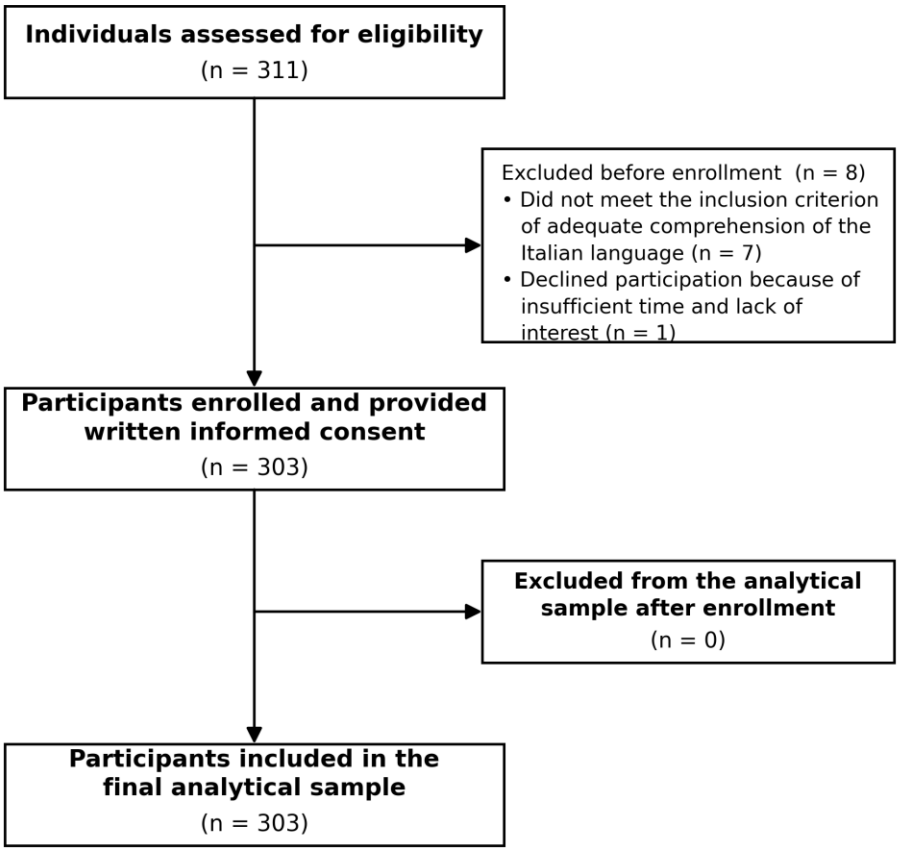
Supplementary File 1. Flow diagram of participant recruitment, Supplementary Methods and Additional Analyses

Content

Supplementary Figure S1. Flow diagram of participant recruitment and inclusion in the final analytical sample.



Note. Of the 311 individuals assessed for eligibility, seven did not meet the inclusion criterion regarding adequate comprehension of the Italian language. Of the remaining 304 eligible individuals invited to participate, one declined because of insufficient time and lack of interest, corresponding to a refusal rate of 0.3%. The final analytical sample therefore comprised 303 participants.

Supplementary Methods

S1. Monte Carlo assessment of sample-size adequacy

The simulation was designed as a stability-based assessment of sample-size adequacy for an exploratory clustering procedure, not as a conventional hypothesis-testing power calculation. Synthetic datasets were generated for sample sizes N = 150, 200, 250, 300, 350, and 400 and true numbers of clusters K = 2–10. Each scenario contained six continuous variables, balanced cluster sizes, independent within-cluster variances equal to 1, and moderate centroid scaling (the code parameter d = 0.50). Cluster centroids were first positioned on a circle in a two-dimensional latent space and then projected across the six observed variables using normalized, exponentially decaying weights.

