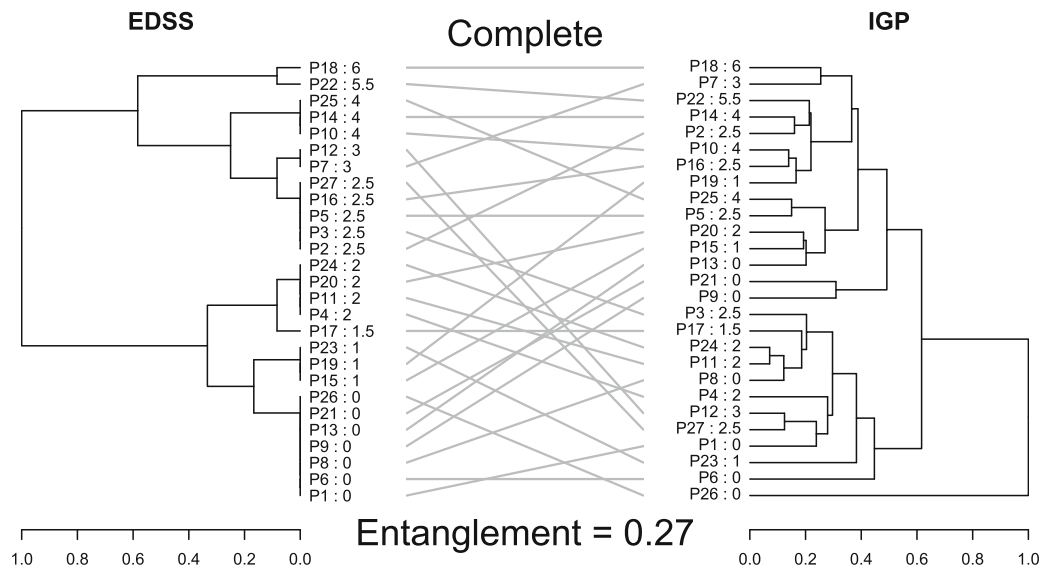


FIGURE 6 Tanglegrams between EDSS data (left) and IGP data (right) with complete linkage

In the following, we present the results of the comparison of the three clustering methods described in Section 3.2.1 using the approach detailed in Section 3.2.2 for selecting the optimal number of clusters and in Section 3.2.3 for validating the clustering structures. Specifically, Section 4.1 discusses the level of entanglement between the dendrograms obtained from both IGP and EDSS data. Section 4.2 gives the results of the estimation of the weighting parameter  $\alpha$  for the `hclustcompro` method. Section 4.3 describes the choice of the optimal number of clusters and Section 4.4 finally details the results of the comparison between the three clustering methods.

#### 4.1 | Entanglement

Figure 6 exhibits the tanglegrams between the dendrograms  $\mathcal{T}_0$  and  $\mathcal{T}_1$  produced by applying the HAC method on  $D_0$  (EDSS data) and  $D_1$  (IGP data) respectively, with the complete linkage criteria. The entanglement is also reported. The leaves of the dendrograms are labeled according to the patient ID and his/her EDSS score.

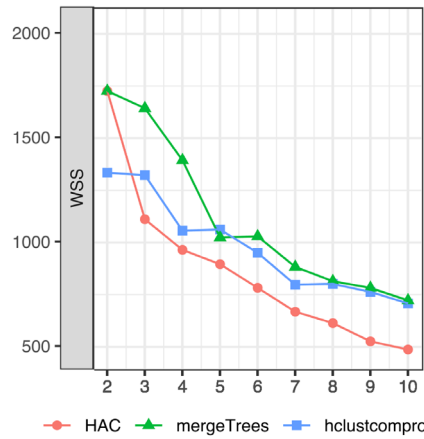
The entanglement is a score between 0 and 1 which measures how different the relative positions of the patients are in the two dendrograms. It is a measure of dissimilarity between dendrograms. Hence, high values indicate large differences while low values suggest that the dendrograms provide similar clustering structures. The observed value of 0.27 suggest that there is an overlap of information between EDSS and IGP data which is not surprising given that the EDSS includes an evaluation of walking disability. Nevertheless, the entanglement is not close to 0 either, which suggests that both sources of data contain unique novel information pertaining to the patient disability.

#### 4.2 | Estimation of the `hclustcompro` weighting parameter

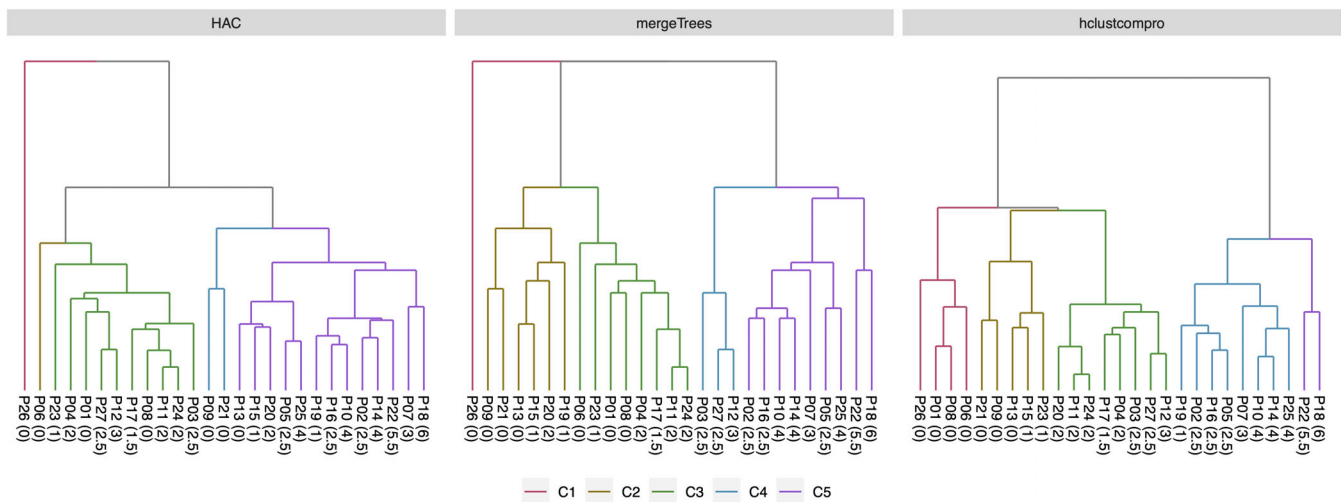
The dissimilarity matrix  $D_\alpha$  which is used by the `hclustcompro` procedure to produce the final dendrogram  $\mathcal{T}_\alpha$  depends on the unknown weighting parameter  $\alpha$  according to Equation (8). We used Equation (10) to get a pointwise estimate of this parameter. These computations led to the estimated value  $\hat{\alpha} = 0.69$ . In essence, this means that we evaluate the dissimilarity between patients as a linear combination of the dissimilarity between their IGP (accounted for with a weight of 69%) and the dissimilarity between their EDSS scores (accounted for with a remaining weight of 31%).

