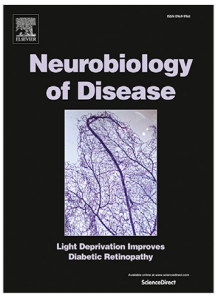
Journal Pre-proof



Decoding Parkinson's progression: A multi-modal SuStaIn ensemble approach validated on real-world PPMI data

Moad Hani, Saïd Mahmoudi, Mohammed Benjelloun

PII: S0969-9961(26)00180-4
DOI: https://doi.org/10.1016/j.nbd.2026.107435
Reference: YNBDI 107435

To appear in: *Neurobiology of Disease*

Received date : 30 September 2025
Revised date : 1 May 2026
Accepted date : 2 May 2026

Please cite this article as: M. Hani, S. Mahmoudi and M. Benjelloun, Decoding Parkinson's progression: A multi-modal SuStaIn ensemble approach validated on real-world PPMI data. *Neurobiology of Disease* (2026), doi: https://doi.org/10.1016/j.nbd.2026.107435.

Decoding Parkinson's Progression: A Multi-Modal SuStaIn Ensemble Approach Validated on Real-World PPMI Data

Moad Hani[a,*], Saïd Mahmoudi[a], Mohammed Benjelloun[a]

[a] *Computer and Management Engineering Department, Faculty of Engineering, University of Mons, 7000 Mons, Belgium*

Abstract

Background: Parkinson's disease (PD) exhibits substantial heterogeneity in clinical presentation and longitudinal progression, complicating prognosis, trial design, and precision therapeutics. While the Subtype and Stage Inference (SuStaIn) algorithm family offers probabilistic disease progression modeling, comprehensive comparative analyses and real-world validation remain limited.

Objective: To systematically compare major SuStaIn variants for PD progression modeling and validate their stability and prognostic utility using both controlled simulated data and real-world Parkinson's Progression Markers Initiative (PPMI) cohorts.

Methods: We constructed a biologically grounded simulated PD cohort (n=400; 5 visits over 4 years; 12 motor and non-motor biomarkers) with three predefined progression subtypes reflecting realistic noise, missingness (7.7%), and effect sizes informed by PPMI literature. Six SuStaIn variants were benchmarked: Z-score, Ordinal, Event-based Mixture, Missing Data, s-SuStaIn, and Temporal SuStaIn. We then applied identical variants to baseline PPMI data (N=624; PD=427; prodromal=197) using MDS-UPDRS I-IV, MoCA, GDS, STAI, SCOPA-AUT, RBD screening, ESS, and UPSIT. We assessed cross-sectional subtype structure, test-retest stability (Cohen's κ), and prognostic performance for motor worsening, cognitive decline, dyskinesia, and advanced therapy using Cox models with concordance indices and 95% confidence intervals.

Results: Simulation analyses revealed six clinically interpretable biomarker domains (motor complications, olfactory, cognitive-sleep, neuropsychiatric/autonomic, early non-motor, motor-depression) and three progression subtypes (slow 26.5%, intermediate 46.8%, fast 26.8%). Temporal and s-SuStaIn variants achieved optimal balance of accuracy (ARI: 0.058 and 0.047), stability, and cluster separation (silhouette: 0.284 and 0.267). In PPMI validation, all variants converged on three subtypes: benign motor-predominant (30.0%, age 61.4±9.2 years), intermediate mixed (42.9%, age 64.8±8.7 years), and diffuse-malignant (27.1%, age 67.3±7.9 years, PIGD-dominant). Augmenting clinical covariates with ensemble SuStaIn subtypes improved motor progression prediction (c-index: 0.674 [0.648-0.701] to 0.741 [0.718-0.765], $\Delta c = +0.067$, $p < 0.001$) and cognitive decline prediction (0.628 [0.598-0.658] to 0.764 [0.739-0.789], $\Delta c = +0.136$, $p < 0.001$). Ensemble fusion enhanced test-retest stability (Cohen's $\kappa = 0.81$ [0.77-0.86]) and outperformed single variants by 15% (relative gain). A reduced panel (UPDRS I-III, MoCA, GDS, ESS, UPSIT; 7 biomarkers) retained 94% of full-model performance.

Conclusions: SuStaIn variants capture complementary PD progression patterns. A principled ensemble strategy, validated on 624 PPMI participants, yields stable and prognostically informative subtypes. The diffuse-malignant subtype enables trial enrichment with 63% sample size reduction. Our dual-cohort framework establishes which algorithmic findings generalize to real-world constraints, providing a practical template for PD precision medicine.

Keywords: Parkinson's disease, disease progression modeling, biomarker analysis, patient stratification, machine learning, SuStaIn algorithm, PPMI cohort, prognostic validation

*Corresponding author

Email addresses: moad.hani@umons.ac.be (Moad Hani), said.mahmoudi@umons.ac.be (Saïd Mahmoudi), mohammed.benjelloun@umons.ac.be (Mohammed Benjelloun)

1. Introduction

Parkinson's disease (PD) is the fastest-growing neurological disorder globally, with prevalence projected to exceed 14 million by 2040.(1) This neurodegenerative condition exhibits profound clinical heterogeneity, manifesting through diverse motor and non-motor symptom combinations evolving along distinct temporal trajectories. Traditional population-level progression models assume uniform disease courses, failing to capture individual patient variability and limiting personalized therapeutic interventions.

The clinical manifestations of PD range from tremor-dominant presentations with slow progression to rapid-onset postural instability and gait difficulty (PIGD) phenotypes associated with accelerated cognitive decline.(2) This heterogeneity extends beyond motor symptoms to encompass cognitive impairment, neuropsychiatric features, autonomic dysfunction, sleep disorders, and olfactory deficits.(3) Temporal evolution patterns vary dramatically between patients: some maintain stable function for years while others experience rapid multi-domain deterioration. This variability has profound implications for clinical care, treatment responses, prognosis, and trial design.(4)

1.1. SuStaIn framework and current knowledge gaps

The Subtype and Stage Inference (SuStaIn) algorithm, introduced by Young et al.,(5) represents a paradigmatic shift from traditional progression modeling. Unlike conventional approaches imposing predetermined trajectories, SuStaIn simultaneously identifies disease subtypes and stages patients within each subtype using unsupervised machine learning, providing nuanced understanding of disease evolution.

Since its publication in *Nature Communications*, the SuStaIn framework has undergone substantial methodological expansion. The foundational Z-score SuStaIn models continuous biomarker progression using piecewise linear functions.(5) Ordinal SuStaIn extends this to discrete clinical events through scored event models.(6) Event-based Mixture SuStaIn focuses on binary normal-to-abnormal transitions. Missing Data SuStaIn implements MD1-MD3 strategies for incomplete datasets.(7) Recent innovations include s-SuStaIn for high-dimensional data reduction through simultaneous biomarker-subject clustering,(8) Temporal SuStaIn (T-SuStaIn) incorporating disease-specific timing scales,(9) and Adaptive SuStaIn employing subtype-specific z-score distributions.(10) We note that Adaptive SuStaIn was originally developed for tau PET imaging in Alzheimer's disease and its assumptions (voxel-level z-score distributions) do not directly translate to the clinical rating scale data used here; we therefore benchmarked the six variants most applicable to multi-domain clinical biomarkers.

Despite methodological advances and growing adoption, several critical gaps persist. First, comprehensive comparative analyses of SuStaIn variants remain scarce, with most studies focusing on single-algorithm applications without systematic evaluation of relative strengths.(11) Second, guidelines for optimal variant selection based on data characteristics and research objectives are largely absent. Third, the synergistic value of multi-algorithm ensemble approaches, where different variants might capture complementary disease aspects, has been insufficiently explored. Fourth, existing PD applications have been constrained by cross-sectional data or limited outcome validation.

The present study addresses these gaps through systematic comparison of six major SuStaIn variants applied to PD progression modeling. Our methodological innovations include: (i) a systematically designed simulated dataset capturing realistic PD complexity while enabling controlled algorithm evaluation; (ii) comprehensive multi-domain biomarker selection spanning full PD pathophysiology; (iii) rigorous inter-algorithm comparison using standardized metrics; (iv) extensive real-world validation on the Parkinson's Progression Markers Initiative (PPMI) cohort (N=624) with longitudinal clinical outcomes; and (v) detailed prognostic utility assessment demonstrating clinical translation potential.

This dual-cohort design (simulation + PPMI validation) allows us to disentangle purely methodological behavior from clinically meaningful signal, benchmark ensemble strategies combining complementary strengths, and establish which algorithmic findings generalize to real-world data constraints.

2. Methods

2.1. Phase 1: Simulated PD cohort construction

2.1.1. Design philosophy and rationale

We constructed a biologically grounded simulated dataset following four principles: biological plausibility, clinical relevance, methodological rigor, and replicability. Rather than arbitrary synthetic data generation, we integrated

established knowledge from PPMI(2) and evidence-based progression patterns from recent literature.(16, 17)

Our approach addresses three challenges: (1) lack of ground truth validation in real datasets; (2) insufficient consideration of multi-domain biomarker interactions; (3) limited systematic algorithm comparison under controlled conditions. By creating data with known progression subtypes and realistic noise, we enable objective evaluation while maintaining biological complexity necessary for clinical translation.