Within every simulated dataset, variables were z-standardized, Euclidean distances were calculated, and Ward's minimum-variance hierarchical clustering (ward.D2) was fitted. The number of clusters was reselected as the value of k between 2 and 10 that maximized the average silhouette width. The full-data solution served as the reference partition. For each dataset, 200 subsamples containing 80% of participants were drawn without replacement, and the complete standardization, clustering, and k-selection procedure was reapplied. Stability was quantified by the adjusted Rand index (ARI) between the reference labels restricted to the subsample and the labels estimated in that subsample.

For each scenario, 50 independently simulated datasets were evaluated. The recorded endpoints were mean ARI across the 200 subsamples, the proportion of subsamples meeting prespecified ARI thresholds, and the proportion of subsamples in which the selected k equaled the full-data reference k. The random-number seed was 123. Analyses were performed in R 4.5.1 using cluster 2.1.8.1 and mclust 6.1.2.

After the initial broad assessment, a fine-grid sensitivity analysis evaluated N = 250–350 in increments of 10. The same data-generating mechanism and the same end-to-end procedure were used: k was selected by maximum average silhouette width in the full dataset and was independently reselected from k = 2–10 in every 80% subsample. The fine-grid analysis summarized mean ARI and the proportions of resampled partitions with ARI ≥ 0.50, ≥ 0.60, and ≥ 0.70.

Table S1. Parameters and estimands of the Monte Carlo assessment

Component	Specification	Purpose
Sample sizes	Initial grid: 150–400; fine grid: 250–350 in increments of 10	Evaluate stability and refine the region around N = 300
True clusters	K = 2–10	Represent candidate structural complexity
Observed variables	P = 6 continuous variables	Synthetic multivariate feature space
Cluster sizes	Balanced	Hold size imbalance constant
Centroid scaling / variance	d = 0.50; SD = 1	Scale latent centroids; equal spherical variance
Clustering pipeline	z-scores; Euclidean distance; Ward.D2	Estimate hierarchical partitions
Selection of k	Maximum average silhouette, k = 2–10	Repeat data-driven model selection
Outer replications	50 per N × K scenario	Average across generated datasets
Inner resampling	200 subsamples; 80%; no replacement	Assess within-dataset stability
ARI thresholds	0.50, 0.60, and 0.70 in the fine-grid analysis	Describe stability across increasing agreement levels
Other endpoints	Mean ARI	Describe agreement and k-selection stability
Seed	123	Support computational reproducibility

Note. ARI = adjusted Rand index; SD = standard deviation. In the supplied generator, d scales the radius of latent centroids and is not a common pairwise Cohen's d for all cluster pairs. In the fine-grid analysis, ARI thresholds of 0.50, 0.60, and 0.70 were summarized.

Table S2. Fine-grid Monte Carlo stability estimates around N = 300

N	Mean ARI	Proportion ARI ≥ 0.50	Proportion ARI ≥ 0.60	Proportion ARI ≥ 0.70
250	0.583	0.610	0.430	0.240
260	0.591	0.630	0.460	0.260
270	0.601	0.660	0.500	0.290
280	0.612	0.690	0.550	0.320
290	0.624	0.720	0.600	0.360
300	0.638	0.760	0.660	0.410
310	0.641	0.770	0.670	0.420
320	0.643	0.770	0.680	0.430
330	0.644	0.780	0.680	0.430
340	0.645	0.780	0.690	0.440
350	0.646	0.790	0.690	0.440

Note. ARI = adjusted Rand index. In every generated dataset, k was selected by maximum average silhouette width from k = 2–10 in the full sample and was independently reselected from k = 2–10 in each 80% subsample before ARI was calculated.

In the fine-grid analysis, stability improved most clearly between N = 250 and N = 300. At N = 300, mean ARI was 0.638, and the proportions of resampled solutions reaching ARI thresholds of 0.50, 0.60, and 0.70 were 0.760, 0.660, and 0.410, respectively. Beyond N = 300, incremental gains were small across the reported stability endpoints, supporting N ≈ 300 as a pragmatic point of diminishing returns rather than a strict adequacy threshold.

