CATÓLICA MEDICAL SCHOOL
LISBOA

Affinity Coefficient vs. Euclidean Distance

in Hierarchical Clustering of Patients with Alcohol Use Disorder

Leonor Bacelar-Nicolau · Áurea Sousa · Sónia Ferreira · Cristina Ribeiro
Ana Paula Nascimento · Helena Bacelar-Nicolau

X Workshop on Computational Data Analysis and Numerical Methods
WCDANM 2026

Guimarães, Portugal · June 11–13, 2026

CATÓLICA MEDICAL SCHOOL
LISBOA

Outline

- ✓ Study context & design
- ✓ Statistical approach
- ✓ The generalised affinity coefficient
- ✓ Dendrogram comparison
- ✓ Partition quality — STAT
- ✓ Results: 2-cluster partitions
- ✓ Partition agreement (ARI)
- ✓ Cluster characterisation
- ✓ Conclusions

Book of abstracts of the X Workshop on Computational Data Analysis and Numerical Methods pages 55-56

Affinity Coefficient vs. Euclidean Distance in Hierarchical Clustering of Patients with Alcohol Use Disorder

Leonor Bacelar-Nicolau[1], Áurea Sousa[2], Sónia Ferreira[3,4], Cristina Ribeiro[5], Ana Paula Nascimento[6] and Helena Bacelar-Nicolau[7]

[1]Center for Interdisciplinary Research in Health (CIIS), Católica Medical School, Universidade Católica Portuguesa, Lisboa, Portugal
[2]Faculty of Sciences and Technology, CEEAplA and OSEAN, Universidade dos Açores, Ponta Delgada, Portugal
[3]Unidade de Tratamento e Reabilitação de Alcoólicos, Unidade Local de Saúde de São José, Lisboa, Portugal
[4]Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[5]Instituto de Medicina Preventiva e Saúde Pública, Clínica Universitária de Medicina Geral e Familiar, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[6]RISE-Health, Center for Translational Health and Medical Biotechnology (TBIO), Escola Superior de Saúde (E2S), Polytechnic of Porto (P. Porto), Porto, Portugal
[7]Faculty of Psychology, Institute of Environmental Health (ISAMB/FM-UL), Universidade de Lisboa, Lisboa, Portugal

Study Context & Design

Clinical Setting – Prospective Cohort Study

65 patients with **Alcohol Use Disorder (AUD)**
Alcohology & New Addictions Service, Lisbon

→ **N = 63 analysed** *(2 excluded as outliers)*

Treatment groups:

NRG (n = 21)	WTG (n = 42)
Neuropsychological Rehabilitation Group	*Weekly Therapeutic Group*

3 assessment moments:

M1 — Baseline
M2 — 3 months
M3 — 6 months

Neuropsychological Battery (13 measures)

General executive function *FAB (Frontal Assessment Battery)*
Cognitive flexibility *TMT (Trail Making Test) · WCST*
Working memory *Letters & Numbers*
Processing speed *Codes*
Verbal fluency *FAS*
Inhibition *Stroop Color-Word Test*

AIM: Use hierarchical cluster analysis to differentiate between 2 clinical groups NRG amd WTG

2 sets of variables:

39 raw score variables
39 difference score variables
M2-M1 M3-M1 M3-M2

S. Ferreira, L. Bacelar-Nicolau, M. Oliveira, S. Pombo, E. Vásquez-Justo, and C. Ribeiro. Executive Functions in Alcohol Use Disorder: The Positive Role of Neuropsychological Rehabilitation — Prospective Cohort Study. *Drug and Alcohol Review*. 2026;45(4):e70154.
https://doi.org/10.1111/dar.70154

Photos: https://unsplash.com/

CATÓLICA MEDICAL SCHOOL
LISBOA

Statistical Approach

4 Hierarchical Clustering Solutions - Complete Linkage

	Raw scores	Difference scores
Affinity coefficient	**S1**	**S3***
Euclidean distance	**S2**	**S4**
	Positive Values	**Positive and Negative Values**

** Only S3 differentiates NRG from WTG*

Quality and Association Metrics

1 STAT coefficients

Evaluates all k = 2 … 62 partitions. Highest value = best-supported partition. Used to select k = 2 as external validation of NRG/WTG assignment.