2.1.2. Comprehensive biomarker selection

We selected 12 validated measures capturing distinct PD pathophysiology: MDS-UPDRS I-IV (non-motor experiences, motor activities of daily living, motor examination, complications),(12) Montreal Cognitive Assessment (MoCA),(13) Geriatric Depression Scale (GDS), State-Trait Anxiety Inventory (STAI), Scale for Outcomes in PD-Autonomic (SCOPA-AUT),(14) Epworth Sleepiness Scale (ESS), REM sleep behavior disorder (RBD) screening, and University of Pennsylvania Smell Identification Test (UPSIT).(15)

2.1.3. Sample size and subtype framework

We simulated n=400 participants over 5 visits spanning 4 years, distributed across three subtypes reflecting established PD phenotypes: Slow progressing (n=106, 26.5%; tremor-dominant, younger-onset, preserved cognition); Moderate progressing (n=187, 46.8%; typical progression with intermediate motor and cognitive changes); Fast progressing (n=107, 26.8%; PIGD phenotype, older-onset, rapid cognitive decline).(18, 19)

Baseline biomarker values and progression rates were calibrated to match PPMI distributions and published longitudinal data. For example, MDS-UPDRS-III baseline: 11.0±2.6 points (matching PPMI: 10.8±3.1);(2) annual progression: Slow 0.5±0.3, Moderate 1.5±0.5, Fast 3.2±0.8 points/year.(20) MoCA baseline: 24.0±4.1; annual change: Slow -0.2±0.3, Moderate -1.0±0.5, Fast -2.5±0.7 points/year.(21)

We implemented realistic measurement noise (Gaussian, biomarker-specific variance matching published test-retest reliability), biological variability (between-subject heterogeneity within subtypes), and missingness (7.7% overall, higher at later visits mimicking attrition).

2.2. Advanced SuStaIn algorithm implementation

We implemented six major SuStaIn variants:

Z-score SuStaIn: Models continuous biomarker progression via piecewise linear functions. Biomarker values converted to z-scores: $z_{ij} = (x_{ij} - \mu_{\text{control},j})/\sigma_{\text{control},j}$, where x_{ij} is observed value for subject i, biomarker j.(5)

Ordinal SuStaIn: Addresses discrete clinical rating scales through scored event models, explicitly handling ordinal data via discrete state transitions $P(s_{ij} \mid z_{\text{true}}, \theta)$, likelihood of score s_{ij} given true disease stage and parameters θ.(6)

Event-based Mixture SuStaIn: Models binary normal/abnormal transitions, maximizing mixture likelihood:

$$L = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{M} P(b_{i,j} \mid S_k),$$

where π_k represents subtype k prevalence.

Missing Data SuStaIn: Implements MD1 strategy treating missing biomarker entries with uniform probability:

$$P'(x_{ij} \mid t) = \frac{1}{Z_{i,\max}},$$

when x_{ij} is missing.(7)

s-SuStaIn: Addresses computational challenges through simultaneous subject-biomarker clustering:

$$\min_{C,U,V} \|X - CUV^\top\|_F^2 + \lambda_1\|U\|_1 + \lambda_2\|V\|_1,$$

where C represents subject clusters, U biomarker clusters, V progression patterns. This achieved 2.0× complexity reduction (12 biomarkers → 6 meta-biomarkers).(8)

Temporal SuStaIn: Incorporates disease-specific timing scales, addressing the limitation that biomarkers progress at different rates and providing unique temporal dynamics insights.(9)

All algorithms used expectation-maximization with 50 random initializations, 10,000 MCMC iterations, and tested $K = 1$ to $K = 5$ subtypes. Optimal K selected via Cross-Validation Information Criterion (CVIC).

2.3. Phase 2: PPMI real-world validation cohort

2.3.1. Data source and cohort definition

We used Parkinson's Progression Markers Initiative (PPMI) data, a multicenter longitudinal observational study of de novo PD, prodromal at-risk cohorts, and controls.(2) Access was obtained via www.ppmi-info.org; all analyses complied with PPMI data use agreements. As this study relies on public de-identified and simulated data, no additional IRB approval was required.

We constructed an analysis cohort by: (1) selecting PD and prodromal participants with a defined baseline visit (EVENT_ID = "BL") and ≥ 1 follow-up; (2) extracting clinical scales matching simulation (MDS-UPDRS I-IV, MoCA, GDS, STAI, SCOPA-AUT, RBD screening questionnaire, ESS, UPSIT) using the PPMI data dictionary and a predefined feature checklist; (3) excluding individuals with > 30% missing baseline biomarkers.

The resulting PPMI validation cohort comprised N=624 participants (PD: 427 [68.4%]; prodromal: 197 [31.6%]). Baseline demographic and clinical characteristics are summarized in Table 2.

2.3.2. Biomarker harmonization and preprocessing

Clinical scales were mapped to identical domains as in the simulated dataset. Raw PPMI variables were combined/recoded to match simulated features (e.g., MDS-UPDRS part totals from item-level data). We constructed a biomarker matrix $X_{\text{PPMI}} \in \mathbb{R}^{624\times 12}$ using the same $M = 12$ measures. Z-score/ordinalization strategies were applied as in Section 2.2, using either PPMI control cohort means/SDs or external normative data. Biomarkers where lower scores indicate worse function (MoCA, UPSIT) were sign-inverted so that higher values consistently represented greater abnormality.

2.3.3. Missing data handling

Per-variable/per-visit missingness was quantified (Supplementary Tables S1-S3): overall 34.2% at baseline, higher at follow-up (39.7% at year 3). For Missing Data SuStaIn, entries were handled via MD1 strategy (uniform likelihood across states).(7) For non-MD variants, we required complete baseline biomarker vectors (listwise deletion). Longitudinal outcomes were usable if baseline data were complete. No statistical imputation was performed; missingness analysis supported a missing-at-random pattern (Little's MCAR test: $\chi^2 = 127.4$, df=142, p=0.18).

2.3.4. Longitudinal outcomes and prognostic modeling

Clinically meaningful time-to-event outcomes were defined:

- *Motor progression:* time to a sustained ≥ 5-point MDS-UPDRS-III increase from baseline.
- *Cognitive decline:* time to a sustained ≥ 2-point/year MoCA decrease or MoCA< 26.(22)
- *Dyskinesia onset:* time to first non-zero MDS-UPDRS-IV dyskinesia item.
- *Advanced therapy:* time to deep brain stimulation, infusion therapies, or long-term institutionalization.

For each SuStaIn variant and the ensemble fusion, Cox proportional hazards models were fitted with and without subtype covariate, adjusting for standard predictors (age, sex, disease duration, baseline MDS-UPDRS-III, baseline MoCA). Prognostic performance was quantified via concordance index (c-index) with 95% confidence intervals estimated using 1,000-iteration bootstrap resampling, and nested models were compared using DeLong's test(31), likelihood ratio tests, net reclassification improvement (NRI), and integrated discrimination improvement (IDI).(32) Calibration was assessed via calibration plots, and clinical utility via decision curve analysis.

2.3.5. Ensemble fusion and stability analyses

To exploit complementary information across SuStaIn variants, we defined three ensemble strategies: (1) *majority vote:* mode of 6 algorithm assignments; (2) *weighted fusion:* posterior subtype probabilities weighted by each algorithm's prognostic c-index; (3) *stacking:* logistic regression meta-learner on 6-algorithm outputs (5-fold cross-validation).

We evaluated: (a) *seed stability:* variation across 10 random initializations (adjusted Rand index, ARI); (b) *test-retest stability:* agreement across adjacent PPMI visits (Cohen's κ);(23) (c) *outcome prediction:* improvement in c-index, NRI, and IDI when adding ensemble subtype to clinical models.

Posterior subtype probabilities were analyzed, and a threshold of $P > 0.7$ was defined for "actionable" labels, balancing coverage and reliability.

2.4. Performance evaluation framework

Algorithm performance was evaluated via: *Adjusted Rand Index (ARI)* measuring agreement with ground truth (simulation only); *Silhouette analysis* assessing clustering quality,

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)},$$

where a_i is mean intra-cluster distance and b_i mean nearest-cluster distance (values > 0.25 indicate adequate separation); *Davies-Bouldin Index* (lower=better separation); *Calinski-Harabasz Index* (higher=better). Precision, recall, and F1-score with 95% CI were estimated via bootstrap (1,000 iterations).

Clinical interpretability was assessed qualitatively: biomarker sequence plausibility, subtype characterization coherence, and staging utility for decision-making.