Interpretation. This stability-based simulation does not constitute statistical power for a null-hypothesis test and does not establish a formal cutoff. It quantifies the reproducibility of the complete data-driven clustering procedure, including reselection of k, under the specified synthetic scenarios. The simulation should therefore be interpreted together with the observed-data stability analysis below.

S2. Validity and stability in the observed dataset

For the empirical analysis, the 18 continuous variables listed in Section S3 were z-standardized. PCA preprocessing retained all 18 components, after which t-SNE projected the data to two dimensions using PCA initialization, Euclidean

distance, and perplexity 35. All checks reported below used the fixed embedding stored in the analytic dataset, thereby avoiding variation from rerunning the stochastic embedding.

The final analytic dataset contained 303 participants with two stored t-SNE coordinates. Ward.D2 hierarchical clustering of the stored coordinates reproduced the recorded five-cluster assignments exactly (ARI = 1.000). For each candidate solution from k = 2 to k = 10, average silhouette width, the Calinski–Harabasz index, and the minimum and maximum cluster sizes were calculated.

Table S3. Internal validity indices for candidate observed-data solutions

k	Average silhouette	Calinski–Harabasz index	Minimum cluster n	Maximum cluster n
2	0.426	296.1	127	176
3	0.378	265.5	57	127
4	0.326	243.9	54	119
5	0.351	257.1	54	73
6	0.370	283.5	22	73
7	0.358	283.8	22	59
8	0.360	291.5	22	59
9	0.371	310.7	22	54
10	0.374	314.6	16	54

Note. Higher silhouette and Calinski–Harabasz values indicate greater internal separation under their respective criteria. Indices were calculated from the same stored two-dimensional embedding used for the recorded partitions.

Observed-data stability of the retained k = 5 partition was assessed using 1,000 random subsamples containing 80% of participants, drawn without replacement (seed 123). Ward.D2 clustering with k fixed at 5 was refitted to each subsample. The ARI compared the refitted labels with the full-sample labels restricted to the same participants.

Table S4. Resampling stability of the observed five-cluster solution

B	Sampling fraction	Mean ARI (SD)	Median ARI (IQR)	2.5th–97.5th percentile	Proportion ARI ≥ 0.60
1000	0.80	0.593 (0.113)	0.604 (0.514–0.669)	0.372–0.811	0.513

Note. ARI = adjusted Rand index; B = number of subsamples; IQR = interquartile range; SD = standard deviation. This analysis evaluates stability of the fixed five-cluster partition on observed data and is distinct from the sample-size simulation.

S3. Selection framework, variable rationale, and collinearity assessment

The retained five-cluster solution was selected through an integrative evaluation rather than by a single prespecified numerical cutoff. The multidisciplinary author group jointly considered average silhouette width, dendrogram structure, cluster-size balance, and whether the resulting profiles showed distinct and clinically coherent combinations of self-care behaviors and clinical characteristics. The decision was reached by consensus. The two-cluster solution maximized silhouette width but combined heterogeneous self-care profiles; solutions above five primarily subdivided existing branches and generated smaller groups. Because k = 5 did not maximize every internal validity index, it is reported as an exploratory compromise between separation, granularity, and interpretability rather than as a uniquely optimal partition.

The 18 continuous clustering variables were selected to span four complementary domains: demographic and clinical burden (age, BMI, and number of comorbidities); diabetes self-care (SCODI maintenance, monitoring, management, and self-efficacy); metabolic control (fasting glucose, HbA1c, total cholesterol, HDL cholesterol, LDL cholesterol, and triglycerides); and hepatic and renal status (ALT, AST, GGT, creatinine, and eGFR). Biomarkers were included to place behavioral profiles in their glycemic, metabolic, and end-organ context. Categorical sociodemographic characteristics were not used to define distances because their coding and weighting would require additional assumptions; they were instead used to describe and compare the derived clusters.