2 Fisher's exact tests

Tests homogeneity between treatment group (NRG/WTG) regarding 2-cluster partition at $\alpha = 0.05$.

3 Tanglegram & entanglement coefficient

Compares full dendrogram structure between affinity and Euclidean solutions. Entanglement: 0 = identical · 1 = maximum disagreement.

4 Adjusted Rand Index (ARI)

Quantify agreement between the best partitions from each method.

CATÓLICA MEDICAL SCHOOL
LISBOA

The Generalised Affinity Coefficient

Why a generalisation is needed

Standard affinity coefficient

Defined for non-negative data. Normalises row profiles to compare shape — ignores magnitude.

Difference scores can be negative

Δ = score(M2) − score(M1) is negative when performance declines — outside the domain of the standard version.

Generalised affinity coefficient

Extended to real-valued data by incorporating the sign of observations. Enables trajectory-based clustering on change scores.

What it captures in this study

Groups by trajectory pattern, not magnitude

Two patients with similar patterns of improvement/decline across measures and time points cluster together — even if their absolute score levels differ.

Reduces influence of correlated variables

Row normalisation mitigates redundancy from correlated neuropsychological measures.

Key contrast with Euclidean distance

Euclidean distance-based clustering is sensitive to magnitude and variance, potentially grouping patients with similar overall score levels but distinct recovery trajectories into the same cluster

H. Bacelar-Nicolau, F. Nicolau, Á. Sousa, and L. Bacelar-Nicolau. Measuring Similarity of Complex and Heterogenous Data in Clustering of Large Data Sets. *Biocybernetics and Biomedical Engineering*. 2009;29(2):9–18. https://www.scopus.com/pages/publications/70450253231

A. P. Nascimento, A. Oliveira, B. M. Faria, R. Pimenta, M. Vieira, C. Prudêncio, and H. Bacelar-Nicolau. Affinity Coefficient for Clustering Autoregressive Moving Average Models. *Computational and Mathematical Methods*. 2024;5540143, 13 pages. https://doi.org/10.1155/2024/5540143

CATOLICA MEDICAL SCHOOL
LISBOA

The Generalised Affinity Coefficient

Assumption: Each k-th data unit $(k = 1, \ldots, n)$ is described by $(x_{k1}, \ldots, x_{kp})$, and represents one point in the whole $\mathbb{R}^p$ space.

Associated with each *k-th* data unit is a profile defined as $\left(x_{k1}/x_{k\bullet}, \ldots, x_{kp}/x_{k\bullet}\right)$, where $x_{k\bullet} = \sum_{j=1}^{p} |x_{kj}|$.

Generalised affinity coefficient

The *Two-way generalised affinity coefficient* for real valued data, between two-data units, k and k' is given by:

$$a_g(k,k') = \sum_{j=1}^{p} \pi_j sign\left(\frac{x_{kj}}{x_{k\bullet}}\right) sign\left(\frac{x_{k'j}}{x_{k'\bullet}}\right) \sqrt{\left|\frac{x_{kj}}{x_{k\bullet}} \cdot \frac{x_{k'j}}{x_{k'\bullet}}\right|}$$

where $x_{kj}, x_{k'j}$ are real values in the whole $\mathbb{R}$ space, π_j are weights such that $0 \leq \pi \leq 1$, $\Sigma\pi_j = 1$. Normalisation constraints are, for each *k-th* profile: $\sum_1^p \left|x_{kj}/x_{k\bullet}\right| = 1$, where $x_{k\bullet} = \sum_{j=1}^{p} |x_{kj}|$.