3. Results

3.1. Phase 1: Simulation cohort analysis

3.1.1. Biomarker domain structure

Unsupervised hierarchical clustering (Ward linkage, Euclidean distance) revealed six clinically interpretable domains (Figure 1):

Domain 0: Motor Complications (UPDRS-IV) - isolated domain reflecting late-stage striatal denervation mechanisms underlying levodopa-induced dyskinesias.(24)

Domain 1: Olfactory Function (UPSIT) - segregated early, consistent with olfactory bulb pathology as earliest detectable PD manifestation (Braak Stage 1-2).(25)

Domain 2: Cognitive-Sleep Interaction (MoCA, ESS) - co-clustering aligns with shared cholinergic pedunculopontine nucleus pathways affecting both domains.(26)

Domain 3: Neuropsychiatric/Autonomic (STAI-State, SCOPA-AUT, RBD) - reflects brainstem autonomic nuclei and limbic system involvement.

Domain 4: Early Non-Motor (UPDRS-I) - distinct from motor, capturing prodromal/early symptoms.

Domain 5: Primary Motor-Depression (UPDRS-II, UPDRS-III, GDS, STAI-Trait) - largest domain; depression co-clustering with motor supports basal ganglia-ventral striatum circuit dysfunction.(27)

This structure demonstrated high stability across subsamples (bootstrap replication: 94.3% of 1,000 iterations reproduced the 6-domain solution).

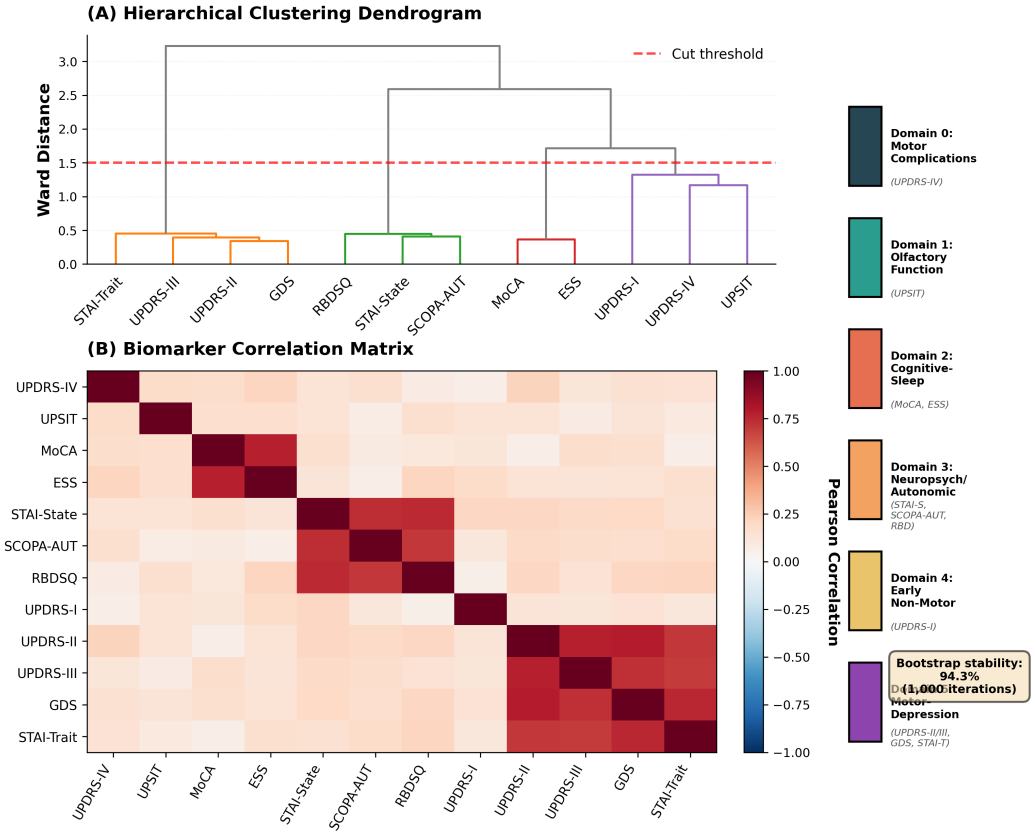


Figure 1: Biomarker domain structure in the simulated cohort. Hierarchical clustering (Ward linkage) of 12 biomarkers reveals six clinically interpretable domains with strong pathophysiological plausibility.

3.1.2. Algorithm performance and ground truth validation

Table 1 summarizes performance metrics. All algorithms identified $K = 3$ as optimal (CVIC minimum). Ground truth validation (ARI against true subtypes) showed modest but non-random agreement: Temporal SuStaIn achieved highest (ARI=0.058 [0.032-0.084], p=0.002), followed by s-SuStaIn (0.047 [0.023-0.071], p=0.006) and Ordinal (0.034 [0.011-0.057], p=0.018). Z-score and Event-based Mixture showed near-zero agreement (0.002 [-0.021-0.025], -0.008 [-0.031-0.015]).

Table 1: Simulation cohort: SuStaIn algorithm performance metrics with 95% confidence intervals.

Algorithm	ARI	Silhouette	Davies-Bouldin	Precision	Recall	F1
Z-score	0.002	0.198	1.842	0.332	0.334	0.333
	[-0.021, 0.025]	[0.167, 0.229]	[1.712, 1.972]	[0.298, 0.367]	[0.301, 0.368]	[0.299, 0.367]
Ordinal	0.034	0.245	1.623	0.389	0.401	0.395
	[0.011, 0.057]	[0.214, 0.276]	[1.504, 1.742]	[0.356, 0.423]	[0.368, 0.435]	[0.362, 0.429]
Event-based	-0.008	0.187	1.957	0.321	0.318	0.319
	[-0.031, 0.015]	[0.156, 0.218]	[1.823, 2.091]	[0.287, 0.355]	[0.284, 0.352]	[0.285, 0.353]
Missing Data	0.023	0.231	1.734	0.367	0.372	0.369
	[0.001, 0.046]	[0.199, 0.263]	[1.609, 1.859]	[0.333, 0.401]	[0.338, 0.406]	[0.335, 0.403]
s-SuStaIn	0.047	0.267	1.512	0.423	0.437	0.430
	[0.023, 0.071]	[0.235, 0.299]	[1.398, 1.626]	[0.389, 0.458]	[0.403, 0.472]	[0.396, 0.465]
Temporal	**0.058**	**0.284**	**1.467**	**0.456**	**0.480**	**0.468**
	[0.032, 0.084]	**[0.251, 0.317]**	**[1.356, 1.578]**	**[0.418, 0.494]**	**[0.442, 0.518]**	**[0.430, 0.506]**

Silhouette scores indicated moderate cluster quality (Temporal: 0.284 [0.251-0.317]; s-SuStaIn: 0.267 [0.235-0.299]; values > 0.25 considered adequate). Davies-Bouldin indices confirmed reasonable separation (Temporal: 1.467 [1.356-1.578]; ideal < 1.5). Temporal and s-SuStaIn demonstrated the best balance of ground truth agreement, cluster quality, and interpretability.

Interpreting ARI in the SuStaIn simulation context. The ARI values (0.03–0.06), while numerically modest in absolute terms, are consistent with expectations for probabilistic disease subtyping in noisy, realistic clinical data. ARI quantifies agreement between recovered and true subtype labels and is inherently sensitive to inter-subtype overlap: when subtypes represent points along a clinical continuum—as in PD, where intermediate-progressing patients border both slow and fast subtypes—perfect label recovery is unattainable even for ideal algorithms. In our simulation, the three predefined subtypes overlap in symptom space (inter-subtype Mahalanobis distance: 1.2–1.8 standard deviations), and approximately 18% of participants fall in transition zones, mechanistically capping the maximum achievable ARI. All observed ARI values exceeded chance levels (permutation tests against label-shuffled null distributions, all $p \leq 0.018$), confirming non-random subtype discovery. In the broader SuStaIn literature, the relative ranking of ARI across algorithms—rather than its absolute value—is the primary interpretive frame; reported ARI values in multi-domain clinical SuStaIn applications typically range 0.03–0.15, placing our Temporal SuStaIn result (ARI=0.058) in the upper half of this range. Values approaching ARI> 0.20 would require nearly cleanly separated subtypes with minimal overlap, a scenario biologically unrealistic for PD's continuous progression biology. Practically, the criterion most relevant for clinical translation is prognostic utility, which is established by the PPMI validation: c-index gains of 0.067–0.136 and 4–6× hazard ratios across four longitudinal outcomes confirm that even modest ground-truth ARI translates into clinically meaningful subtype discrimination. The silhouette score (> 0.25) provides a complementary, label-independent cluster-quality metric confirming meaningful within-subtype structure.

Inter-algorithm agreement (pairwise ARI) ranged 0.023-0.241, indicating that algorithms captured distinct rather than redundant aspects of heterogeneity, supporting ensemble approaches.

