Interpretable Machine Learning Evaluates Differential Therapy Effects in Parkinsonian Gait

Artur Chudzik[1] **Henryk Josiński**[2] **Andrzej Przybyszewski**[1*]
[1]Faculty of Computer Science, Polish-Japanese Academy of Information Technology, Warsaw, Poland
[2]Department of Computer Graphics, Vision and Digital Systems, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland
{artur.chudzik,przy}@pja.edu.pl henryk.josinski@polsl.pl

Abstract

Parkinson's disease therapy is usually adjusted based on extensive clinical examinations, but the effects of medication and deep brain stimulation (DBS) on gait dynamics remain difficult to quantify precisely. We use interpretable, use-inspired probabilistic machine learning to measure therapy-related gait changes in a repeated-measures motion-capture dataset comprising 376 walking trials from 19 patients with Parkinson's disease. Patients were recorded under four therapy conditions: medication off/stimulation off, medication off/stimulation on, medication on/stimulation off, and medication on/stimulation on. They performed natural and fast walking, and a study neurologist scored motor severity using UPDRS-III. From 1.0 s sliding windows of foot-marker motion, we compute short-window largest Lyapunov exponent (LLE)-derived local-divergence features that capture local changes in gait trajectory dynamics. We summarize these features as two candidate LLE-derived digital therapy-response biomarkers: left-right asymmetry and left-right coupling. Bayesian hierarchical regression then estimates medication, stimulation, and task effects while accounting for repeated measurements within the same patients and patient-specific baseline differences. The main result is that during natural walking, LLE asymmetry decreased under combined therapy with a large paired effect size ($d_z = -0.83$). Coupling features changed with therapy in a complementary way, especially for stimulation-related changes in bilateral coordination. Standard gait metrics, focused UPDRS-III correlations, parameter sensitivity, and leave-one-patient-out analyses provided convergent evidence and supported the robustness of the main result. These findings suggest that short-window LLE asymmetry tracks therapy-related reduction of lateralized gait dysregulation, while LLE coupling tracks treatment-dependent reorganization of left-right coordination. By modeling candidate LLE-derived biomarkers across controlled medication and DBS states, we show that the two therapies are expressed through complementary gait-dynamics signatures, supporting interpretable nonlinear gait biomarkers for objective therapy-response assessment.

1 Introduction

Parkinson's disease (PD) is an increasingly urgent clinical and public-health problem. As populations age, the number of people living with PD is expected to rise substantially, while objective therapy-response assessment remains constrained by episodic visits and coarse rating scales. This gap is particularly severe for gait measurements, which are commonly evaluated using UPDRS-III item

*Corresponding author: przy@pja.edu.pl.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

scores and basic mobility tests such as the Timed Up and Go (TUG). Gait impairment in PD is disabling, associated with falls and loss of independence, and often responds unpredictably to therapy. Although dopaminergic medication and deep brain stimulation (DBS) are established treatments for cardinal motor symptoms, their effects on gait can be variable, patient-specific, and difficult to evaluate objectively during routine care.

The measurement problem is not only clinical but also computational. Modern motion capture, wearable sensors, and video-based pose estimation can produce rich motor time series, but the field still lacks standardized, shareable, therapy-state-resolved datasets and agreed-upon digital gait biomarkers. Recent reviews emphasize that the absence of common gait protocols, harmonized analysis pipelines, and minimum sets of quantified biomarkers limits comparison across studies and slows clinical translation [5, 7]. This is especially problematic for DBS, where gait effects may depend on electrode location, stimulation parameters, medication state, and patient-specific symptom lateralization [6, 8].