Standard affinity coefficient

The standard affinity coefficient, $a(k,k')$, between any pair of data units $k, k' \in D$ described by $(x_{k1}, \ldots, x_{kp})$ and $(x_{k'1}, \ldots, x_{k'p})$, respectively, is given by:

$$a(k,k') = \sum_{j=1}^{p} \pi_j \sqrt{\frac{x_{kj}}{x_{k\bullet}} \cdot \frac{x_{k'j}}{x_{k'\bullet}}}$$

where x_{kj} and $x_{k'j}$ are general non-negative real values of the $n \times p$ data matrix $\mathbf{X}$; $x_{k\bullet} = \sum_{j=1}^{p} x_{kj}$, $x_{k'\bullet} = \sum_{j=1}^{p} x_{k'j}$; π_j are weights satisfying the usual conditions $0 \leq \pi_j \leq 1$, $\sum_{j=1}^{p} \pi_j = 1$; *and* $(x_{kj}/x_{k\bullet}, j = 1, \ldots, p)$ represents the *k-th* data unit *profile*.

Euclidean distance

The Euclidean distance between two data units k and k' is:

$$d(k,k') = \sqrt{\sum_{j=1}^{p} (x_{kj} - x_{k'j})^2}$$

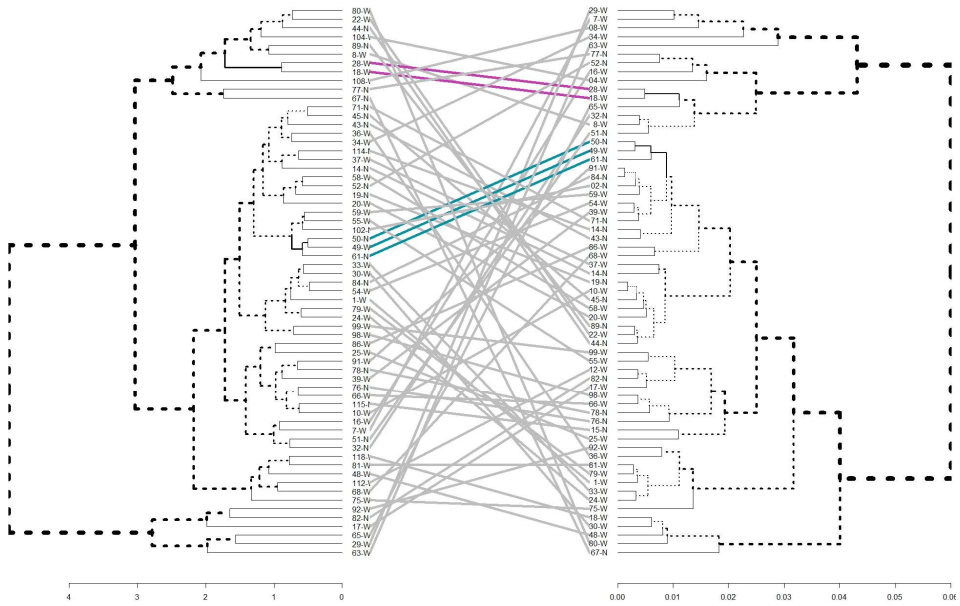
(standardized data)