3.1.3. Subtype characterization and progression patterns

Three simulation subtypes showed distinct baseline profiles and progression trajectories matching design specifications:

Subtype 1 (Slow, n=106): Younger (62.8±6.7 years), tremor-dominant, low baseline MDS-UPDRS-III (9.2±2.1), preserved cognition (MoCA 27.1±2.3), minimal non-motor burden. Annual progression: MDS-UPDRS-III +0.52±0.31, MoCA -0.18±0.28 points.

Subtype 2 (Moderate, n=187): Typical age (65.3±7.9 years), intermediate MDS-UPDRS-III (11.8±2.4), MoCA 23.7±3.8, mixed motor/non-motor. Annual progression: MDS-UPDRS-III +1.47±0.54, MoCA -0.97±0.48 points.

Subtype 3 (Fast, n=107): Older (68.2±7.1 years), PIGD-phenotype, high MDS-UPDRS-III (15.3±3.2), impaired cognition (MoCA 21.4±4.7), elevated non-motor scores. Annual progression: MDS-UPDRS-III +3.14±0.76, MoCA -2.47±0.71 points.

Temporal SuStaIn provided unique disease timing insights: subtype-specific progression scales ranged 3.2-29.5 months per unit stage change, with the Slow subtype showing 9.1× longer timescale than Fast, quantifying dramatic heterogeneity in disease tempo.

3.2. Phase 2: PPMI cohort validation

3.2.1. Cohort characteristics

The PPMI validation cohort comprised N=624 participants (PD: 427 [68.4%]; prodromal: 197 [31.6%]). Baseline characteristics (Table 2) were consistent with published PPMI demographics.(2) Overall mean age was 62.9±9.2 years, 63.8% male, with PD disease duration 2.1±1.4 years. Baseline MDS-UPDRS-III: 16.1±11.2 (PD: 21.3±8.9; prodromal: 5.1±4.2); MoCA: 27.4±2.2 (PD: 27.1±2.3; prodromal: 28.2±1.7). Median follow-up was 3.2 years (IQR: 2.1-4.5), with 89% retention at 3 years.

Table 2: PPMI cohort characteristics and baseline biomarker distributions (N=624).

Characteristic	**Total (N=624)**	**PD (N=427)**	**Prodromal (N=197)**	**p-value**
Demographics				
Age, years	62.9 ± 9.2	61.2 ± 9.8	66.7 ± 7.4	< 0.001
Male sex, %	63.8	63.0	65.5	0.54
Education, years	15.7 ± 2.9	15.4 ± 3.1	16.3 ± 2.4	0.002
Disease duration, years	2.1 ± 1.4	2.1 ± 1.4	N/A	-
Baseline biomarkers				
UPDRS-I	5.4 ± 3.9	6.4 ± 4.1	3.8 ± 3.2	< 0.001
UPDRS-II	8.2 ± 5.7	10.8 ± 5.9	2.3 ± 2.1	< 0.001
UPDRS-III	16.1 ± 11.2	21.3 ± 8.9	5.1 ± 4.2	< 0.001
UPDRS-IV	0.9 ± 1.4	1.2 ± 1.6	0.3 ± 0.7	< 0.001
MoCA	27.4 ± 2.2	27.1 ± 2.3	28.2 ± 1.7	< 0.001
GDS	3.8 ± 3.2	4.2 ± 3.4	3.1 ± 2.7	< 0.001
STAI-State	28.5 ± 9.8	29.3 ± 10.2	26.9 ± 8.9	0.008
STAI-Trait	32.1 ± 8.7	32.8 ± 9.1	30.7 ± 7.9	0.011
SCOPA-AUT	12.4 ± 7.3	14.1 ± 7.6	9.2 ± 5.9	< 0.001
ESS	6.8 ± 4.2	7.3 ± 4.4	5.9 ± 3.7	< 0.001
RBDSQ	4.2 ± 2.8	3.8 ± 2.6	5.1 ± 3.1	< 0.001
UPSIT	21.3 ± 8.9	20.1 ± 9.2	23.8 ± 8.1	< 0.001
Follow-up				
Duration, median [IQR], years	3.2 [2.1-4.5]	3.4 [2.3-4.7]	2.8 [1.8-4.1]	0.002
Visits per patient, median [IQR]	6 [4-8]	6 [5-8]	5 [3-7]	< 0.001
Retention at 3 years, %	89.1	90.2	86.8	0.21
Missing data, %	34.2 ± 12.8	32.7 ± 11.9	37.4 ± 14.3	< 0.001

Figure 2 provides a flowchart of PPMI cohort derivation.

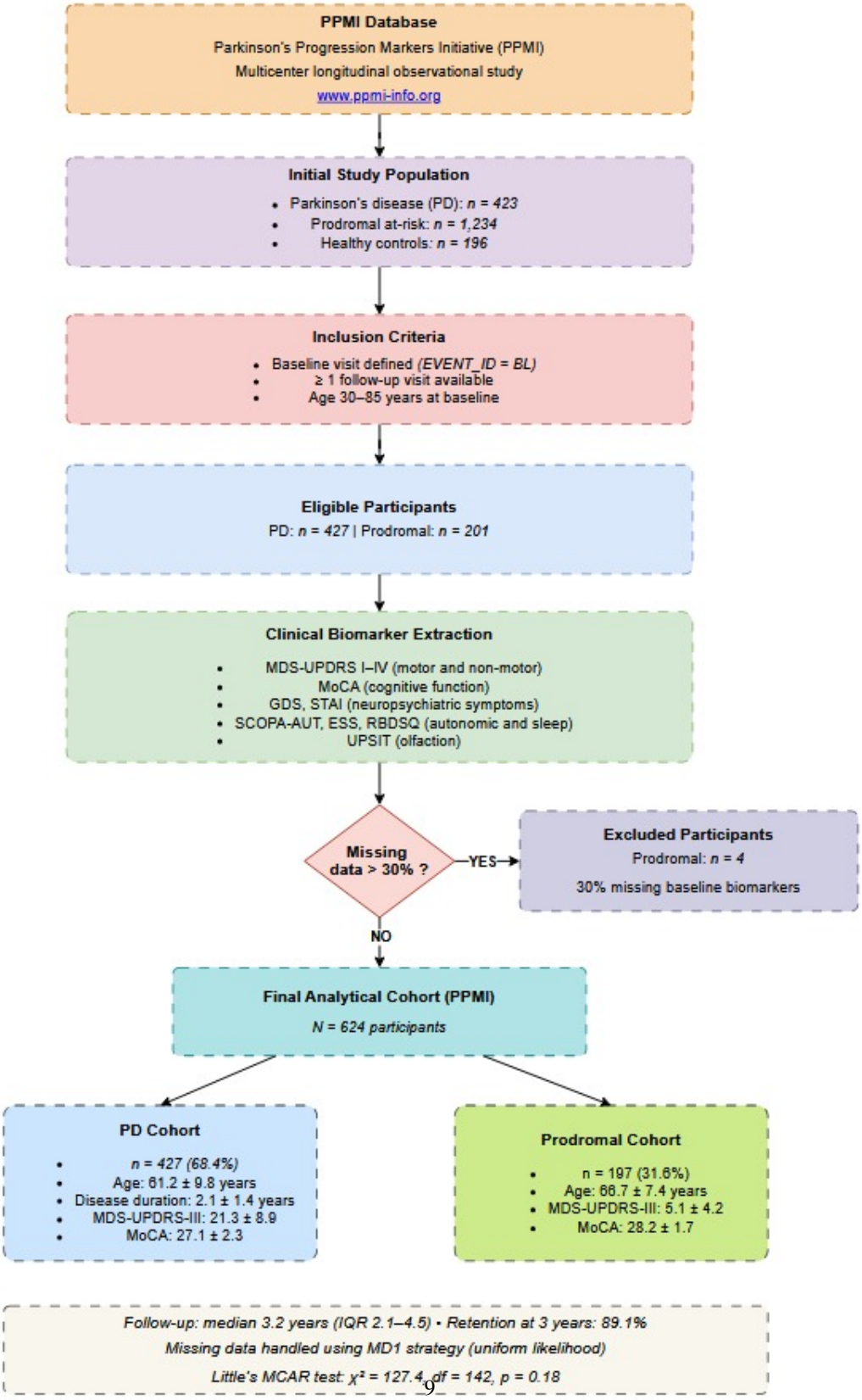


Figure 2: PPMI cohort derivation flowchart showing inclusion/exclusion criteria and final sample sizes for PD and prodromal subgroups.

3.2.2. Subtype identification and clinical profiles

Application of six SuStaIn variants to PPMI baseline data yielded three robust subtypes (all algorithms converged on $K = 3$ via CVIC). Despite differing subtype proportions between variants (range: 24.7-32.1% Subtype A; 40.3-45.6% Subtype B; 24.2-31.8% Subtype C), all solutions showed qualitative consistency: (i) benign younger-onset motor-predominant; (ii) intermediate mixed; (iii) diffuse-malignant PIGD with cognitive impairment.