In this work, we address this gap by introducing a new motion-capture-derived dataset of Parkinsonian gait collected under a within-patient medication-by-stimulation design. Nineteen patients with PD were recorded across up to four therapy states: MED OFF/STIM OFF, MED OFF/STIM ON, MED ON/STIM OFF, and MED ON/STIM ON. Clinical motor severity was assessed using the Unified Parkinson's Disease Rating Scale part III (UPDRS-III) by a study neurologist. This four-condition design is uncommon and valuable because it provides a within-patient view of gait under medication and stimulation manipulations, rather than only comparing treated and untreated groups or patients and controls. It allows us to ask not only whether gait is abnormal, but how the same patient's gait dynamics change as medication and subthalamic nucleus deep brain stimulation (STN-DBS) are switched on and off.

We propose that interpretable machine learning should focus on therapy response rather than only diagnosis. For a clinician, the key question is often not simply whether a patient has abnormal gait, but whether therapy is improving gait, whether DBS is changing bilateral coordination, and whether residual asymmetry remains. We therefore extract short-window LLE-derived local-divergence features from foot-marker trajectories and summarize them as two clinically interpretable candidate biomarker families: left-right asymmetry and bilateral coupling.

Nonlinear dynamical measures offer a promising framework for quantifying such aspects of gait. Human walking is a continuous, cyclic, bilaterally coordinated dynamical process, and small alterations in local trajectory structure may reveal changes that are not captured by average gait parameters. Lyapunov-exponent-based measures have been widely used to quantify the divergence of nearby trajectories reconstructed from kinematic or inertial gait signals [1–3].

The largest Lyapunov exponent is commonly used to quantify the exponential divergence of nearby trajectories in reconstructed state space. Positive LLE values are conventionally interpreted as a signature of local chaotic divergence and sensitivity to initial conditions, whereas smaller or negative values indicate weaker divergence or locally convergent trajectory evolution. In this study, LLE is estimated in short sliding windows to capture time-local changes in gait dynamics rather than averaging stability over an entire recording.

LLE-based gait analysis has substantial precedent, but its implementation is not standardized. Prior studies differ in signal choice, filtering, normalization, time-series length, embedding parameters, divergence-curve fitting regions, and temporal-neighbour exclusion rules. Work on phase-dependent stability and short gait bouts also shows why one whole-trial value can miss local structure [17, 18], while long-term divergence can be sensitive to the noise structure of stride intervals [19]. A recent systematic scoping review found that Lyapunov exponent measures were among the most frequently used nonlinear dynamic metrics in walking studies, with Rosenstein's algorithm being the most common estimation approach, but also emphasized substantial heterogeneity in analysis pipelines and incomplete reporting of attractor-reconstruction parameters such as embedding dimension, time delay, and Theiler windows [4]. This motivates an explicit, implementation-aware interpretation of LLE-derived gait features rather than treating absolute LLE values as directly interchangeable across studies.

Because PD commonly begins asymmetrically, the clinically important signal may not be the absolute level of local divergence alone, but the imbalance of local chaotic instability between the two sides of the body. Bilateral asymmetry of short-window LLE features may therefore provide a marker of

lateralized motor dysregulation. Similarly, the temporal coupling between left- and right-sided LLE profiles may capture changes in bilateral coordination and the alternating structure of gait. These two feature classes may reflect complementary aspects of Parkinsonian motor impairment: asymmetry as a marker of lateralized therapy response, and coupling as a marker of left-right coordination.

The distinction between dopaminergic medication and DBS is also clinically important. Medication may reduce global or bilateral motor asymmetry, whereas the effect of DBS may depend on the exact electrode location, stimulation parameters, and the lateralization of symptoms. Thus, medication and stimulation may influence different components of gait dynamics. Medication-related effects may be most evident in LLE-derived asymmetry, whereas stimulation-related effects may be more strongly reflected in left-right LLE coupling. A factorial MED-by-STIM design provides a unique opportunity to dissociate these effects and to test whether LLE-derived features reflect treatment-specific mechanisms rather than merely distinguishing an untreated from a treated state.