#### 4.3 | Choice of the optimal number of clusters

Figure 7 displays the variation of WSS computed from Equation (12), as the number of clusters grows from 2 to 10.



**FIGURE 7** Within-cluster sum of squares (WSS) as a function of the number of clusters. Colors encode the clustering methods



**FIGURE 8** Dendrograms obtained by combining each of the three clustering methods (HAC, mergeTrees, and hclustcompro) with complete linkage criteria. Clusters are matched between dendrograms by ordering them as much as possible from smallest to largest median EDSS

It is computed from partitions generated by cutting the dendrograms in Figure 8 at various heights in order to achieve the desired number of clusters.

In this section, we analyze both figures to determine a unique global optimal number of clusters.

1. *HAC on IGP data.* Figure 7 suggests that the optimal number of clusters should be three, regardless of the linkage criterion. In effect, an elbow is visible on the WSS curves. The left dendrogram displayed in Figure 8 seems to recommend splitting patients into only three groups of unbalanced sizes. Within groups, the distributions of EDSS scores are mostly random, which means that the grouping structure is not representative of the overall disability.
2. *mergeTrees.* Figure 7 indicates that five clusters should be formed. Figure 8 shows that the distribution of EDSS scores within groups is less heterogeneous than the ones obtained that of from HAC. However, some clusters group patients with large differences in overall disability and, conversely, some patients with similar overall disability end up into different groups.
3. *hclustcompro.* Figure 7 suggests that the optimal number of clusters should be in the range  $K \in \{4, 5, 6\}$ . In Figure 8, we can appreciate how naturally the dendrogram naturally suggests to form five clusters of relatively similar sizes and homogeneous distributions of EDSS scores.

Considering all these observations, and because it allows a one-to-one comparison of the resulting clusters between all clustering methods, we systematically generated five clusters for all methods. The resulting partitions can be visually appreciated in Figure 8 by the color coding scheme.

#### 4.4 | Comparison of the clustering methods

Table 3 provides a description of the clusters from the perspective of internal data, that is, data that have been used to generate the clusters. Specifically, each cluster is summarized by

- Its size ( $n$ ): This is the number of patients in the cluster ( $n$ ).
- Its proportion of within-cluster inertia ( $p_W^{(k)}$ ): This is computed according to Equation (13) and amounts to the ratio of within-cluster inertia to the distance between the cluster medoid and the overall medoid of the data; small values indicate a well-isolated cluster with concentrated members.
- Its median EDSS score: This informs about the average level of overall disability of a given cluster.

In Table 3, for each clustering method, we ordered as much as possible the clusters from smallest to largest median EDSS scores. It was not always possible to achieve a monotonic ranking because sometimes clusters with more similar median EDSS scores ended up too far from each other in the final dendrogram.

Looking at the cluster sizes, we can notice that both HAC and mergeTrees produce singleton clusters, that is, clusters with a single patient. We can furthermore observe that this is not an artifact due to the choice of the number of clusters. In effect, regardless of the value of  $K$ , Figure 8 clearly shows that both methods do produce a singleton cluster at the first bifurcation. Singleton clusters falls more within the realm of outlier detection than the one of clustering. In effect, they relieve another cluster from an observation that was too far from the others in the group to create another group only for that outlier observation. Hence both clusters benefit from a reduction of inertia as one can appreciate from the columns  $p_W^{(k)}$  in Table 3. However, from the perspective of clustering individuals, that is, of finding homogeneous groups of individuals, it makes no sense to create a cluster for one individual. This is in line with the elbow method which is usually used to gain insight about the optimal number of clusters from the WSS curve while the minimum WSS is trivially achieved for  $K = n$ , that is, when all individuals are within their own cluster, which clearly yields a useless partition. We can finally see from column EDSS that HAC fails to produce a sequence of clusters with increasing median EDSS. By contrast, mergeTrees and hclustcompro produce a sequence of clusters with increasing median EDSS. Furthermore, we can appreciate how the later approach would never generate singleton clusters, no matter which number of clusters we split the individuals into (see Figure 8). This translates into nonzero proportions of within-cluster inertia for all clusters in Table 3.

We shall now provide a performance assessment of the clustering methods at the granularity of the partition rather than the individual clusters. Table 4 naturally propose aggregated proportions of within-cluster IGP inertia for the entire partition ( $p_W$ ) and adds the Dunn index as well which is high when clusters are well-separated from each other but composed of very close individuals in terms of IGP data. Given that clusters are matched by median EDSS scores between methods and that we have argued that singleton clusters are undesirable in the scope of establishing a meaningful partition and wrongly improve internal cluster validation metrics, we also decided to present in Table 4 the same metrics but computed only the last three clusters C3, C4, and C5 which were never singleton clusters for any method.

TABLE 3 Summary of the clusters in terms of size ( $n$ ), proportion of within-cluster IGP inertia ( $p_W^{(k)}$ ) and median EDSS

Clustering	C1			C2			C3			C4			C5		
	$n$	$p_W^{(1)}$	EDSS	$n$	$p_W^{(2)}$	EDSS	$n$	$p_W^{(3)}$	EDSS	$n$	$p_W^{(4)}$	EDSS	$n$	$p_W^{(5)}$	EDSS
HAC	1	0.0	0	1	0.0	0	10	50.5	2	2	28.6	0	13	43.8	2.5
mergeTrees	1	0.0	0	6	63.6	1	8	55.2	1.5	3	22.3	2.5	9	38.0	4
hclustcompro	4	64.7	0	5	63.6	0	8	52.0	2	8	33.9	3	2	43.5	5.5

Note: For a given cluster, its proportion of within-cluster inertia is computed on the IGP data according to Equation (13).