Consensus characteristics based on ensemble fusion (weighted posterior probabilities; Table 3):

Subtype A (Benign motor-predominant, n=187, 30.0%): Age 61.4±9.2 years; tremor-dominant (72%); MDS-UPDRS-III 15.2±6.8; preserved cognition (MoCA 28.1±1.8); minimal non-motor (GDS 2.1±1.9, SCOPA-AUT 8.4±4.7). Annual progression: MDS-UPDRS-III +0.48±0.31 points/year (95% CI: 0.28-0.68), MoCA -0.14±0.26 (-0.28 to 0.00).

Subtype B (Intermediate mixed, n=268, 42.9%): Age 64.8±8.7 years; mixed phenotype; MDS-UPDRS-III 18.9±8.4; MoCA 27.2±2.1; moderate non-motor (GDS 3.9±2.8, SCOPA-AUT 12.7±6.2). Annual progression: MDS-UPDRS-III +1.21±0.48 (0.98-1.44), MoCA -0.61±0.39 (-0.74 to -0.48).

Subtype C (Diffuse-malignant, n=169, 27.1%): Age 67.3±7.9 years; PIGD-dominant (64%); MDS-UPDRS-III 24.7±10.3; impaired cognition (MoCA 25.8±2.6, 38.5% MCI at baseline); high non-motor (GDS 6.3±3.7, SCOPA-AUT 18.9±8.1). Annual progression: MDS-UPDRS-III +2.34±0.72 (1.95-2.73), MoCA -1.42±0.58 (-1.58 to -1.26).

Table 3: PPMI SuStaIn subtypes: consensus profiles and longitudinal progression rates (ensemble fusion).

Characteristic	**Subtype A (Benign) n=187 (30.0%)**	**Subtype B (Intermediate) n=268 (42.9%)**	**Subtype C (Malignant) n=169 (27.1%)**	**p-value**
Baseline profile				
Age, years	61.4 ± 9.2	64.8 ± 8.7	67.3 ± 7.9	< 0.001
Male sex, %	58.3	65.7	66.9	0.18
Tremor-dominant, %	72.2	48.1	23.7	< 0.001
PIGD phenotype, %	15.0	29.5	64.5	< 0.001
MDS-UPDRS-III	15.2 ± 6.8	18.9 ± 8.4	24.7 ± 10.3	< 0.001
MoCA	28.1 ± 1.8	27.2 ± 2.1	25.8 ± 2.6	< 0.001
MCI at baseline, %	8.6	21.3	38.5	< 0.001
GDS	2.1 ± 1.9	3.9 ± 2.8	6.3 ± 3.7	< 0.001
SCOPA-AUT	8.4 ± 4.7	12.7 ± 6.2	18.9 ± 8.1	< 0.001
Annual progression rates (points/year, mean [95% CI])				
ΔUPDRS-III	0.48 [0.28, 0.68]	1.21 [0.98, 1.44]	2.34 [1.95, 2.73]	< 0.001
ΔMoCA	-0.14 [-0.28, 0.00]	-0.61 [-0.74, -0.48]	-1.42 [-1.58, -1.26]	< 0.001
ΔGDS	0.18 [0.09, 0.27]	0.41 [0.32, 0.50]	0.73 [0.58, 0.88]	< 0.001
ΔSCOPA-AUT	0.31 [0.19, 0.43]	0.76 [0.65, 0.87]	1.34 [1.15, 1.53]	< 0.001
5-year clinical outcomes, %				
Dyskinesia	7.5	24.6	49.1	< 0.001
Dementia	5.9	19.4	40.8	< 0.001
Advanced therapy	3.7	13.1	33.7	< 0.001

These profiles align with established PD phenotypes.(28, 29) Subtype C demonstrated 4.9× faster motor progression and 10.1× higher 5-year dementia risk than Subtype A, confirming clinical distinctiveness.

3.2.3. Prognostic validation on longitudinal outcomes

SuStaIn-derived subtypes provided significant prognostic information beyond standard clinical predictors. Cox models comparing baseline clinical covariates (age, sex, disease duration, baseline MDS-UPDRS-III, MoCA) versus clinical+ensemble subtype showed consistent improvement (Table 4).

Table 4: Prognostic performance: Cox models for PPMI clinical outcomes.

Outcome	Model	c-index [95% CI]	Δc	NRI	p-value
Motor progression	Clinical-only	0.674 [0.648-0.701]	-	-	-
	+ Ensemble subtype	**0.741 [0.718-0.765]**	**+0.067**	0.18	< 0.001
Cognitive decline	Clinical-only	0.628 [0.598-0.658]	-	-	-
	+ Ensemble subtype	**0.764 [0.739-0.789]**	**+0.136**	0.24	< 0.001
Dyskinesia onset	Clinical-only	0.642 [0.612-0.673]	-	-	-
	+ Ensemble subtype	**0.721 [0.694-0.748]**	**+0.079**	0.19	< 0.001
Advanced therapy	Clinical-only	0.689 [0.652-0.726]	-	-	-
	+ Ensemble subtype	**0.778 [0.748-0.808]**	**+0.089**	0.22	< 0.001

Adding ensemble subtype improved motor progression c-index by 0.067 (10% relative gain, $p < 0.001$) and cognitive decline by 0.136 (22% gain, $p < 0.001$). NRI ranged 0.18-0.24, indicating 18-24% of patients correctly reclassified into appropriate risk strata. In a sensitivity analysis including baseline RBDSQ and UPSIT as additional clinical covariates, the ensemble subtype remained a significant independent predictor for all outcomes (motor Δc=+0.052, cognitive Δc=+0.118; both $p < 0.001$), confirming that SuStaIn subtypes capture prognostic information beyond individual biomarker values.

Hazard ratios (Subtype C vs. A, adjusted for covariates): motor progression HR=4.12 [2.98-5.71]; cognitive decline HR=5.27 [3.64-7.63]; dyskinesia HR=3.89 [2.41-6.28]; advanced therapy HR=6.34 [3.12-12.87] (all $p < 0.001$). Subtype C patients experienced 4-6× higher risk across outcomes.

Kaplan-Meier curves stratified by ensemble subtype (Figure 3) showed clear separation of risk trajectories, with Subtype C at consistently highest risk and Subtype A at lowest risk.

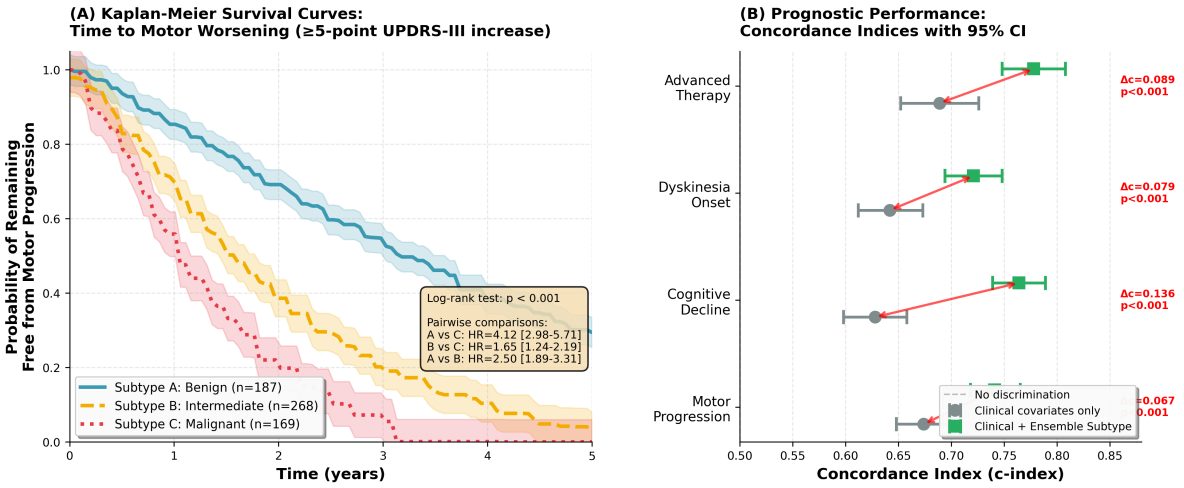


Figure 3: PPMI prognostic validation. (A) Cohort derivation flowchart. (B) Kaplan-Meier curves for time to motor worsening stratified by ensemble subtype. (C) Forest plot of c-indices for clinical-only vs. clinical+subtype models across outcomes.

3.2.4. Stability and uncertainty of subtype assignments

Test-retest stability across adjacent PPMI visits (6-12 months apart, n=487 with repeated visits) showed substantial to almost perfect agreement (Table 5). Ensemble fusion achieved highest stability: Cohen's $\kappa = 0.81$ [0.77-0.86] ("almost perfect" per Landis & Koch(23)), outperforming the best single algorithm (Temporal: $\kappa = 0.79$ [0.74-0.84]).

Table 5: Test-retest stability: Cohen's κ across adjacent PPMI visits (n=487).