Small clinical datasets create an additional modeling challenge: even when many gait windows or repeated recordings are available, the number of patients remains small and measurements from the same patient are not independent. Hierarchical Bayesian models are well suited to this setting because they allow each patient to have an individual baseline while estimating shared group-level therapy effects and uncertainty [9, 10]. This makes the task a clinically realistic small-data machine learning problem: the target is not diagnostic classification, but interpretable estimation of therapy response.

Therapy-state-resolved gait recordings in PD are difficult to collect because they require coordinated access to motion-capture infrastructure, DBS patients, controlled medication and stimulation states, and clinical annotation. Our use-inspired contribution is the design and validation of an interpretable pipeline aligned with this clinical factorial experiment: a therapy-state-resolved dataset, bilateral local-divergence profiles, clinically readable endpoints, and probabilistic repeated-measures inference. We provide de-identified participant-session-task features and analysis code to support reproducibility and comparison of future therapy-response biomarkers.

Here, we propose that short-window LLE gait features can provide clinically interpretable signatures of therapy-dependent motor dysregulation in PD. We hypothesize that untreated gait shows greater bilateral asymmetry of local trajectory divergence than treated gait within the same patients, reflecting lateralized motor impairment. We further hypothesize that combined therapy reduces this asymmetry, while medication- and stimulation-related effects may be expressed through partially distinct signatures. Specifically, medication-related effects are expected to be evident in LLE-derived asymmetry, whereas stimulation-related effects may be more strongly reflected in left-right LLE coupling. More broadly, we argue that asymmetry and coupling provide complementary digital biomarkers: asymmetry as a clinically interpretable marker of therapy-related improvement, and coupling as a marker of treatment-dependent modulation of bilateral gait coordination. Figure 1 summarizes the proposed pipeline: foot-marker motion capture is transformed into short-window LLE features, which are then modeled using Bayesian hierarchical machine learning to separate medication and stimulation effects.

2 Methods

Cohort and Experimental Design. We analyzed a motion-capture cohort of 19 participants with Parkinson's disease. Data collection was approved by the relevant institutional ethics committee, and all participants provided informed consent. Each participant was recorded under up to four therapy conditions forming a within-subject 2×2 design: medication off and stimulation off (MED OFF / STIM OFF), medication off and stimulation on (MED OFF / STIM ON), medication on and stimulation off (MED ON / STIM OFF), and medication on and stimulation on (MED ON / STIM ON). Recordings included two walking tasks: natural-speed walking and fast walking. Clinical motor severity was assessed on site by a study neurologist using UPDRS-III item scores and total scores matched to the corresponding therapy condition.

Experimental Setup and Data Acquisition. Data were acquired in the Human Motion Laboratory at PJAIT Bytom using a multimodal gait laboratory setup with Vicon motion capture and synchronized laboratory subsystems. Motion capture was recorded with 10 Vicon MX-T40 cameras at 100 fps. Synchronized video was recorded using 4 Basler Pilot piA 1920-32gc cameras at 25 fps. Participants