CATÓLICA MEDICAL SCHOOL
LISBOA

Dendrogram Comparison — Tanglegrams

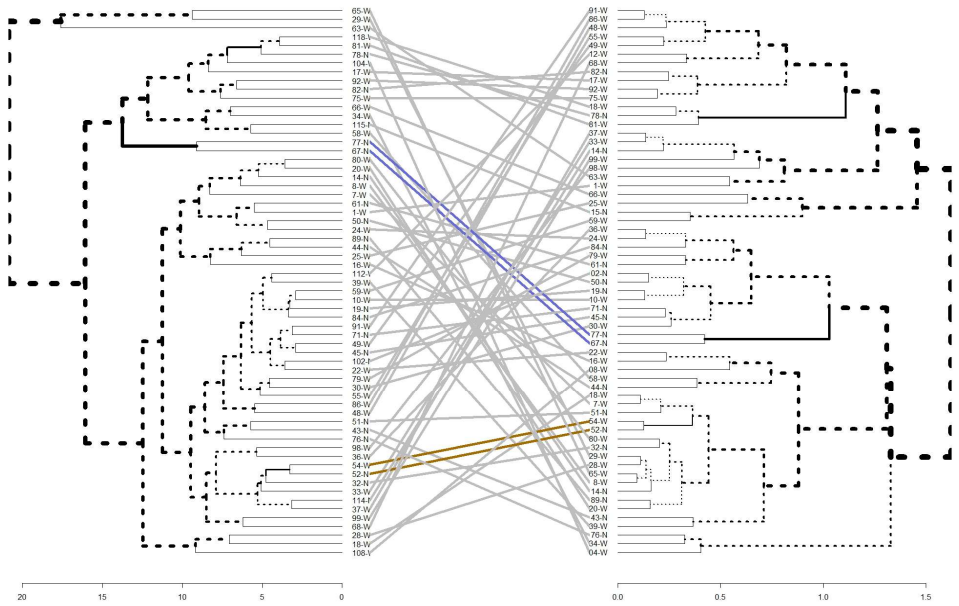
Entanglement coefficient: 0 = identical tree structure · 1 = maximum disagreement

Raw scores: Euclidean (left) vs Affinity (right)



Entanglement = **0.485** Moderate-high disagreement

Differences: Euclidean (left) vs Affinity (right)