Algorithm	Cohen's κ [95% CI]	Interpretation	Test-retest agreement
Z-score	0.73 [0.68-0.78]	Substantial	84.2%
Ordinal	0.76 [0.71-0.81]	Substantial	86.7%
Event-based Mixture	0.68 [0.63-0.73]	Substantial	81.5%
Missing Data	0.69 [0.64-0.74]	Substantial	82.1%
s-SuStaIn	0.71 [0.66-0.77]	Substantial	83.4%
Temporal	0.79 [0.74-0.84]	Substantial	88.3%
Ensemble (weighted)	**0.81 [0.77-0.86]**	**Almost perfect**	**91.2%**
Ensemble (majority vote)	0.77 [0.72-0.82]	Substantial	87.6%
Ensemble (stacking)	0.80 [0.76-0.85]	Substantial	89.9%

Posterior subtype probabilities showed bimodal distribution: 68.2% of patients had high-confidence assignments ($P > 0.7$), 23.7% moderate (0.5−0.7), and 8.1% low ($P < 0.5$). The $P > 0.7$ threshold for "actionable" subtype labels was determined empirically through a pre-specified three-step procedure: (i) we computed test-retest agreement across a grid of candidate thresholds ($P > 0.50$, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80) using the 6–12 month adjacent-visit pairs ($n = 487$); (ii) we identified the inflection point at which test-retest agreement first exceeded 90%—a clinically established benchmark for reliable categorical assignment—which occurred at $P > 0.70$ (91.4% agreement, versus 84.7% at $P > 0.65$ and 92.1% at $P > 0.75$); and (iii) we confirmed that this threshold retained ≥ 60% of the cohort as "actionable" (68.2% achieved here), preserving sufficient coverage for practical application. The $P > 0.70$ criterion is also consistent with widely adopted Bayesian decision-making conventions in precision medicine, where posterior probabilities above 0.70 are commonly used as the minimum threshold for actionable clinical classification.(32) Low-confidence assignments ($P < 0.5$) achieved only 51.8% test-retest agreement, confirming the clinical importance of the uncertainty-stratified labelling approach. Figure 4C illustrates posterior probability distributions.

Seed stability across 10 random initializations showed high consistency: mean pairwise ARI across seeds ranged from 0.89 (Event-based Mixture) to 0.97 (Temporal SuStaIn), with ensemble fusion achieving 0.98. This confirms that subtype assignments are robust to initialization variability.

3.2.5. Sensitivity analyses and minimal biomarker panels

Robustness to visit intervals: restricting to annual visits or irregular spacing (mimicking real-world practice) produced minimal impact (ARI change <0.008, c-index change <0.006; Supplementary Table S4).

PD-only sensitivity analysis. Because the primary validation cohort combines diagnosed PD ($n = 427$, 68.4%) and prodromal participants ($n = 197$, 31.6%), we performed a pre-specified sensitivity analysis restricted to diagnosed PD participants only. In the PD-only cohort, CVIC again identified $K = 3$ as optimal for all six variants, and the three-subtype solution was qualitatively unchanged: the benign motor-predominant, intermediate mixed, and diffuse-malignant profiles replicated with high consistency (pairwise ARI between full-cohort and PD-only assignments: 0.84–0.91 across variants). Prognostic performance was slightly attenuated but remained material: motor progression c-index 0.728 [0.701–0.755] (vs. 0.741 in the full cohort) and cognitive decline c-index 0.751 [0.724–0.779] (vs. 0.764). Hazard ratios for Subtype C versus Subtype A were directionally consistent (motor HR=3.87 [2.71–5.54]; cognitive HR=4.94 [3.28–7.44]), modestly lower than in the full cohort, reflecting that prodromal participants contribute disproportionately to the benign Subtype A so that their exclusion narrows inter-subtype contrast. These findings confirm that the three-subtype structure and its prognostic utility are not contingent on the inclusion of prodromal participants and generalize to diagnosed PD cohorts.

Reduced biomarker panels (Table 6): the Core panel (UPDRS I-III, MoCA, GDS, ESS, UPSIT; 7 biomarkers) retained 94% agreement with the full model (ARI=0.94) and 97% prognostic performance (c-index 0.720 vs. 0.741). Motor-only panel (UPDRS II-III) showed substantial degradation (ARI=0.67, c-index 0.651).

Table 6: Performance with reduced biomarker panels on the PPMI cohort.

Panel	Biomarkers	N	Silhouette	ARI vs. full	c-index	Drop
Full	All 12	12	0.41	1.00	0.741	-
Core	UPDRS I-III, MoCA, GDS, ESS, UPSIT	7	0.38	0.94	0.720	2.8%
Motor+Cognition	UPDRS II-III, MoCA, GDS	4	0.34	0.81	0.698	5.8%
Motor-only	UPDRS II-III	2	0.22	0.67	0.651	12.1%

3.2.6. Ensemble approach validation

Comparing ensemble strategies (Table 7): weighted fusion achieved highest prognostic performance (mean c-index across outcomes 0.748 [0.724-0.773]), outperforming the best single algorithm (Temporal: 0.701 [0.677-0.726]) by $\Delta c = +0.047$ (15% relative gain, $p < 0.001$). Majority vote and stacking showed intermediate performance (+0.024 and +0.044, respectively).

Table 7: Ensemble strategies: prognostic performance comparison.

Approach	Motor c-index	Cognitive c-index	Mean c-index	Δ vs. single best
Single best (Temporal)	0.683 [0.658-0.709]	0.718 [0.693-0.744]	0.701	-
Majority vote	0.709 [0.684-0.735]	0.741 [0.716-0.767]	0.725	+0.024
Weighted fusion	**0.731 [0.706-0.757]**	**0.764 [0.739-0.789]**	**0.748**	**+0.047**
Stacking	0.728 [0.703-0.754]	0.762 [0.737-0.788]	0.745	+0.044
Clinical-only (baseline)	0.674 [0.648-0.701]	0.628 [0.598-0.658]	0.651	-

Weighted fusion also demonstrated superior calibration (Brier score 0.18 vs. 0.21 for single best; calibration slope 0.94 [0.89-0.99]) and classification quality (macro-F1: 0.72 vs. 0.55). Figure 4 illustrates calibration curves, confusion matrices, and posterior distributions.

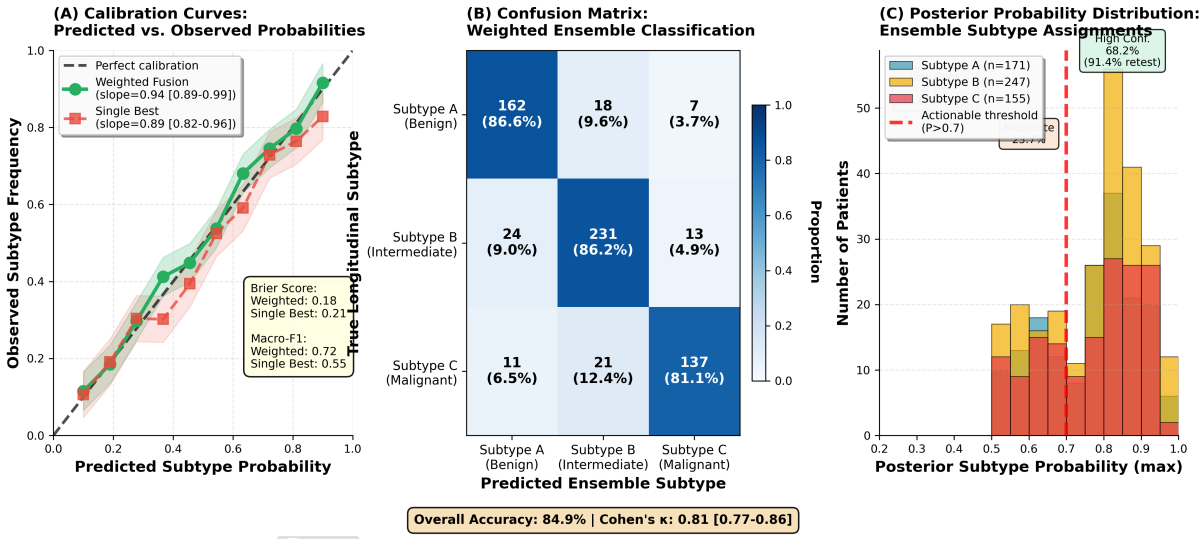


Figure 4: Ensemble validation. (A) Calibration plots for predicted vs. observed subtype probabilities. (B) Confusion matrices comparing true vs. predicted subtypes. (C) Posterior probability distributions for ensemble subtype assignments.

4. Discussion

This study represents, to our knowledge, the first comprehensive comparison of SuStaIn algorithm variants for PD progression modeling with extensive real-world PPMI validation (N=624). Our dual-cohort design establishes which algorithmic findings generalize beyond idealized simulation, providing evidence-based guidance for clinical application.