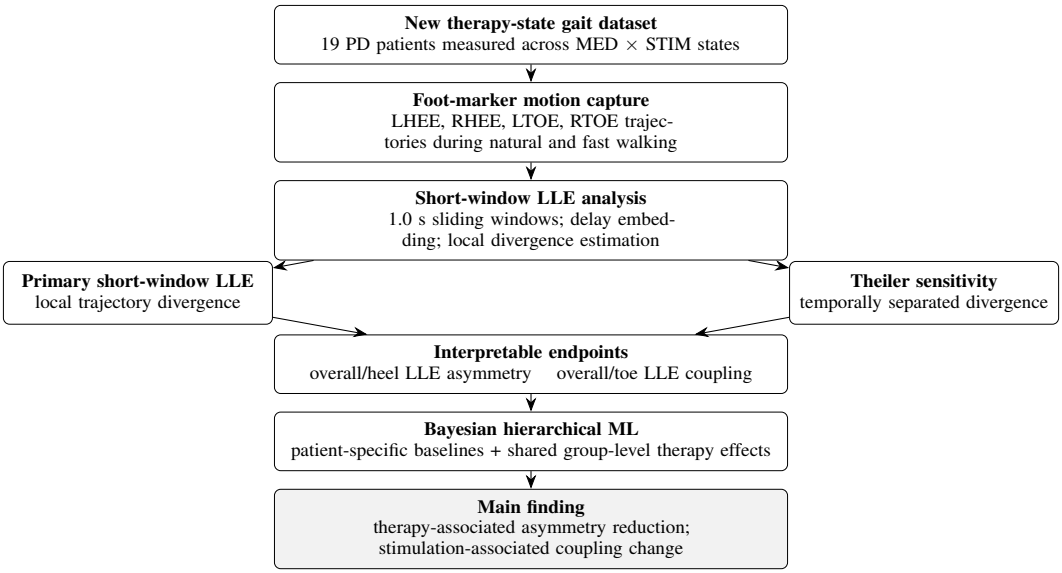

Figure 1: Conceptual overview of the proposed therapy-response biomarker pipeline. A new therapy-state-resolved gait dataset is transformed into short-window LLE features and summarized as asymmetry and coupling endpoints. Bayesian hierarchical machine learning then estimates medication and stimulation effects while accounting for repeated observations within patients.

were instrumented using the standard Vicon full-body Plug-in Gait marker model with 39 markers. Ground reaction forces were recorded using two Kistler force plates embedded in a 6 m walkway at 1000 Hz. Electromyography was recorded using a Noraxon MyoSystem 1400A system at 1000 Hz with 16 electrodes, 8 on each side of the body, placed over lower-limb and hip muscles. The present analysis uses the foot-marker motion-capture trajectories.

Motion-Capture Signals. Motion-capture recordings were stored as C3D files and processed from the lower-limb foot markers LHEE, RHEE, LTOE, and RTOE. For each marker, we extracted the vertical coordinate $z_i(t)$ at the native point sampling rate. Marker samples with invalid residuals, non-finite coordinates, or all-zero coordinates were treated as missing and linearly interpolated when enough valid samples were available. Each marker trajectory was smoothed with a Savitzky–Golay filter using window length 11 frames and polynomial order 3 before local divergence estimation.

Short-Window LLE-Derived Local Divergence Features. For each marker i, we reconstructed a delay-coordinate state space from the smoothed vertical trajectory using Equation 1,

$$\mathbf{x}_i(t) = [z_i(t), z_i(t+\tau), \ldots, z_i(t+(d-1)\tau)]. \tag{1}$$

We used fixed embedding parameters $d = 3$ and $\tau = 25$ frames for all four markers. LLE-derived features were estimated in 1.0 s sliding windows advanced every 0.1 s. The parameters were selected once at the dataset level rather than separately by marker, patient, task, or therapy condition. This choice avoids condition-specific parameter tuning and therefore prevents therapy labels from influencing the feature definition. The embedding dimension $d = 3$ provides the minimal delay reconstruction capable of representing nonlinear state-space structure beyond a planar trajectory, while $\tau = 25$ frames corresponds to 0.25 s at the 100 Hz motion-capture sampling rate. Thus, each 1.0 s window contains 100 marker samples, each embedded state vector spans 0.5 s of motion, and 50 embedded state vectors remain for local neighbour search and divergence estimation. This preserves a clinically local time scale while retaining enough state-space samples to estimate a window-level slope. Within each window, nearest neighbours were identified in reconstructed state space. For each embedded point $\mathbf{x}(t)$, we found the nearest admissible neighbour $\mathbf{x}(t')$ and computed the one-step logarithmic divergence in Equation 2,

$$\delta(t) = \log \|\mathbf{x}(t+1) - \mathbf{x}(t'+1)\|. \tag{2}$$

The short-window LLE-derived feature was then estimated as the slope of a linear regression fitted to the sequence of one-step local-divergence values within the window. This is inspired by Rosenstein-style nearest-neighbour divergence, but it is not a conventional per-window Rosenstein LLE estimate. Rosenstein's estimator fits the slope of a mean multi-horizon divergence curve. We therefore use the term *short-window LLE-derived local-divergence feature* throughout.