TABLE 4 Performance metrics of IGP data at the partition level

Clustering	Using all clusters		Using only clusters that are never singletons	
	$P_w$	DI	$P_w$	DI
HAC	35.0	0.428	43.8	0.428
mergeTrees	40.2	0.288	41.4	0.329
hclustcompro	51.6	0.214	42.1	0.467

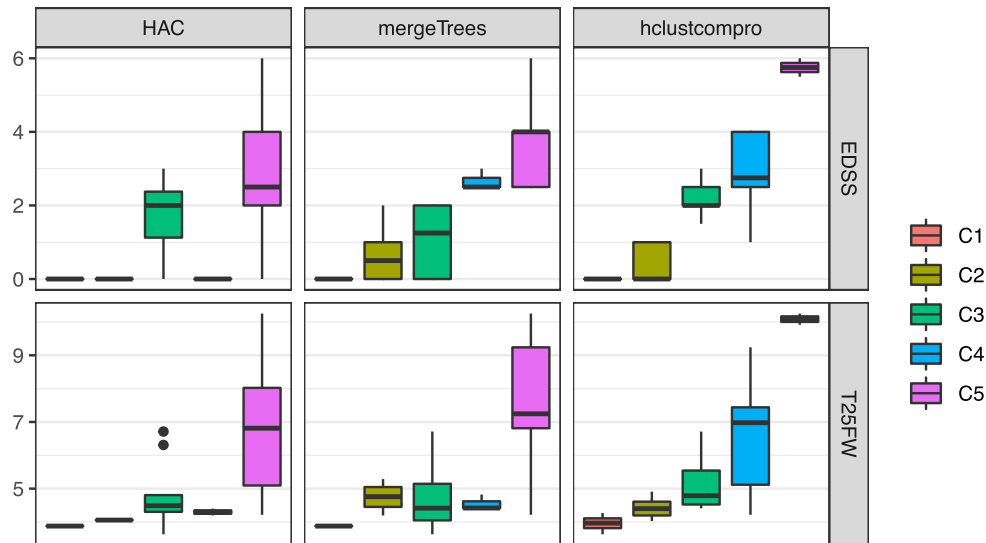


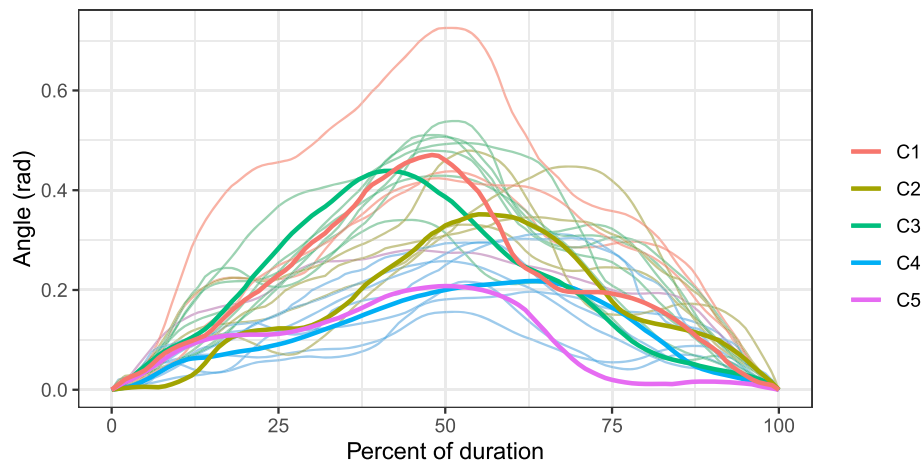
FIGURE 9 Cluster description in terms of standard clinical indicators of overall (EDSS) and gait-specific (T25FW) disability

As expected, when looking at the performance metrics computed using all five clusters, the HAC method outperforms the semi-supervised methods because it only accounts for IGP data and thus does a better job at finding groups with similar IGP characteristics. We see that `mergeTrees` comes as second-best which is largely due to the presence of a singleton cluster. This interpretation is confirmed when looking at the performance metrics computed using only clusters that are never singletons. From this perspective, `hclustcompro` is as good as HAC in terms of inertia and clearly outperforms both HAC and `mergeTrees` when comparing Dunn indices.

Figure 9 displays the intra-cluster distributions of EDSS scores and the duration of the T25FW test for the three clustering methods (HAC, `hclustcompro`, `mergeTrees`). It provides a description of the clusters from the perspective of standard clinical indicators of global disability (EDSS) and gait impairment (T25FW), which complements the description provided in Table 3.

The intra-cluster distributions of EDSS scores and of walking time during the T25FW test presented in Figure 9 are used to assess the clinical relevance of the generated partitions. We also asked five neurologists to interpret the results and their comments led to the following observations:

1. HAC. Clusters formed by HAC gather patients with very heterogeneous MS severity. For example, cluster C5 contains patients with normal neurological functions (EDSS = 0) and patients that cannot walk 100 m without walking aid (EDSS = 6). The patients in this cluster also present very different walking time during the T25FW test as illustrated in Figure 9.
2. `mergeTrees`. The partition provided by `mergeTrees` gathers patients with more similar MS severity. It tends to present higher EDSS scores in the last two clusters C4 and C5 than in the first three clusters C1, C2, and C3. However, some patients with similar EDSS scores are split into different groups whereas some patients with different MS severity are grouped together. In addition, the resulting clusters do not present either significant or ordered differences in terms of walking speed.



**FIGURE 10** Variation of the angle of hip rotation during a gait cycle. Clusters are those produced by `hclustcompro` with complete linkage

3. `hclustcompro`. The partition provided by `hclustcompro` is more homogeneous. Cluster C1 regroups patients with EDSS score of 0 (ie, identical to healthy individuals from a neurological standpoint). Figure 9 also shows that they correspond to the patients with the lowest walking time during the T25FW test. Cluster C2 regroups patients with EDSS scores 0 and 1. They tend to have slightly higher walking times during the T25FW test w.r.t. the patients in cluster C1. When contrasted with cluster C1, this suggests that the IPG could be able to separate patients with apparent normal neurological condition on the basis of their gait impairment. Cluster C5 is also of high interest as it regroups the two patients with worst neurological assessment. They are also the two patients performing the worst at the T25FW test. Clusters C3 and C4 are less homogeneous in terms of clinical features, but they regroup patients with an intermediate degree of severity of the pathology and C4 regroups patients with slightly higher EDSS scores than patients in C3, although not significantly.

According to these observations, the method `hclustcompro` leads to the best clustering of the cohort. As a final illustration of these results, the variation of the rotation angle of the hip during a gait cycle is extracted from the IGP according to Equation (11) and displayed in Figure 10. Bold curves represent the medoids of each cluster. Figure 10 reveals that patients in the clusters C1, C2, and C3 tend to present a higher amplitude of hip rotation than patients in clusters C4 and C5. This is interesting because we have seen before that C1 to C3 also regroup patients with milder overall and gait-specific disabilities.