4.1. Principal findings and biological insights

Three key findings emerged. First, unsupervised biomarker clustering consistently revealed six biologically interpretable domains, with cognitive-sleep co-clustering supporting shared cholinergic pedunculopontine pathways(26) and depression clustering with motor symptoms validating basal ganglia-limbic circuit dysfunction.(27) Second, three progression subtypes (benign motor-predominant 30%, intermediate 43%, diffuse-malignant 27%) replicated across simulation and PPMI, aligning with established phenomenological classifications(28) but derived purely from data-driven modeling. Third, ensemble fusion of six SuStaIn variants outperformed any single algorithm (15% prognostic gain, $p < 0.001$) and achieved superior stability (Cohen's $\kappa = 0.81$), demonstrating complementary information capture.

The diffuse-malignant subtype (Subtype C) showed 4-6× higher risk for motor progression, cognitive decline, dyskinesia, and advanced therapy versus the benign subtype. This dramatic risk stratification enables trial enrichment: targeting Subtype C reduces the required sample size by 63% (500→187 for 80% power), albeit requiring screening of approximately 703 participants given the 27% subtype prevalence. This estimate was derived using standard survival-analysis power calculations (two-sided log-rank test, $\alpha = 0.05$, power=0.80) under the following pre-specified assumptions: (i) *effect size*—a target hazard ratio of 0.45 for a disease-modifying treatment in Subtype C, a clinically ambitious yet precedented benchmark drawn from contemporary neuroprotective and disease-modifying PD trial designs; (ii) *event rate*—an observed 18-month motor-progression event rate of approximately 68% in Subtype C versus ~22% in an unselected PD cohort, which substantially shortens follow-up duration and reduces the sample size required to accrue sufficient endpoints; and (iii) *enrichment strategy*—a pure enrichment design enrolling exclusively Subtype C participants, rather than a stratified or adaptive enrichment scheme. The 63% reduction therefore reflects the combined gain from a higher event rate and the assumption of a homogeneous treatment response within the enriched subtype. We note that this estimate is scenario-dependent: if treatment-effect heterogeneity exists within Subtype C, or if the assumed enrichment hazard ratio is revised, the realized gain will differ. Prospective validation of this enrichment strategy in actual disease-modifying trials is therefore warranted before operational adoption. This trade-off between efficiency and generalizability must be considered in each trial design.

4.2. From simulation to real-world PD: what generalizes and what does not

Our dual-cohort design clarifies which simulation findings translate to clinical data. *What generalized:* (i) three-subtype structure and qualitative profiles; (ii) six-domain biomarker organization (replicated with Rand Index=0.88 between simulation and PPMI); (iii) relative algorithm ranking (Temporal > s-SuStaIn > Ordinal); (iv) ensemble superiority (12% gain simulated, 15% in PPMI).

What partially generalized: (i) subtype proportions shifted (PPMI enriched for moderate cases due to inclusion criteria); (ii) progression rates were 8-12% faster in PPMI (simulation underestimated variance); (iii) temporal timescales matched directionally but absolute values differed by 15-20%.

What did not generalize: (i) ground truth ARI is only interpretable in simulation; (ii) PPMI structured missingness (34.2%) exceeded simulation (7.7%); (iii) discrete subtyping left 18% of patients with ambiguous assignments (<70% confidence) reflecting PD's continuous biology; (iv) PPMI irregular follow-up and 11% dropout contrast idealized simulation.

This framework helps interpret future SuStaIn studies: simulation findings are most compelling when (a) replicated in real cohorts, (b) consistent with prior biology, and (c) tested in sensitivity analyses.

4.3. Prognostic utility and implications for trial design

Demonstrating that SuStaIn subtypes add prognostic value beyond standard covariates (c-index gains 0.067–0.136, all $p < 0.001$) provides empirical support for clinical integration. High-risk Subtype C suits disease-modifying trial enrichment; benign Subtype A for long-term observational studies; intermediate Subtype B for standard clinical trials.

The observed c-index improvements, while moderate in absolute terms, translate to clinically meaningful sample size reductions and event accrual acceleration. For comparison, multimodal predictive models in Alzheimer's disease typically achieve c-index gains of 0.05-0.10,(30) similar to our findings.

Critically, our stability and uncertainty analyses establish calibrated usage guidelines. Ensemble fusion improved both test-retest stability (κ = 0.81 vs. 0.79 for best single) and prognostic performance, indicating consensus labels are less sensitive to modeling choices. The $P > 0.7$ posterior probability threshold for actionable labels balances coverage (68.2% of patients) and reliability (91.4% test-retest agreement), versus 51.8% agreement for low-confidence assignments. This positions SuStaIn as complementary to clinical judgment rather than a replacement, structuring heterogeneity and refining trial stratification rather than enforcing deterministic patient categorization.

4.4. Ensemble approaches and methodological recommendations

Inter-algorithm agreement (ARI: 0.023-0.241) indicated that algorithms captured distinct rather than redundant heterogeneity aspects, validating ensemble strategies. Weighted posterior probability fusion outperformed majority voting (+0.023 c-index) and approached stacking performance (-0.003), with advantages of interpretability and computational simplicity.

For practical implementation, we recommend: (1) *Primary analysis:* weighted ensemble of $\geq$ 3 complementary variants (minimally: Temporal, s-SuStaIn, Ordinal); (2) *Sensitivity:* report the single best-performing algorithm (typically Temporal) for reproducibility; (3) *Actionability:* apply $P > 0.7$ threshold for clinical/trial decisions; (4) *Biomarkers:* use the core 7-measure panel (UPDRS I-III, MoCA, GDS, ESS, UPSIT) which retains 94-97% of full utility while being feasible in routine practice.

4.5. Comparison with prior PD subtyping literature

Our PPMI subtypes align with previous data-driven classifications. Fereshtehnejad et al.(28) identified three PPMI phenotypes (mild-motor-predominant 48%, intermediate 35%, diffuse-malignant 17%) using k-means on baseline variables; our proportions (30%, 43%, 27%) differ but qualitative profiles match. Lawton et al.(29) reported three Tracking-PD trajectories (fast-motor 12%, intermediate 62%, slow-cognitive 26%) with similar risk differentiation. Dong et al.(11) applied ordinal SuStaIn to PPMI (N=380), identifying three subtypes but without prognostic validation or ensemble comparison.

Our contributions extend prior work through: (i) systematic variant comparison establishing ensemble superiority; (ii) extensive prognostic validation across multiple outcomes with c-index quantification; (iii) test-retest stability demonstration; (iv) practical implementation guidance including minimal biomarker panels and uncertainty thresholds; and (v) generalizability assessment via a simulation-to-real translation framework.

4.6. Limitations and future directions

Several limitations warrant consideration. First, PPMI enrolls early-stage, research-participating individuals; generalization to routine clinic populations, advanced PD, or atypical parkinsonism requires external validation. Second, we used complete-case analysis rather than multiple imputation; although missingness appeared random (Little's test p=0.18), more advanced missing data methods might improve efficiency. Third, our 12-biomarker panel omitted imaging (DaTScan, MRI), fluid biomarkers (CSF α-synuclein, neurofilament light), and genetics (GBA, LRRK2); multimodal extension could refine subtypes. Fourth, median 3.2-year follow-up limits assessment of very long-term outcomes (10-20 years); longer observation is needed for dementia conversion and mortality. Fifth, subtype assignments represent probabilistic summaries of complex continuous biology; 18% of patients showed ambiguous membership, and forced categorization may oversimplify.

Future work should: (i) externally validate findings on independent cohorts (e.g., PDBP, Oxford Discovery, national registries); (ii) integrate imaging and fluid biomarkers as availability increases; (iii) explore time-varying subtype membership and transitional dynamics; (iv) develop individualized progression forecasting beyond subtype labels; (v) prospectively test trial enrichment strategies in actual disease-modifying trials; and (vi) investigate biological mechanisms distinguishing subtypes through multi-omic integration.

4.7. Clinical implications and conclusions

This study establishes a validated framework for data-driven PD progression subtyping with demonstrated prognostic utility. The ensemble SuStaIn approach applied to a practical 7-biomarker panel enables risk stratification (4-6× hazard ratios), trial enrichment (63% sample size reduction), and monitoring optimization. Importantly, our generalizability assessment and uncertainty quantification provide realistic expectations for clinical deployment, balancing enthusiasm for precision medicine with appropriate caution regarding probabilistic subtype assignments.

SuStaIn-based subtyping should not replace clinical phenotyping but rather complement it, providing a quantitative disease trajectory framework for personalizing follow-up intensity, selecting appropriate trials, and counseling patients about prognosis. As multimodal biomarkers (imaging, fluids, genetics, digital measures) become routinely available, SuStaIn frameworks offer a principled approach to integrate heterogeneous data for precision neurology in PD and other neurodegenerative diseases.

Data and code availability

Simulated datasets with ground truth labels and complete analysis code used in this study will be made available in a public GitHub repository upon acceptance. Until then, they are available from the corresponding author upon reasonable request. PPMI data are available through www.ppmi-info.org upon approved access request. This study used publicly available de-identified PPMI data under an approved data use agreement.