The largest Lyapunov exponent quantifies the rate of divergence of nearby trajectories in reconstructed state space. Here, the short-window derived slopes quantify how local divergence changes over time, between body sides, and across therapy states. Their absolute values should not be interpreted as conventional whole-trial LLE estimates.

We computed two nearest-neighbour variants. The primary short-window LLE variant excluded only self-matches and retained temporally adjacent neighbours. This follows the intended use of the feature as a local marker of trajectory continuity and divergence within brief gait segments. The Theiler-excluded variant additionally excluded neighbours within 0.1 s and was used as a sensitivity analysis based on temporally separated neighbours.

Derived Endpoints. Window-level marker LLE values were summarized into side-specific, asymmetry, and coupling endpoints. For each window, left and right mean LLE were defined as the average of heel and toe markers on the corresponding side. Heel asymmetry was defined as LHEE minus RHEE, toe asymmetry as LTOE minus RTOE, and overall asymmetry as the left mean minus the right mean. These asymmetry values are therefore differences between window-level LLE-derived slopes. For participant-session-task summaries, signed asymmetry time series were converted to absolute values before averaging, so that the endpoint quantified the magnitude of left-right imbalance rather than its direction.

Coupling endpoints were defined as within-recording Pearson correlations between left- and right-sided LLE profiles: heel coupling between LHEE and RHEE, toe coupling between LTOE and RTOE, and overall coupling between the averaged left- and right-sided LLE time series. Coupling is therefore a unitless correlation between left- and right-sided LLE profiles. Its sign describes whether the two sides change together or in opposing phases, not whether a Lyapunov exponent is positive or negative. Negative coupling indicates anti-correlated left-right local divergence profiles, consistent with the alternating phase structure of gait. Within-trial summaries included means, medians, standard deviations, interquartile ranges, percentiles, extrema, positive fractions, sign-switching rates, and lag-1 autocorrelations. Because some participant-session-task combinations contained multiple walking trials, trial summaries were averaged within each participant, therapy condition, and task before statistical modeling. Thus, the 376 processed walking trials yielded 144 participant-session-task rows, which were used as the inferential analysis dataset.

Classical Statistical Analysis. All analyses were performed separately by walking task and LLE variant. We first summarized endpoint distributions within each therapy condition. We then tested paired therapy contrasts using within-subject Wilcoxon signed-rank tests. The prespecified clinical contrast compared MED OFF / STIM OFF with MED ON / STIM ON for natural-walking overall LLE asymmetry in the primary variant. As a secondary clinical contrast, we compared MED OFF / STIM OFF with MED ON / STIM OFF to assess medication-sensitive asymmetry reduction.

To separate medication and stimulation effects, we computed within-subject factorial contrasts across participants with complete observations in all four therapy conditions. The medication main effect was defined as the average of MED ON conditions minus the average of MED OFF conditions. The stimulation main effect was defined as the average of STIM ON conditions minus the average of STIM OFF conditions. The interaction was defined as $(S04 - S03) - (S02 - S01)$, where $S01$ through $S04$ denote the four therapy sessions in the order listed above. For all inferential rows, we report paired-sample effect size d_z and Benjamini–Hochberg false-discovery-rate correction within each analysis section, task, and LLE variant.

Clinical, Standard-Metric, and Robustness Validation. We extracted standard gait metrics from detected events and compared their OFF/OFF to ON/ON changes with changes in the LLE-derived endpoints using Spearman correlations. Convergent validity was tested against matched changes in UPDRS-III total, gait-item, and postural-stability scores, with FDR correction across six planned comparisons and secondary speed-adjusted checks. Robustness analyses used seven one-factor feature-extraction perturbations, phase-randomized surrogates, leave-one-patient-out analyses, and

Bayesian posterior predictive checks. Appendix A and the supplementary material provide the configurations, complete outputs, and null results.