Entanglement = **0.562** Higher disagreement on diffs

CATÓLICA MEDICAL SCHOOL
LISBOA

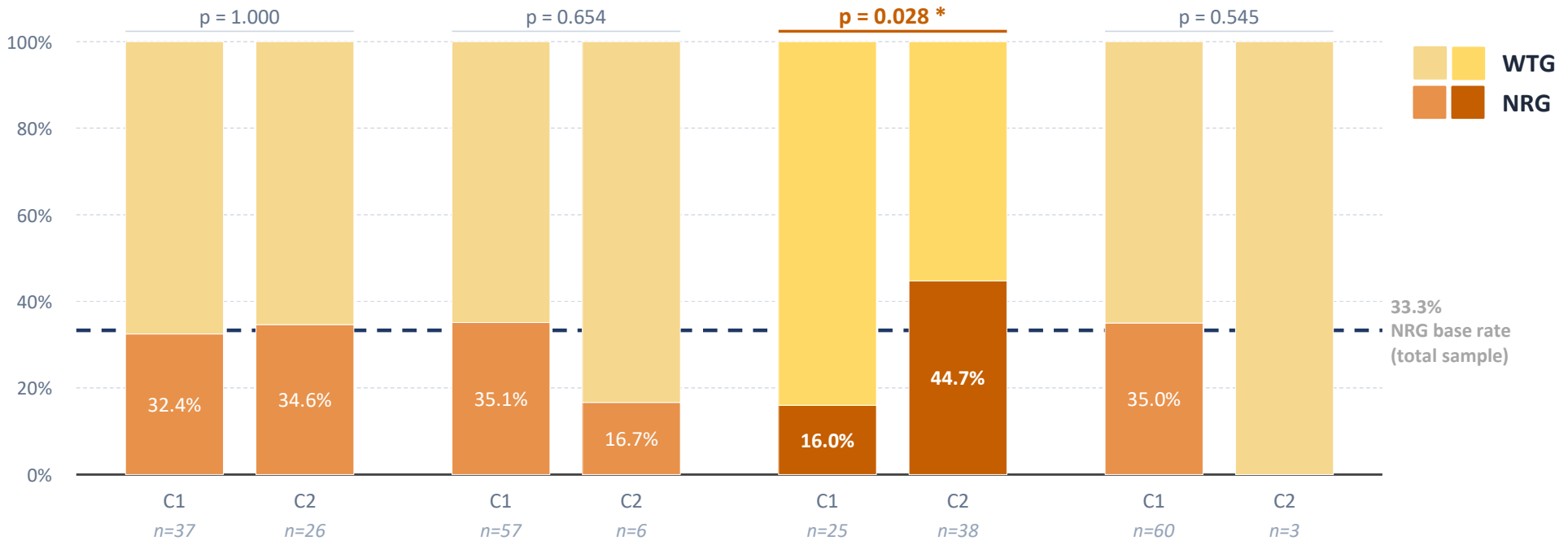
Partition Quality — STAT Coefficient

Solution	STAT at k = 2 (level 61)	Global STAT max (level — optimal k)
S1: Affinity (raw)	18.27	20.01 (level 60 — k = 3)
S2: Euclidean (raw)	25.54	33.37 (level 52 — k = 11)
S3: Affinity (diffs)	**14.65**	**22.45 (level 59 — k = 4)**
S4: Euclidean (diffs)	21.74	27.19 (level 56 — k = 7)

- STAT values not directly comparable across solutions — different proximity scales
- S3 uniquely achieves its STAT optimum at k = 4 — independent data-driven support for this partition
- k = 2 examined as external validation: can each method recover the known NRG/WTG assignment?

CATÓLICA MEDICAL SCHOOL
LISBOA

2-Cluster Solutions — NRG/WTG Composition



* Significant at 5%
All p-values from
Fisher Exact Tests (2-sided)

CATÓLICA MEDICAL SCHOOL
LISBOA

2-Cluster Solutions — NRG/WTG Composition (Table)

	WTG		NRG				
Solution / Cluster	**n**	**Row %**	**n**	**Row %**	**Total n**	**Total %**	**Fisher p**
S1: Affinity – Raw scores							
Cluster 1	25	67.6%	12	32.4%	37	58.7%	
Cluster 2	17	65.4%	9	34.6%	26	41.3%	p = 1.000
S2: Euclidean – Raw scores							
Cluster 1	37	64.9%	20	35.1%	57	90.5%	
Cluster 2	5	83.3%	1	16.7%	6	9.5%	p = 0.654
S3: Affinity – Differences							
Cluster 1	21	84.0%	4	16.0%	25	39.7%	
Cluster 2	21	55.3%	17	44.7%	38	60.3%	**p = 0.028***
S4: Euclidean – Differences							
Cluster 1	39	65.0%	21	35.0%	60	95.2%	
Cluster 2	3	100.0%	0	0.0%	3	4.8%	p = 0.545
Total	42	66.7%	21	33.3%	63	100%	

** p = 0.028, significant at α = 0.05*

CATÓLICA
MEDICAL SCHOOL
LISBOA

Partition Agreement — Adjusted Rand Index (ARI)

Method comparison (same data representation)

Comparison between 2-cluster partitions	ARI	Interpretation
S1 vs S2 (Affinity vs Euclidean — Raw)	**0.005**	*Near zero — no agreement*
S3 vs S4 (Affinity vs Euclidean — Diffs)	**0.071**	*Low agreement — structurally distinct*

Both comparisons show near-zero ARI — affinity and Euclidean solutions are structurally distinct, regardless of data representation.

Agreement with clinical group assignment

Partition	ARI	Interpretation
S1: Affinity (Raw scores)	**-0.011**	*No agreement*
S2: Euclidean (Raw scores)	**-0.038**	*No agreement*
S3: Affinity (Diffs)	**0.025**	*Marginal — highest*
S4: Euclidean (Diffs)	**-0.023**	*No agreement*

All ARI values are close to zero — no clustering solution recovers the NRG/WTG assignment at the binary level.