## 5 | DISCUSSION

In this article, we proposed a semi-supervised clustering method for QTS coined `hclustcompro` which combines (i) the principles of hierarchical clustering to provide a hierarchy of possible clustering configurations, (ii) the idea of compromise-based clustering<sup>21,22</sup> for guiding the clustering using supplementary information, and (iii) the DTW dissimilarity adapted for unit QTS.<sup>12</sup> The `hclustcompro` method linearly combines the dissimilarity matrices from the quaternion data and from the supplementary data using a weighting parameter. This parameter is estimated by minimizing an objective criterion based on the cophenetic distance and is naturally interpreted as the proportion of each source of information injected into the clustering process.

We also compared `hclustcompro` with unsupervised HAC and with `mergeTrees`<sup>54</sup> (an ensemble clustering approach) in the context of gait analysis. All three methods were applied to cluster a data set of 27 patients diagnosed with MS. In this study, QTS measuring the hip rotation over time were used to compute an individual gait pattern as the main source of information for clustering and supplementary information came in the form of a clinical score representing the overall disability of a patient. The results demonstrate both the importance and usefulness of injecting even a small amount of prior information (when available) as the two semi-supervised methods provide more clinically interpretable clusters in contrast to the more traditional mono-source HAC performed only on the IGP data. The results also tend to demonstrate that `hclustcompro` outperforms `mergeTrees`, especially when in terms of clinical interpretability of the

resulting partition. This is in line with the observations made by the authors of `mergeTrees` themselves who described a loss of efficiency of `mergeTrees` when using too heterogeneous sources of information.<sup>54</sup>

To the best of our knowledge, this is the first description of semi-supervised clustering of QTS, and its first application in the context of gait analysis. It provides interesting results and several research perspectives can be identified.

As a word of caution, we shall remind that the MS data set used in this study is a rather small sample. Hence, our results and conclusions should be confirmed on a larger cohort. Adding new patients to the analysis may lead to the formation of new groups and to the identification of more precise gait characteristics shared by patients of the same cluster. Also, an investigation of how well the IPG correlates with state-of-the-art spatio-temporal parameters characterizing gait provided by other devices, such as Gaitrite,<sup>55</sup> may improve the interpretation of the results and lead to the identification of other gait characteristics shared by patients of the same cluster.

Compromise-based HAC leans toward factorial methods for the joint analysis of multiple data tables such as STATIS which searches for a compromise table that is the most representative according to some criterion.<sup>56,57</sup> It could therefore be a natural methodological perspective to extend `hclustcompro` to the case of more than two sources of information. This provides an interesting development perspective in gait analysis. It might indeed be possible to take into account parameters in addition to the shape of the IGP, for example, the intra-individual walking variability which has been described to be associated with the risk of falling among the elderly.<sup>58</sup>

As hierarchical clustering is a distance-based method, results are by construction influenced by the choice of the dissimilarity measure. In this article, we chose QDTW which is to date the only dissimilarity suitable for QTS. There are several other dissimilarities described for time series of Euclidean observations, which are elastic, based on models or on specific representations of the data. The generalization of these methods to QTS would make it possible to extend the range of possible choices which would ultimately allow to explore different geometries for quaternion data.

## ACKNOWLEDGEMENTS

The authors gratefully thank the ARSEP foundation (Fondation pour l'Aide à la Recherche sur la Sclérose En Plaques) for funding this study as well as the Observatoire Français de la Sclérose en Plaques Group (OFSEP), the clinical research staff of the University Hospital of Nantes, Dr Emanuelle Lepage and Dr Anne Kerbrat from the University Hospital Center of Rennes, France, Dr Amélie Dos Santos and Dr Sandrine Wiertelowski from the University Hospital Center of Nantes, and the patients who all immensely helped in collecting the data for this project.

## DATA AVAILABILITY STATEMENT

The authors elect to not share data.

## ORCID

Pierre Drouin  <https://orcid.org/0000-0002-6735-5768>

Aymeric Stamm  <https://orcid.org/0000-0002-8725-3654>

## REFERENCES

1. Muro-De-La-Herran A, Garcia-Zapirain B, Mendez-Zorrilla A. Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors*. 2014;14:3362-3394.
2. Switonski A, Michalczyk A, Josinski H, Polanski A. Dynamic time warping in gait classification of motion capture data. *Int J Comput Inf Eng*. 2012;6:1289-1294.
3. Switonski A, Josinski H, Wojciechowski K. Dynamic time warping in classification and selection of motion capture data. *Multidimens Syst Signal Processg*. 2019;30:1437-1468.
4. Piórek M, Josiński H, Michalczyk A, Świtoński A, Szczesna A. Quaternions and joint angles in an analysis of local stability of gait for different variants of walking speed and treadmill slope. *Inf Sci*. 2017;384:263-280.
5. Szczesna A. Quaternion entropy for analysis of gait data. *Entropy*. 2019;21:79.
6. Josiński H, Świtoński A, Michalczyk A, Grabiec P, Pawlyta M, Wojciechowski K. Assessment of local dynamic stability in gait based on univariate and multivariate time series. *Comput Math Methods Med*. 2019;2019:6917658.
7. Motl RW, Cohen JA, Benedict R, et al. Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis. *Mult Scler J*. 2017;23:704-710.
8. Aggarwal CC, Reddy CK, eds. *Data Clustering: Algorithms and Applications*. Boca Raton: CRC Press; 2014.
9. Baghdadi A, Cavuoto LA, Jones-Farmer A, Rigdon SE, Esfahani ET, Megahed FM. Monitoring worker fatigue using wearable devices: a case study to detect changes in gait parameters. *J Qual Technol*. 2019;53:47-71.
10. Pulido-Valdeolivas I, Gómez-Andrés D, Martín-Gonzalo JA, et al. Gait phenotypes in paediatric hereditary spastic paraplegia revealed by dynamic time warping analysis and random forests. *PLoS One*. 2018;13:1-28.