Bayesian Hierarchical Machine Learning. We used Bayesian hierarchical linear regression as an interpretable probabilistic machine learning model for therapy-dependent LLE features. This choice was motivated by three properties of the dataset. First, observations are repeated within patients and therefore not independent. Second, the number of patients is small, so high-capacity predictive models would be prone to overfitting. Third, the clinically relevant output is not only a point estimate but also uncertainty about medication, stimulation, and task effects. The treatment indicators are binary. Thus, the MED, STIM, and MED-by-STIM terms represent all four cells of the 2×2 design and do not impose a linear biological dose-response relation.

For each endpoint, we modeled standardized biomarker values using the likelihood in Equation 3,

$$y_n \sim \mathcal{N}(\mu_n, \sigma). \tag{3}$$

Here, y_n is the standardized value of one LLE-derived endpoint for observation n, such as overall LLE asymmetry or toe LLE coupling. Thus, the LLE-derived biomarker is the modeled outcome, not an input predictor. The linear predictor is defined in Equation 4,

$$\mu_n = \alpha + u_{p[n]} + \beta_{\mathrm{MED}}\mathrm{MED}_n + \beta_{\mathrm{STIM}}\mathrm{STIM}_n + \beta_{\mathrm{INT}}(\mathrm{MED}_n \times \mathrm{STIM}_n) + \beta_{\mathrm{FAST}}\mathrm{FAST}_n, \tag{4}$$

where μ_n is the model-predicted mean biomarker value for observation n, α is the group-level intercept, and $u_{p[n]}$ is a patient-specific random intercept for the patient associated with observation n. The binary indicators MED_n and STIM_n equal 0 when medication or stimulation is off and 1 when medication or stimulation is on, respectively. The product $\mathrm{MED}_n \times \mathrm{STIM}_n$ represents the medication-by-stimulation interaction. The binary task indicator FAST_n equals 0 for natural-speed walking and 1 for fast walking, adjusting therapy-effect estimates for systematic differences between walking tasks. The coefficients β_{MED}, β_{STIM}, β_{INT}, and β_{FAST} estimate average changes in the LLE-derived outcome associated with each design component, on the standardized outcome scale.

In plain terms, each Bayesian model asks whether an LLE-derived biomarker changes when medication is on, when DBS stimulation is on, when both are on together, or when the task is fast rather than natural walking, while allowing each patient to have their own baseline biomarker level. Patient intercepts were drawn from the group-level prior in Equation 5,

$$u_p \sim \mathcal{N}(0, \sigma_{\mathrm{patient}}). \tag{5}$$

This hierarchy allows each patient to have an individual baseline while borrowing information from the full cohort. It estimates group-level therapy effects with baseline heterogeneity, not patient-specific treatment slopes. Random therapy slopes would be weakly identified from at most eight task-by-state rows per patient, with missing cells for some patients. Weakly regularizing priors are shown in Equation 6,

$$\alpha \sim \mathcal{N}(0, 1), \quad \beta_k \sim \mathcal{N}(0, 0.5), \quad \sigma \sim \mathrm{HalfNormal}(1), \quad \sigma_{\mathrm{patient}} \sim \mathrm{HalfNormal}(0.5). \tag{6}$$

Models were fitted separately for the primary LLE variant and the Theiler-excluded LLE variant. The main endpoints were selected a priori: overall LLE asymmetry, heel LLE asymmetry, overall LLE coupling, and toe LLE coupling. We additionally fitted task-modulated models with MED-by-task and STIM-by-task terms to visualize model-predicted biomarker trajectories separately for natural and fast walking.