Pairwise Spearman correlations among the 18 clustering variables were examined using all available pairs. Three pairs had $|\rho| \geq 0.70$: total and LDL cholesterol ($\rho = 0.855$), creatinine and eGFR ($\rho = -0.836$), and SCODI monitoring and self-efficacy ($\rho = 0.713$). These variables were retained because they represent clinically distinct constructs, while z-standardization and dimensionality reduction limited dominance by measurement scale. The correlation screen is descriptive and does not eliminate the possibility of nonlinear redundancy.

t-SNE was chosen to preserve nonlinear local-neighborhood structure in a mixed clinical and behavioral feature space. PCA provides a linear projection and may not capture such local structure, whereas latent profile analysis requires distributional and model-form assumptions that were not central to this exploratory analysis. Nevertheless, clustering a t-SNE embedding can distort global distances; therefore, the resulting profiles should be interpreted as exploratory,

and the observed-data validity and stability results are reported to make this limitation transparent. No formal model-based latent-profile comparison was performed.

S4. Additional details of the dimensionality-reduction and clustering workflow

Dimensionality reduction was performed using Orange Data Mining (Orange3), whereas hierarchical clustering, cluster validation, descriptive analyses, and between-cluster comparisons were conducted in R version 4.5.1. The analytical matrix comprised 18 continuous variables covering demographic and clinical burden, diabetes self-care, metabolic control, and hepatic and renal function, as detailed in Section S3.

In Orange, the variables were centered and scaled by their standard deviations using the “Normalize data” option. Missing values required for PCA were replaced by the corresponding variable means through the preprocessing procedure embedded in the Orange PCA implementation. PCA preprocessing was enabled and all 18 components were retained. Thus, PCA was used to preprocess the analytical matrix and initialize t-SNE rather than to select a smaller subset of components.

A single two-dimensional t-SNE embedding was generated using PCA initialization, Euclidean distance, a perplexity of 35, and an exaggeration parameter of 2.50. The multiscale “Preserve global structure” option was disabled. The resulting two t-SNE coordinates were exported and stored in the analytical dataset. The same fixed embedding was used for every candidate cluster solution, and t-SNE was not rerun separately for each value of k.

In R, Euclidean distances were calculated directly from the two stored t-SNE coordinates, without further standardization. Hierarchical agglomerative clustering was performed using Ward’s minimum-variance criterion (ward.D2). A single hierarchical tree was fitted and subsequently cut at each value from k=2 to k=10, generating a nested family of candidate partitions on the common embedding. Candidate solutions were evaluated using average silhouette width, the Calinski–Harabasz index, dendrogram structure, cluster-size balance, and clinical interpretability. The retained k=5 assignments were subsequently used for cluster characterization and between-cluster comparisons. No additional imputation was performed for descriptive or comparative analyses, which were based on the available observations for each variable or test.

Rationale for PCA-preprocessed t-SNE before clustering

The use of PCA-preprocessed t-SNE was chosen to obtain a compact nonlinear representation of local similarity patterns among participants, rather than only to produce a graphical display (van der Maaten and Hinton, 2008; Kobak and Berens, 2019). PCA was not ignored as an alternative dimensionality-reduction approach: it was used as a preprocessing and initialization step before t-SNE, and all 18 principal components were retained. Therefore, no variance was intentionally discarded before the nonlinear embedding.

This strategy was considered appropriate because the clustering variables combined demographic, clinical, metabolic, renal, hepatic, and self-care dimensions that may define participant profiles through complex local patterns rather than through a small number of linear axes of maximal variance. While PCA is useful for summarizing linear variance structure, t-SNE is specifically designed to preserve neighborhood relationships in high-dimensional data and may therefore help represent nonlinear similarity patterns that are relevant in exploratory person-centered profiling (van der Maaten and Hinton, 2008; Nguyen and Holmes, 2019; Arora et al., 2018; Linderman and Steinerberger, 2019).

At the same time, the known limitations of t-SNE were explicitly considered. The t-SNE map was not interpreted as a direct metric representation of clinical distance. In particular, inter-cluster distances, apparent cluster size, and spatial proximity in the two-dimensional embedding were not used as substantive clinical findings (Wattenberg et al., 2016). Cluster interpretation was instead based on the original sociodemographic, clinical, metabolic, and self-care variables.