11. Steinmetzer T, Bonninger I, Priwitzer B, et al. Clustering of human gait with Parkinson's disease by using dynamic time warping. *Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*; 2018:1-6; IEEE.
12. Jablonski B. Quaternion dynamic time warping. *IEEE Trans Signal Process.* 2011;60:1174-1183.
13. Dinler D, Tural MK. A survey of constrained clustering. In: Celebi M, Aydin K, (eds). *Unsupervised learning algorithms*. Cham: Springer International Publishing; 2016:207-235.
14. Lampert T, Dao TBH, Lafabregue B, et al. Constrained distance based clustering for time-series: a comparative and experimental study. *Data Mining Knowl Discov.* 2018;32:1663-1707.
15. Bade K, Nürnberger A. Creating a cluster hierarchy under constraints of a partially known hierarchy. *Proceedings of the 2008 SIAM International Conference on Data Mining*; 2008:13-24; SIAM.
16. Davidson I, Ravi SS. Agglomerative hierarchical clustering with constraints: theoretical and empirical results. *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*; 2005:59-70; Springer.
17. Zhao H, Qi Z. Hierarchical agglomerative clustering with ordering constraints. *Proceedings of the 2010 3rd International Conference on Knowledge Discovery and Data Mining*; 2010:195-199; IEEE.
18. Zheng L, Li T. Semi-supervised hierarchical clustering. *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*; 2011:982-991; IEEE.
19. Cornuéjols A, Wemmer C, Gañarski P, Bennani Y. Collaborative clustering: why, when, what and how. *Inf Fusion.* 2018;39:81-95.
20. He G, Pan Y, Xia X, He J, Peng R, Xiong NN. A fast semi-supervised clustering framework for large-scale time series data. *IEEE Trans Syst Man Cybern Syst.* 2021;51:4201-4216.
21. Ma X, Dhavala S. Hierarchical clustering with prior knowledge; 2018. <https://arxiv.org/abs/1806.03432>.
22. Bellanger L, Coulon A, Husi P. PerioClust: a simple hierarchical agglomerative clustering approach including constraints. In: Chadji-padelis T, Lausen B, Markos A, Lee TR, Montanari A, Nugent R, eds. *Data Analysis and Rationality in a Complex World*. Cham: Springer International Publishing; 2021:1-8.
23. Coulon A, Bellanger L, Husi P. SPARTAAS: statistical pattern recognition and dating using archaeological artefacts assemblages. *Comprehens R Arch Netw (CRAN)*. 2021.
24. Lance GN, Williams WT. A general theory of classificatory sorting strategies. *Comput J.* 1967;9:373-380.
25. Ah-Pine J, Wang X. Similarity based hierarchical clustering with an application to text collections. In: Boström H, Knobbe A, Soares C, Papapetrou P, eds. *Advances in Intelligent Data Analysis XV*. Cham: Springer International Publishing; 2016:320-331.
26. Piórek M. Analysis of Chaos for quaternion time series. *Analysis of Chaotic Behavior in Non-linear Dynamical Systems*. New York: Springer; 2019:73-88.
27. Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering - a decade review. *Inf Syst.* 2015;53:16-38.
28. Forner-Cordero A, Koopman HJFM, van der Helm FCT. Describing gait as a sequence of states. *J Biomech.* 2006;39:948-957.
29. Blazkiewicz M, Lann Vel Lace K, Hadamus A. Gait symmetry analysis based on dynamic time warping. *Symmetry.* 2021;13:836.
30. Adhikary S, Ghosh A. Dynamic time warping approach for optimized locomotor impairment detection using biomedical signal processing. *Biomed Signal Process Control.* 2022;72:103321.
31. Varatharajan R, Manogaran G, Priyan MK, Sundarasekar R. Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. *Clust Comput.* 2018;21:681-690.
32. Geler Z, Kurbalija V, Ivanović M, Radovanović M, Dai W. Dynamic time warping: Itakura vs Sakoe-Chiba. *Proceedings of the 2019 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*; 2019:1-6; IEEE.
33. Müller M, Röder T, Clausen M, Eberhardt B, Krüger B, Weber A. Documentation Mocap database HDM05. *Inst Inform II Univ Bonn.* 2007;CG-2007-2, 1610-8892.
34. Sokal RR, Rohlf FJ. The comparison of dendrograms by objective methods. *Taxon.* 1962;11:33-40.
35. LaRocca NG. Impact of walking impairment in multiple sclerosis. *Patient Patient-Center Outcomes Res.* 2011;4:189-201.
36. Fischer JS, Jak AJ, Kniker JE, Rudick RA, Cutter G. *Multiple Sclerosis Functional Composite (MSFC): Administration and Scoring Manual*. New York: National Multiple Sclerosis Society; 2001.
37. Severini G, Manca M, Ferraresi G, et al. Evaluation of clinical gait analysis parameters in patients affected by multiple sclerosis: analysis of kinematics. *Clin Biomech.* 2017;45:1-8.
38. Pau M, Mandaresu S, Pilloni G, et al. Smoothness of gait detects early alterations of walking in persons with multiple sclerosis without disability. *Gait Posture.* 2017;58:307-309.
39. Rueterbories J, Spaich EG, Larsen B, Andersen OK. Methods for gait event detection and analysis in ambulatory systems. *Med Eng Phys.* 2010;32:545-552.
40. Helwig NE, Hong S, Hsiao-Weckler ET, Polk JD. Methods to temporally align gait cycle data. *J Biomech.* 2011;44:561-566.
41. Ploeger HE, Bus SA, Nollet F, Brehm MA. Gait patterns in association with underlying impairments in polio survivors with calf muscle weakness. *Gait Posture.* 2017;58:146-153.
42. Engelhard MM, Dandu SR, Patek SD, Lach JC, Goldman MD. Quantifying six-minute walk induced gait deterioration with inertial sensors in multiple sclerosis subjects. *Gait Posture.* 2016;49:340-345.
43. Kim CM, Eng JJ. Magnitude and pattern of 3D kinematic and kinetic gait profiles in persons with stroke: relationship to walking speed. *Gait Posture.* 2004;20:140-146.
44. Zhang B, Yan K, Jiang S, Wei D. Walking stability analysis by age based on dynamic time warping. *Proceedings of the 2008 8th IEEE International Conference on Computer and Information Technology*; 2008:544-548; IEEE.
45. Sangalli LM, Secchi P, Vantini S, Vitelli V. K-mean alignment for curve clustering. *Comput Stat Data Anal.* 2010;54:1219-1233.

46. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*. 1983;33:1444.
47. Meyer-Moock S, Feng YS, Maeurer M, Dippel FW, Kohlmann T. Systematic literature review and validity evaluation of the expanded disability status scale (EDSS) and the multiple sclerosis functional composite (MSFC) in patients with multiple sclerosis. *BMC Neurol*. 2014;14:58.
48. Whitaker JN, McFarland HF, Rudge P, Reingold SC. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. *Mult Scler J*. 1995;1:37-47.
49. Hulot A, Chiquet J, Rigaiil G. mergeTrees: aggregating trees. Comprehensive R archive network (CRAN), 2019.
50. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics*. 1971;27:857-871.
51. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol 344. Hoboken, New Jersey, U.S.: John Wiley & Sons; 1990:1-67.
52. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. *J Stat Softw*. 2014;61:1-36.
53. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. *J Intell Inf Syst*. 2001;17:107-145.
54. Hulot A, Chiquet J, Jaffrézic F, Rigaiil G. Fast tree aggregation for consensus hierarchical clustering. *BMC Bioinform*. 2020;21:120.
55. Sosnoff JJ, Weikert M, Dlugonski D, Smith DC, Motl RW. Quantifying gait impairment in multiple sclerosis using GAITRite™ technology. *Gait Posture*. 2011;34:145-147.
56. Escoufier Y. Operators related to a data matrix. In: JR Barra (ed). *Recents Developpements In Statistics: North-Holland Publishing Company*. 1977;125-131
57. Lavit C, Escoufier Y, Sabatier R, Traissac P. The act (statis method). *Comput Stat Data Anal*. 1994;18:97-119.
58. Reelick MF, Kessels RPC, Faes MC, Weerdesteijn V, Esselink RAJ, Rikkert MGMO. Increased intra-individual variability in stride length and reaction time in recurrent older fallers. *Aging Clin Exp Res*. 2011;23:393-399.
59. Hartigan JA. Consistency of single linkage for high-density clusters. *J Am Stat Assoc*. 1981;76:388-394.

**How to cite this article:** Drouin P, Stamm A, Chevreuril L, et al. Semi-supervised clustering of quaternion time series: Application to gait analysis in multiple sclerosis using motion sensor data. *Statistics in Medicine*. 2023;42(4):433-456. doi: 10.1002/sim.9625

## APPENDIX A. MERGETREES: A CONSENSUS-BASED CLUSTERING METHOD

This method have been described in the context of OMICS data by Hulot et al.<sup>54</sup> Let consider  $m$  data sets observed on the same  $n$  individuals. Let  $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_m\}$  the set of  $m$  dendrograms obtained from these data sets with any hierarchical agglomerative method. The method mergeTrees builds a consensus tree  $C(\mathcal{T})$  according to the following rule: *For any individuals  $i$  and  $j$  in  $\{1, \dots, n\}$ ,  $i \neq j$ , if  $i$  and  $j$  are not in the same cluster in at least one of the trees of  $\mathcal{T}$  at a given height  $h$ , then they are not in the same cluster in  $C(\mathcal{T})$  at height  $h$ .* The algorithm is also designed in such a way that  $C(\mathcal{T})$  follows the rules of anonymity, neutrality, and unanimity. Anonymity implies that the result does not depend on the order of the trees in  $\mathcal{T}$ . Neutrality implies that changing the labels of the leaves in the trees in  $\mathcal{T}$  simply relabels the leaves in  $C(\mathcal{T})$ . Finally, unanimity implies that merging a set composed of the same tree  $\mathcal{T}$  results in  $C(\mathcal{T}) = \mathcal{T}$ . It have to be mentioned that the heights in the dendrograms need to be comparable. Indeed, if all the divisions in a given tree  $\mathcal{T}_\alpha$  happen at a higher height than the divisions of any of the other trees, the consensus dendrogram will result in  $C(\mathcal{T}) = \mathcal{T}_\alpha$ . Rescaling step such as data normalization to make the dendrograms comparable could be therefore needed. One final remark can be made about the fact that the consensus tree will not be binary in the case divisions happens at the same height in multiple trees in the set. As example three branches may be linked at the same node in  $C(\mathcal{T})$ .

## APPENDIX B. CLUSTERING GAIT DATA IN MULTIPLE SCLEROSIS: ADDITIONAL RESULTS

This section presents the results obtained with the three methods HAC, mergeTrees, and hclustompro, following the methodological pipeline presented in Figure 4, using single and average linkage criteria.

### B.1 Entanglement

Figure B1A,B depicts the entanglement between the dendrograms  $\mathcal{T}_0$  and  $\mathcal{T}_1$  produced by applying the HAC method on  $D_0$  (EDSS data) and  $D_1$  (IGP data) respectively. The clustering structure provides by the two dendrograms are more similar with single linkage compared to average and complete linkage (cf. Figure 6). However, the entanglement for these three criteria is not close to 0, suggesting that the two data sources provide different information about the patient disability.

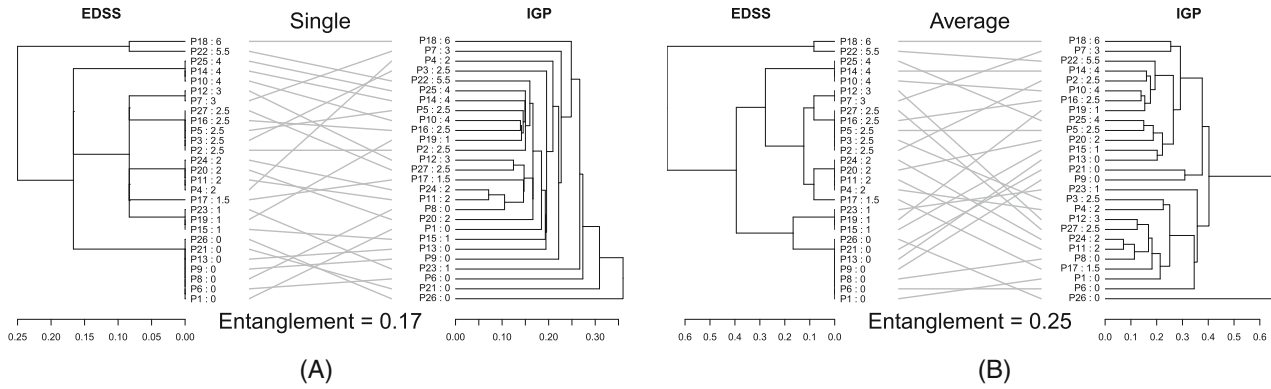


FIGURE B1 Tanglegrams between EDSS data (left) and IGP data (right) by linkage criterion