Posterior summaries are reported as posterior means, 95% credible intervals, and posterior direction probabilities such as $P(\beta_{\mathrm{MED}} < 0)$ for asymmetry reduction or $P(\beta_{\mathrm{STIM}} > 0)$ for coupling increase. The key Bayesian quantities are summarized in Table 4, and model-predicted trajectories are shown in Figure 2. All Bayesian models were run on a CPU workstation with a 13th Gen Intel Core i9-13900KF, 32 logical CPU cores, and 31.8 GB RAM; no GPU was used. The 12 reported Bayesian models, each using 4 chains with 1000 tuning iterations and 1000 posterior draws, completed in approximately 1.5 minutes.

3 Results

Cohort and Data Availability. The analysis included 19 participants with PD (15 male, 4 female; age 60.1 ± 9.6 years). Motion-capture availability was high but not perfectly balanced across sessions.

For natural walking, participant-session counts were 18, 18, 19, and 18 for MED OFF / STIM OFF, MED OFF / STIM ON, MED ON / STIM OFF, and MED ON / STIM ON, respectively. For fast walking, the corresponding counts were 17, 18, 18, and 18. UPDRS-III total scores decreased across therapy conditions: 52.7 ± 16.4 in MED OFF / STIM OFF, 29.6 ± 11.8 in MED OFF / STIM ON, 24.1 ± 14.1 in MED ON / STIM OFF, and 17.9 ± 9.8 in MED ON / STIM ON. Throughout the Results, p denotes the nominal p-value, q denotes the FDR-adjusted p-value, and d_z denotes the paired-sample standardized effect size; these definitions apply to all tables below.

Combined Therapy Reduced Normal-Walking Overall LLE Asymmetry. In the primary LLE variant, natural-walking overall LLE asymmetry decreased monotonically across therapy conditions, as shown in Table 1. The prespecified paired comparison between MED OFF / STIM OFF and MED ON / STIM ON showed a significant reduction in overall LLE asymmetry after false-discovery-rate (FDR) correction ($n = 17$, mean difference $= -0.0103$, Wilcoxon $p = 0.0038$, FDR $q = 0.0427$, $d_z = -0.83$). The secondary medication-sensitive contrast, MED OFF / STIM OFF versus MED ON / STIM OFF, was also significant after correction ($n = 18$, mean difference $= -0.0071$, $p = 0.0066$, $q = 0.0474$, $d_z = -0.75$). Thus, the main clinical signal was a reduction of lateralized local gait divergence during natural walking.

Table 1: Natural-walking endpoint means by therapy condition in the primary LLE variant. Values are means across available participants. Asymmetry endpoints are absolute left-right differences in short-window LLE-derived slope units. Coupling endpoints are unitless Pearson correlations (r).

Endpoint	OFF/OFF	OFF/ON	ON/OFF	ON/ON
Overall LLE asymmetry	0.0515	0.0464	0.0447	0.0406
Heel LLE asymmetry	0.0551	0.0520	0.0492	0.0458
Toe LLE asymmetry	0.0574	0.0536	0.0544	0.0505
Overall LLE coupling	-0.4092	-0.3625	-0.3932	-0.3199
Heel LLE coupling	-0.2654	-0.2246	-0.2643	-0.2169
Toe LLE coupling	-0.3448	-0.2454	-0.2606	-0.2245

Natural Walking Provided the Clearest Therapy-Response Signal. After FDR correction, the FDR-significant paired contrasts in the primary LLE variant were concentrated in natural walking, as summarized in Table 2. In addition to overall LLE asymmetry, toe LLE coupling showed robust therapy-related changes. Toe coupling increased from MED OFF / STIM OFF to MED OFF / STIM ON ($n = 18$, mean difference $= 0.0993$, $p = 0.0019$, $q = 0.0427$, $d_z = 0.94$), to MED ON / STIM OFF ($n = 18$, mean difference $= 0.0843$, $p = 0.0047$, $q = 0.0427$, $d_z = 0.78$), and to MED ON / STIM ON ($n = 17$, mean difference $= 0.1270$, $p = 0.0026$, $q = 0.0427$, $d_z = 0.86$). Fast walking showed lower overall asymmetry and higher coupling than natural walking, with the same directional reduction in overall asymmetry from MED OFF / STIM OFF to MED ON / STIM ON (0.0377 to 0.0274), but the paired contrast did not survive section-level FDR correction ($p = 0.0214$, $q = 0.1925$, $d_z = -0.65$). We therefore treat fast walking as a secondary task-speed comparison rather than as a mechanistic therapy effect.