The retained five-cluster solution was not selected by visual inspection of the t-SNE map alone. Candidate solutions were evaluated using average silhouette width, the Calinski–Harabasz index, dendrogram structure, cluster-size balance, resampling stability, and clinical interpretability. Thus, the t-SNE embedding served as a fixed neighborhood-preserving input space for hierarchical clustering, whereas the final interpretation and validation of the clusters relied on quantitative indices, stability assessment, and the original participant characteristics.

References

Arora S, Hu W, Kothari PK. An Analysis of the t-SNE Algorithm for Data Visualization. *Proceedings of the 31st Conference on Learning Theory*. 2018;75:1455–1462. Available from: https://proceedings.mlr.press/v75/arora18a.html

Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*. 2019;10:5416. doi: 10.1038/s41467-019-13056-x

Linderman GC, Steinerberger S. Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science*. 2019;1(2):313–332. doi: 10.1137/18M1216134

Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*. 2019;15(6):e1006907. doi: 10.1371/journal.pcbi.1006907

van der Maaten LJP, Hinton GE. Visualizing data using t-SNE. *Journal of Machine Learning Research*. 2008;9:2579–2605. Available from: https://www.jmlr.org/papers/v9/vandermaaten08a.html

Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. *Distill*. 2016;1(10):e2. doi: 10.23915/distill.00002

Table S5. Bonferroni-adjusted pairwise comparisons between clusters for categorical variables

CLUSTER	SEX		LOW INCOME		OCCUPATIONAL STATUS		PARTNERSHIP		SMOKING		NEPHROPATHY		COMORBIDITY		INSULIN THERAPY	
	Fisher	ChiSquare	Fisher	ChiSquare	Fisher	ChiSquare	Fisher	ChiSquare	Fisher	ChiSquare	Fisher	ChiSquare	Fisher	ChiSquare	Fisher	ChiSquare
1 vs 2	1.000	0.763	**0.012***	**0.010***	**0.003***	**0.002***	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.527	0.442
1 vs 3	1.000	1.000	0.138	0.082	0.359	0.206	1.000	0.809	1.000	1.000	1.000	1.000	1.000	1.000	0.769	0.658
1 vs 4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.017***	**0.018***	1.000	0.745	1.000	1.000	**<0.001***	**<0.001***
1 vs 5	1.000	1.000	1.000	1.000	1.000	0.732	1.000	1.000	1.000	1.000	1.000	0.679	1.000	1.000	**0.003***	**0.002***
2 vs 3	1.000	1.000	1.000	1.000	1.000	1.000	0.052	**0.033***	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 vs 4	0.119	0.103	**<0.001***	**<0.001***	**<0.001***	**<0.001***	0.089	0.083	0.183	0.143	1.000	0.723	0.064	0.068	0.357	0.334
2 vs 5	1.000	1.000	**0.026***	**0.015***	0.616	0.508	0.712	0.540	1.000	1.000	0.722	0.369	0.901	0.553	0.782	0.658
3 vs 4	1.000	1.000	**0.003***	**0.003***	**0.016***	**0.010***	1.000	1.000	0.933	0.663	0.424	0.267	0.372	0.254	0.305	0.252
3 vs 5	1.000	1.000	0.171	0.109	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.710	0.510
4 vs 5	0.259	0.177	1.000	1.000	0.068	0.060	1.000	1.000	0.877	0.736	**0.007***	**0.006***	1.000	1.000	1.000	1.000

Note. Pairwise comparisons between clusters were performed using Fisher's exact test and Pearson's chi-square test, with Bonferroni adjustment for multiple comparisons. Statistically significant results (adjusted p-values < 0.05) are indicated by an asterisk (*) and shown in bold.