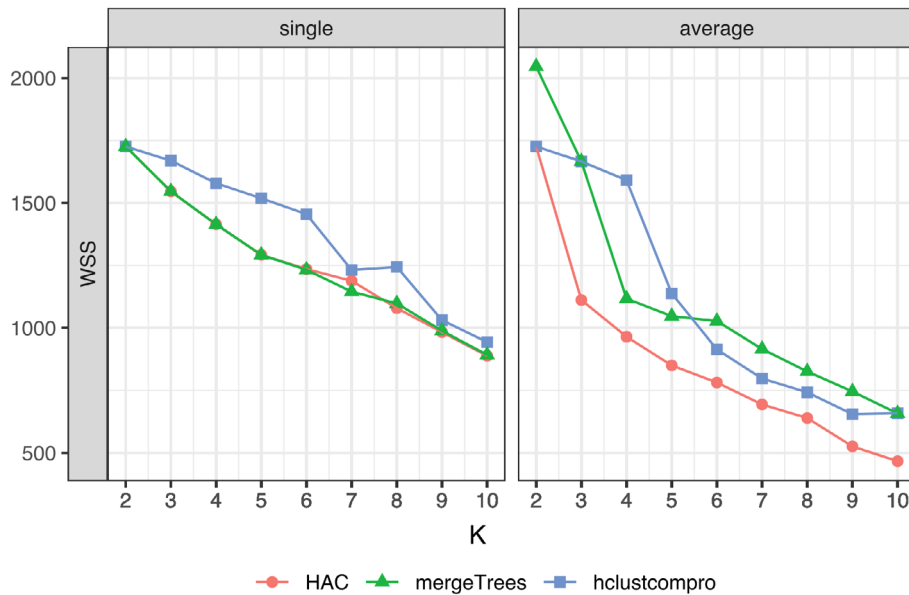


FIGURE B2 Within-cluster sum of squares (WSS) as a function of the number of clusters. Colors encode the clustering methods while columns correspond to different graph linkage criteria

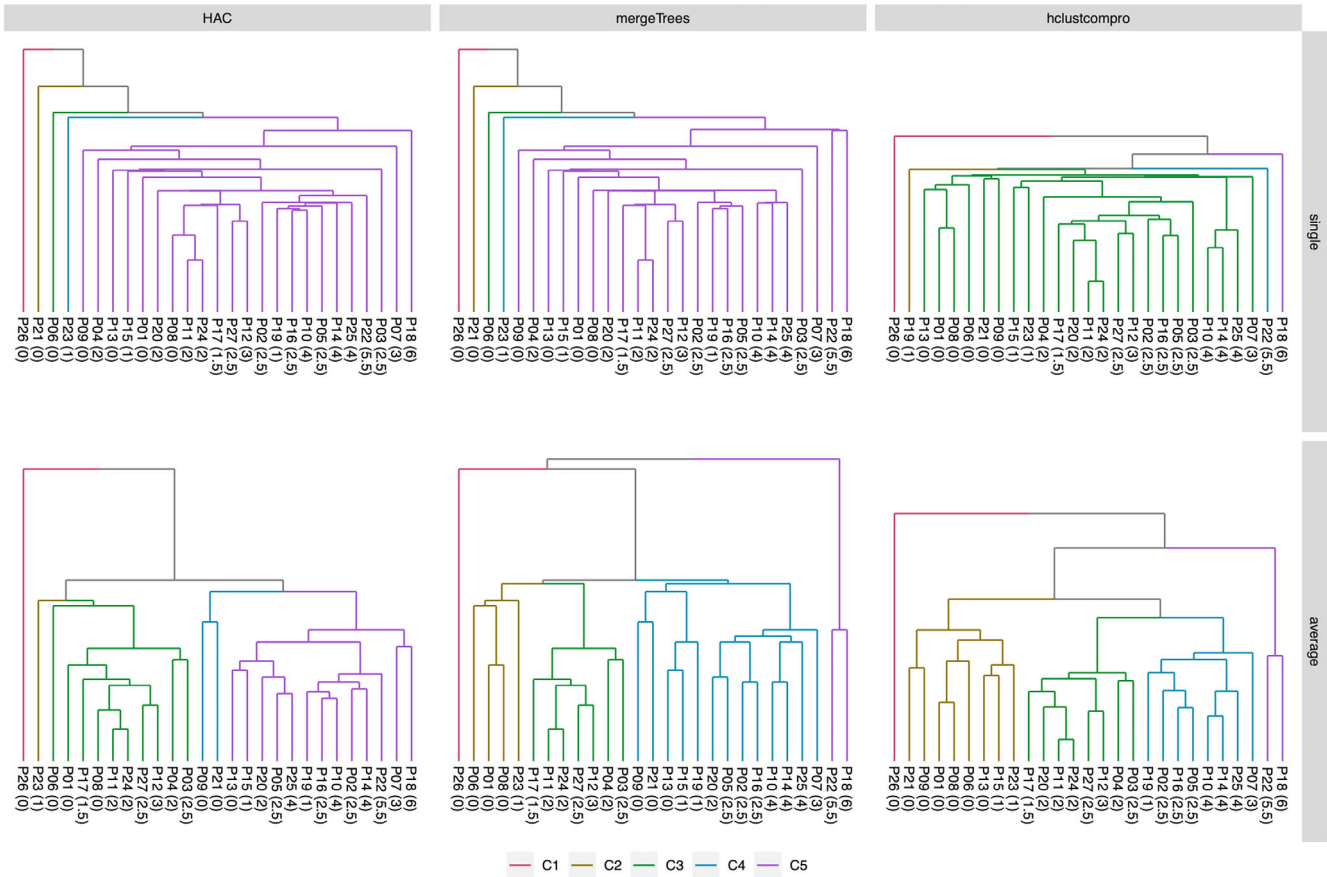
**B.2 Estimation of the hclustcompro weighting parameter**

The procedure of Equation (10) led to the estimated values  $\hat{\alpha} = 0.59$  (resp. 0.67) for the single linkage criterion (resp. average linkage). In essence, this means that we evaluate the dissimilarity between patients as a linear combination of the dissimilarity between their IGP (accounted for with a weight of 59% and 67% depending on the considered linkage criterion) and the dissimilarity between their EDSS scores (accounted for with a remaining weight of 41%, 33%). Interestingly, hclustcompro seems to put more information from the IGP data into the final dissimilarity matrix for linkage criteria that produced a large entanglement between EDSS data and IGP data. This suggests that hclustcompro naturally leans toward the IGP data when both data sources would separately produce different clustering structures.

**B.3 Choice of the optimal number of clusters**

Figure B2 displays the variation WSS, computed from Equation (12), as the number of clusters grows from 2 to 10. They are computed from partitions generated by cutting the dendrograms in Figure B3 at various heights in order to achieve the desired number of clusters.

In this section, we analyze both figures to determine a unique global optimal number of clusters. First, let us focus on the results obtained from single linkage. The corresponding dendrograms, visible in the first row of Figure B3, are symptomatic of the well-known tendency of this criterion to produce chained clusters.<sup>59</sup> This makes it hard to find a



**FIGURE B3** Dendrograms obtained by combining each of the three clustering methods (HAC, mergeTrees, and hclustcompro) with one of the three graph linkage criteria (single, average, and complete). Clusters are matched between dendrograms by ordering them as much as possible from smallest to largest median EDSS

suitable height threshold and usually leads to choosing either a very large number of clusters or only one. Given also that single linkage was the graph linkage criterion that produced an IGP-only dendrogram  $\mathcal{T}_1$  most similar to the EDSS-only dendrogram  $\mathcal{T}_0$ , we discard this linkage criterion in the remainder of the analysis. We can therefore focus on the results produced by average linkage, which we now comment by clustering method.