S3 achieves the highest ARI (0.025) — consistent with Fisher p = 0.028: a statistically detectable but modest structural association.

ARI = 0: agreement at chance level
ARI = 1: perfect agreement
ARI < 0: less agreement than chance

CATÓLICA MEDICAL SCHOOL
LISBOA

Cluster Characterisation — S3 2-Cluster Solution

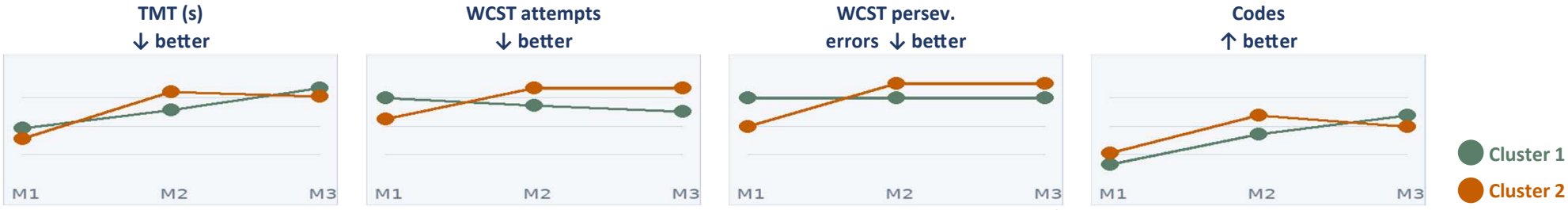
Cluster 1 (n = 25 · 84% WTG · 16% NRG)

- Steady continuous improvement in processing speed (Codes) and cognitive flexibility (TMT) across all 3 moments
- No improvement in WCST perseverative errors — card sorting flexibility does not change
- Better TMT trajectory than C2 in the M2→M3 interval (−29s vs −3s)

Cluster 2 (n = 38 · 55% WTG · 45% NRG)

- Large early improvement in WCST from M1→M2: fewer attempts (−13), fewer perseverative (−3) and non-perseverative errors (−4)
- Starts worse on WCST at baseline (more attempts, more perseveration); ends better than C1 on all WCST measures
- TMT improves markedly M1→M2 (−40s) then stabilises — early gain rather than continuous improvement

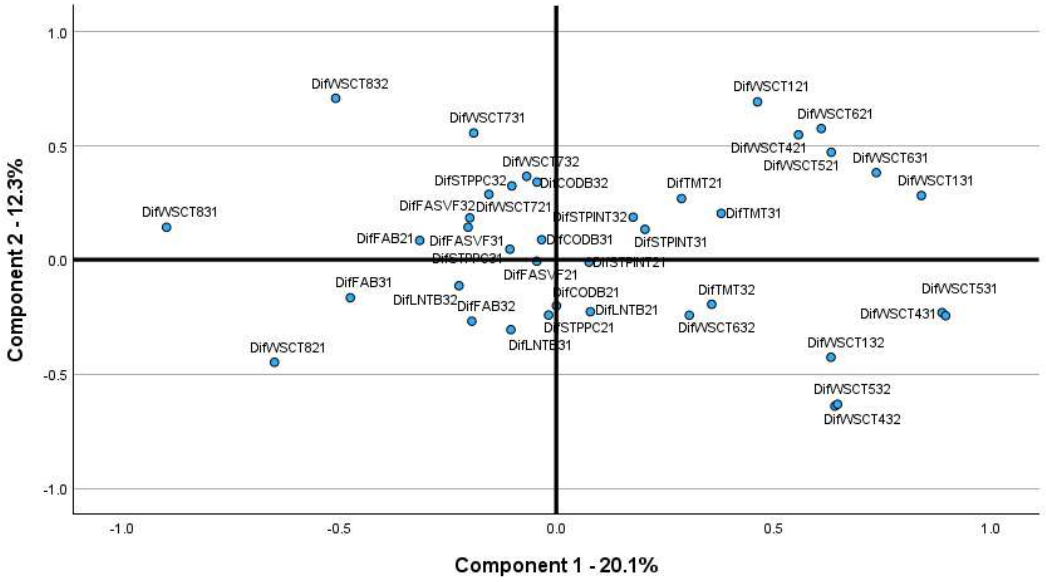
Selected key trajectory contrasts (medians)