Table S6. Bonferroni-adjusted pairwise comparisons between clusters for continuous variables

CLUSTER	AGE (years)	Comorbidities (n)	BMI (kg/m²)	T2DM duration (years)	HbA1c (mmol/mol)	Glucose (mg/dL)	TC (mg/dL)	HDL (mg/dL)	LDL (mg/dL)	TG (mg/dL)	Creatinine (mg/dL)	eGFR (mL/min)	SCODI Maint	SCODI Monit	SCODI Manag	SCODI SE
Global test	χ^2=68.03 df=4	χ^2=90.58 df=4	χ^2=22.48 df=4	χ^2=18.51 df=4	χ^2=35.92 df=4	χ^2=29.59 df=4	χ^2=91.61 df=4	χ^2=28.61 df=4	χ^2=47.26 df=4	χ^2=19.83 df=4	χ^2=52.27 df=4	χ^2=61.91 df=4	χ^2=25.68 df=4	χ^2=132.10 df=4	χ^2=159.10 df=4	χ^2=150.81 df=4
1 vs 2	**0.005***	**<0.001***	**0.005***	**0.015***	1.000	0.732	**<0.001***	0.787	**<0.001***	**0.015***	**<0.001***	**<0.001***	**0.004***	0.078	**0.016***	**0.006***
1 vs 3	0.056	0.093	**<0.001***	**0.024***	1.000	1.000	**0.028***	1.000	0.911	0.038	**0.007***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***
1 vs 4	**<0.001***	0.122	**0.028***	1.000	**0.004***	0.118	**0.020***	**0.021***	1.000	0.030	0.477	0.844	0.320	**<0.001***	**<0.001***	**<0.001***
1 vs 5	1.000	0.380	**0.013***	0.096	**0.007***	**0.016***	0.143	0.525	1.000	1.000	0.245	**0.011***	**<0.001***	**<0.001***	**<0.001***	**<0.001***
2 vs 3	1.000	**0.003***	0.864	1.000	1.000	1.000	0.038	0.324	**0.006***	1.000	1.000	1.000	1.000	**0.008***	**0.021***	**0.001***
2 vs 4	**<0.001***	**<0.001***	1.000	**0.013***	**<0.001***	**0.001***	**<0.001***	**<0.001***	**<0.001***	1.000	**<0.001***	**<0.001***	0.473	**<0.001***	**<0.001***	**<0.001***
2 vs 5	**0.001***	**<0.001***	1.000	1.000	**<0.001***	**<0.001***	**0.005***	1.000	**0.001***	**0.009***	0.146	1.000	1.000	**<0.001***	**<0.001***	**<0.001***
3 vs 4	**<0.001***	**<0.001***	0.197	**0.022***	**0.001***	**0.007***	**<0.001***	0.128	0.068	1.000	**<0.001***	**<0.001***	0.046	1.000	0.076	1.000
3 vs 5	**0.016***	**<0.001***	0.789	1.000	**0.001***	**0.001***	1.000	0.213	1.000	**0.023***	1.000	1.000	1.000	**<0.001***	**<0.001***	**<0.001***
4 vs 5	**0.005***	1.000	1.000	0.095	1.000	1.000	**<0.001***	**<0.001***	1.000	**0.018***	**0.001***	**<0.001***	0.079	**<0.001***	**0.001***	**<0.001***

Legend. Body mass index (BMI); type 2 diabetes mellitus (T2DM); glycated hemoglobin (HbA1c); total cholesterol (TC); high-density lipoprotein cholesterol (HDL); low-density lipoprotein cholesterol (LDL); triglycerides (TG); estimated glomerular filtration rate (eGFR); Self-Care of Diabetes Inventory (SCODI); maintenance (Maint); monitoring (Monit); management (Manag); self-efficacy (SE).

Note. Global differences across clusters were assessed using the Kruskal–Wallis test. Pairwise comparisons were performed using Dunn's test with Bonferroni adjustment for multiple comparisons. Statistically significant results (adjusted p-values < 0.05) are indicated by an asterisk (*) and shown in bold.