1. HAC on IGP data. Figure B3 suggests that the optimal number of clusters should be three. In effect, an elbow is visible on the WSS curves. The dendrogram displayed in the first column of Figure B3 with average linkage presents a structure relatively similar to that obtained with complete linkage (see Figure 8), and can be interpreted in the same way as in Section 4.3.
2. mergeTrees. Figure B2 indicates that four or five clusters should be chosen. Figure B3 shows that the method with average linkage produces a dendrogram very different from the one obtained with complete linkage (see Figure 8). However, as in the case of complete linkage, some patients in the same cluster present different overall disability, while some patients with similar overall disability are split in different groups.
3. hclustcompro. Figure B2 suggests that five clusters should be optimal. In effect, we can see that the slope of the WSS curve becomes less steep from this value. In Figure B3, we can appreciate how naturally the dendrograms suggest to split the patients into five groups. The dendrogram presents a singleton cluster.

As for complete linkage, we systematically generated 5 clusters for all methods. The resulting partitions can be visually appreciated in Figure B3 by the color coding scheme.

**B.4 Comparison of the clustering methods**

Table B1 describes the clusters from the perspective of internal data and is constructed in the same way as Table 3

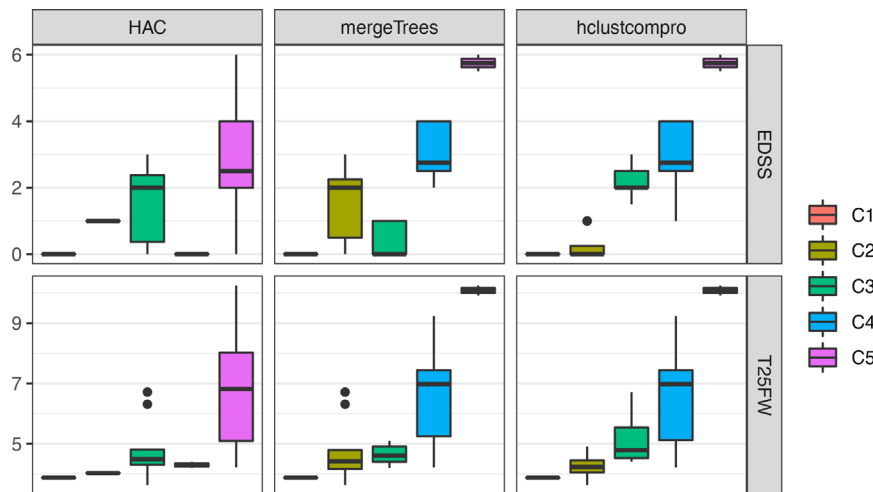
**TABLE B1** Summary of the clusters in terms of size ( $n$ ), proportion of within-cluster IGP inertia ( $p_W^{(k)}$ ) and median EDSS (EDSS)

Clustering	Linkage	C1			C2			C3			C4			C5		
		$n$	$p_W^{(1)}$	EDSS	$n$	$p_W^{(2)}$	EDSS	$n$	$p_W^{(3)}$	EDSS	$n$	$p_W^{(4)}$	EDSS	$n$	$p_W^{(5)}$	EDSS
HAC	Average	1	0.0	0	1	0.0	1	10	47.0	2	2	28.6	0	13	43.8	2.5
mergeTrees	Average	1	0.0	0	11	52.6	2	5	65.5	0	8	37.0	3	2	43.5	5.5
hclustcompro	Average	1	0.0	0	8	69.8	0	8	52.0	2	8	33.9	3	2	43.5	5.5

Note: For a given cluster, its proportion of within-cluster inertia is computed on the IGP data according to Equation (13).

**TABLE B2** Performance metrics of IGP data at the partition level

Clustering	Linkage	Using all clusters		Using only clusters that are never singletons	
		$p_W$	DI	$p_W$	DI
HAC	Average	34.0	0.377	42.5	0.377
mergeTrees	Average	42.7	0.288	49.2	0.288
hclustcompro	Average	45.9	0.192	42.1	0.467

**FIGURE B4** Cluster description in terms of standard clinical indicators of overall (EDSS) and gait-specific (T25FW) disability

Looking at the cluster sizes, we can notice that all the methods produce at least one singleton clusters, that is, clusters with a single patient. `hclustcompro` produces a sequence of clusters with increasing median EDSS, which is not observed for both `HAC` and `mergeTrees`.

Table B2 provides a performance assessment of the clustering methods at the granularity of the partition. It is constructed in the same way as Table 4.

When comparing the performance metrics in Table B2 with the ones displayed in Table 4, we can see that the clustering results obtained with complete linkage are better or at least similar to those obtained with average linkage. This is observed for all the three methods, and for the performance metrics computed using all five clusters and using only clusters that are never singletons.

Figure B4 displays the intra-cluster distributions of EDSS scores and the duration of the T25FW test for the three clustering methods (`HAC`, `hclustcompro`, `MergeTrees`) using average linkage criteria.

1. `HAC`. The clinical relevance of the clustering results is similar to that observed using complete linkage.
2. `mergeTrees`. The intra-cluster distributions of EDSS scores is more heterogeneous than those observed with complete linkage (see Figure 9). The resulting clusters do not present either significant or ordered differences in terms of walking speed.

3. `hclustcompro`. Clusters C3, C4, and C5 are the same for both average and complete linkage criteria, and are described in Section 4.4. With average linkage, cluster C1 is a singleton cluster. This isolated patient (P27) is also found in a singleton cluster in all other partitions except the one obtained with `hclustcompro` with complete linkage. Cluster C2 gathers patients with EDSS scores between 0 and 1, which presents the lowest walking time during the T25FW test.

## B.5 Conclusion

The clustering results obtained with HAC, `mergeTrees`, and `hclustcompro` using the criteria single and complete linkage are displayed in this section. In short, all the three methods produce chained clustering structure with single linkage, which are almost impossible to interpret. The clustering results produced by `mergeTrees` and `hclustcompro` using complete linkage present higher internal quality and clinical relevance than those obtained with average linkage. The clustering results are similar for HAC. For these reason, we chose to present only the results obtained with complete linkage in the main body of this article.