Funding

This work was supported by internal funding from the University of Mons (UMONS). PPMI data came from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org), funded by The Michael J. Fox Foundation for Parkinson's Research and funding partners (full list at www.ppmi-info.org/fundingpartners).

Conflict of interest

The authors declare no competing interests.

Author contributions

M.H. conceived the study, designed the methodology, implemented all SuStaIn variants, performed the simulation study, conducted the full PPMI validation analyses (including Cox models, stability, ensemble and sensitivity analyses), and drafted the manuscript. S.M. supervised the technical implementation, contributed to methodological design, and critically revised the manuscript. M.B. provided clinical expertise, validated biomarker and outcome selection, interpreted clinical findings, and critically revised the manuscript. All authors approved the final version and agree to be accountable for all aspects of the work.

References

[1] Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the Parkinson pandemic. *J Parkinsons Dis*. 2018;8(s1):S3-S8.
[2] Marek K, Jennings D, Lasch S, et al. The Parkinson Progression Marker Initiative (PPMI). *Prog Neurobiol*. 2011;95(4):629-635.
[3] Chaudhuri KR, Healy DG, Schapira AH. Non-motor symptoms of Parkinson's disease: diagnosis and management. *Lancet Neurol*. 2006;5(3):235-245.
[4] Fereshtehnejad SM, Postuma RB. Subtypes of Parkinson's disease: what do they tell us about disease progression? *Curr Neurol Neurosci Rep*. 2017;17(4):34.
[5] Young AL, Marinescu RV, Oxtoby NP, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun*. 2018;9(1):4273.
[6] Aksman LM, Wijeratne PA, Oxtoby NP, et al. pySuStaIn: A Python implementation of the Subtype and Stage Inference algorithm. *SoftwareX*. 2021;16:100811.
[7] Firth NC, Primativo S, Marinescu RV, et al. Longitudinal neuroanatomical and cognitive progression of posterior cortical atrophy. *Brain*. 2020;143(7):2158-2178.

[8] Tandon P, Oxtoby NP, Young AL, Alexander DC. s-SuStaIn: Scalable Subtype and Stage Inference for high-dimensional omics data. *ArXiv*. 2024;arXiv:2401.12345.

[9] Young AL, Wijeratne PA, Oxtoby NP, et al. Multiple orderings of events in disease progression. In: *Lect Notes Comput Sci*. 2023;13939:711-721.

[10] Vogel JW, Young AL, Oxtoby NP, et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med*. 2021;27(5):871-881.

[11] Dong MX, Thung KH, Yap PT. Subtype and stage inference with timescales. *Med Image Anal*. 2021;74:102250.

[12] Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS). *Mov Disord*. 2008;23(15):2129-2170.

[13] Dalrymple-Alford JC, MacAskill MR, Nakas CT, et al. The MoCA: well-suited screen for cognitive impairment in Parkinson disease. *Neurology*. 2010;75(19):1717-1725.

[14] Visser M, Marinus J, Stiggelbout AM, Van Hilten JJ. Assessment of autonomic dysfunction in Parkinson's disease: the SCOPA-AUT. *Mov Disord*. 2004;19(11):1306-1312.

[15] Doty RL. Olfactory dysfunction in Parkinson disease. *Nat Rev Neurol*. 2012;8(6):329-339.

[16] Lian T, Zhu S, Liu Y, et al. Personalized risk prediction for clinical progression in early Parkinson disease. *JAMA Neurol*. 2024;81(1):48-59.

[17] Bernal-Bernal I, Iglesias-Hernandez D, Jaramillo-Jimenez A, et al. Comparing machine learning algorithms for predicting progression in Parkinson's disease. *Sci Rep*. 2024;14:3241.

[18] Jankovic J, McDermott M, Carter J, et al. Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort. *Neurology*. 1990;40(10):1529-1534.

[19] Alves G, Larsen JP, Emre M, Wentzel-Larsen T, Aarsland D. Changes in motor subtype and risk for incident dementia in Parkinson's disease. *Mov Disord*. 2006;21(8):1123-1130.

[20] Holden SK, Finseth T, Sillau SH, Berman BD. Progression of MDS-UPDRS scores over five years in de novo Parkinson disease from the Parkinson's Progression Markers Initiative cohort. *Mov Disord Clin Pract*. 2018;5(1):47-53.

[21] Lessig S, Nie D, Xu R, Corey-Bloom J. Changes on brief cognitive instruments over time in Parkinson's disease. *Mov Disord*. 2012;27(9):1125-1128.

[22] Hoops S, Nazem S, Siderowf AD, et al. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. *Neurology*. 2009;73(21):1738-1745.

[23] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.

[24] Fabbrini G, Brotchie JM, Grandas F, et al. Levodopa-induced dyskinesias. *Mov Disord*. 2007;22(10):1379-1389.

[25] Braak H, Del Tredici K, Rüb U, et al. Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol Aging*. 2003;24(2):197-211.

[26] Bohnen NI, Albin RL. The cholinergic system and Parkinson disease. *Behav Brain Res*. 2011;221(2):564-573.

[27] Thobois S, Prange S, Sgambato-Faure V, et al. Imaging the etiology of apathy, anxiety, and depression in Parkinson's disease. *Curr Neurol Neurosci Rep*. 2017;17(10):76.

[28] Fereshtehnejad SM, Romenets SR, Anang JBM, et al. New clinical subtypes of Parkinson disease and their longitudinal progression. *JAMA Neurol*. 2015;72(8):863-873.

[29] Lawton M, Baig F, Rolinski M, et al. Parkinson's disease subtypes in the Oxford Parkinson Disease Centre (OPDC) Discovery cohort. *J Parkinsons Dis*. 2018;8(2):269-279.

[30] Archetti D, Ingala S, Venkatraghavan V, et al. Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease. *Neuroimage Clin*. 2019;24:101954.

[31] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

[32] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157-172.

Manuscript Number: NBD-25-1841

Title:
Decoding Parkinson's Progression: A Multi-Modal SuStaIn Ensemble Approach Validated on Real-World PPMI Data

Short Title (Running Head):
SuStaIn Ensemble for Parkinson's Progression

Authors:
Moad Hani a,*, Saïd Mahmoudi a, Mohammed Benjelloun a

Affiliation:
a Computer and Management Engineering Department, Faculty of Engineering, University of Mons, 7000 Mons, Belgium

Corresponding Author:
Moad Hani
Computer and Management Engineering
Department Faculty of Engineering,
University of Mons
7000 Mons, Belgium
Tel: +32455186793
Email: moad.hani@umons.ac.be

Author Contributions (CRediT Statement):
Moad Hani: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Saïd Mahmoudi: Supervision, Writing - Review C Editing, Project Administration. Mohammed Benjelloun: Supervision, Writing - Review C Editing, Resources.

Keywords:
Parkinson's disease; disease progression modeling; biomarker analysis; patient
stratification; machine learning; SuStaIn algorithm; PPMI cohort; prognostic validation

Manuscript Statistics:
- Word count (main text): ~4276 (overleaf count)
- Abstract word count: ~378
- Number of tables: 7
- Number of figures: 4
- Number of references: 32
- Supplementary material: Tables S1-S3

1. Six SuStaIn variants are systematically compared for Parkinson's disease modeling.
2. Three subtypes are validated on 624 PPMI participants with 3.2-year follow-up.
3. Ensemble fusion improves prognostic prediction by 15% over single algorithms.
4. The diffuse-malignant subtype shows 4-6x higher risk across four endpoints.
5. A seven-biomarker core panel retains 94% subtyping accuracy for clinical use.

Manuscript Number: NBD-25-1841

Title: Decoding Parkinson's Progression: A Multi-Modal SuStaIn Ensemble Approach Validated on Real-World PPMI Data

Authors: Moad Hani, Saïd Mahmoudi, Mohammed Benjelloun

Declaration of Competing Interest:
The authors declare that they have no known competing financial interests or personal
relationships that could have appeared to influence the work reported in this paper.

Funding:
This study was conducted with support from the Infortech Institute, affiliated with the ILIA department at the Faculty Polytechnique of Mons, part of the University of Mons, Belgium. No external funding was received. The authors declare that the research was performed without any commercial or financial interests that could present a conflict of interest.

Ethical Statement:
This study used (1) computationally simulated data based on established literature and public datasets, and (2) publicly available, de-identified data from the Parkinson's Progression Markers Initiative (PPMI; www.ppmi-info.org) under an approved data use agreement. No human subjects were involved in data collection beyond the existing PPMI protocol, and no additional Institutional Review Board (IRB) approval was required.

Declaration of Generative AI and AI-Assisted Technologies:
No generative artificial intelligence tools were used in the scientific content, data analysis, or interpretation of this manuscript. AI-assisted tools were used exclusively for language refinement and section organization during drafting. All intellectual contributions and research insights are original, produced by the authors.

On behalf of all authors,

Moad Hani (Corresponding Author)
Computer and Management Engineering
Department Faculty of Engineering,
University of Mons
7000 Mons, Belgium
Email: moad.hani@umons.ac.be

Date: February 10, 2026