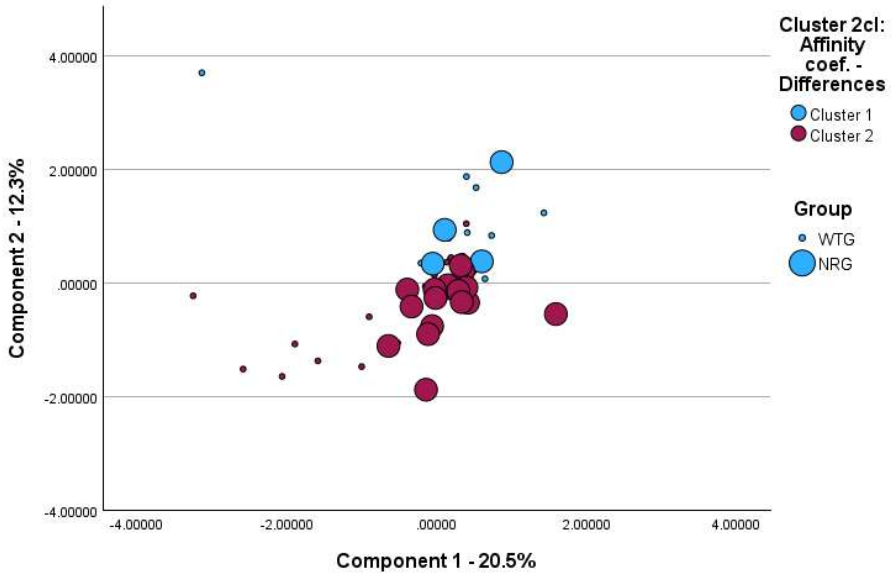
Higher position = better performance on each chart

Cluster Characterisation — S3 2-Cluster Solution with PCA

Principal Components Plot - Variables



Principal Components Plot - Patients



Conclusions

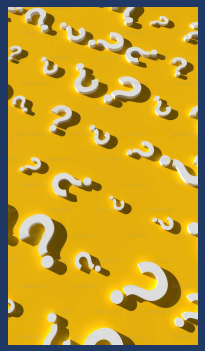
High entanglement — affinity and Euclidean produce substantially different tree structures
Raw scores: entanglement = 0.485 · Differences: entanglement = 0.562. Greater disagreement when applied to difference scores.

Only S3 (affinity on difference scores) differentiates NRG from WTG
The 2-cluster partition of S3 is the only one reaching significance (Fisher $p = 0.028$, $\alpha = 0.05$): C1 = 84% WTG · C2 = 45% NRG. S4 on the same data: $p = 0.545$.

The difference is attributable to the proximity measure, not the data representation
S3 and S4 both use difference scores. Only S3 (affinity) detects the NRG/WTG structure — because it groups by trajectory pattern rather than magnitude of change.

ARI confirms structural distinctiveness between methods
Near-zero ARI between all method pairs. S3 achieves the highest agreement with clinical grouping (ARI = 0.025) — consistent with Fisher $p = 0.028$.

The generalised affinity coefficient extends the method to real-valued data
Difference scores may be negative. The generalised version handles this naturally, enabling trajectory-based clustering unavailable with the standard affinity or Euclidean distance.



Thank you!

CATÓLICA MEDICAL SCHOOL
LISBOA

Leonor Bacelar-Nicolau
lnicolau@ucp.pt

Work in Progress!
Any suggestions are most welcome!

Funding: FCT UID/04279/2025 · UIDB/00685/2025

Photos: https://unsplash.com/