Table 2: FDR-significant paired contrasts in the primary LLE variant. Δ is condition B minus condition A; p is the nominal Wilcoxon p-value; q is the FDR-adjusted p-value; and d_z is the paired-sample standardized effect size. Identical q values can result from the monotonic Benjamini–Hochberg adjustment. Asymmetry differences are in short-window LLE-derived slope units; coupling differences are changes in unitless Pearson correlation (r).

Endpoint	Contrast	Δ	p	q	d_z
Toe LLE coupling	OFF/OFF to OFF/ON	0.0993	0.0019	0.0427	0.94
Toe LLE coupling	OFF/OFF to ON/ON	0.1270	0.0026	0.0427	0.86
Overall LLE asymmetry	OFF/OFF to ON/ON	-0.0103	0.0038	0.0427	-0.83
Toe LLE coupling	OFF/OFF to ON/OFF	0.0843	0.0047	0.0427	0.78
Overall LLE asymmetry	OFF/OFF to ON/OFF	-0.0071	0.0066	0.0474	-0.75

Standard Gait Measures Confirmed Improvement and Clarified Endpoint Meaning. In the natural-walking OFF/OFF to ON/ON comparison ($n = 17$), combined therapy increased walking speed by 0.4325 m/s ($d_z = 2.41$), step length by 0.2058 m ($d_z = 2.13$), and stride length by 0.4375 m ($d_z = 2.02$). Overall LLE-derived asymmetry changes were strongly associated with cadence ($\rho = -0.816$, $q < 0.001$), step time ($\rho = 0.775$, $q = 0.0017$), and stride time ($\rho = 0.801$, $q < 0.001$). Coupling was less reducible to gross speed or length. These endpoints are therefore complementary continuous-trajectory summaries, not replacements for standard gait measures. Table 5 and the supplementary CSV files provide the complete results.

UPDRS-III Changes Provided Initial Convergent Validity. For the same contrast, overall coupling change correlated with UPDRS-III total change ($\rho = -0.637$, $q = 0.036$) and postural-stability change ($\rho = -0.521$, $q = 0.046$). Toe coupling showed similar associations, and overall asymmetry correlated with item 30 (gait). Table 6 reports all six tests. Adjustment for walking-speed change preserved their directions but increased focused q values to about 0.09. These results support initial convergence, not clinical validity.

The Main Findings Were Stable Across Multiple Checks. The combined-therapy asymmetry reduction retained its direction and nominal significance across all eight feature-extraction settings; medication-only asymmetry and combined-therapy toe coupling retained direction in all eight and significance in seven. The Theiler stimulation result was less stable and remains secondary. Every leave-one-patient-out analysis preserved the key directions. Phase-randomized surrogates reduced asymmetry and moved coupling toward zero, supporting temporal and cross-marker organization without establishing equivalence to conventional whole-trial LLE. Table 7 and the supplement provide details.

Theiler Sensitivity Preserved the Asymmetry Direction and Emphasized Coupling. The Theiler-excluded sensitivity analysis preserved the direction of the primary asymmetry effect but attenuated the corrected paired OFF/OFF to ON/ON result. Natural-walking overall LLE asymmetry decreased from 0.0421 in MED OFF / STIM OFF to 0.0388 in MED OFF / STIM ON, 0.0369 in MED ON / STIM OFF, and 0.0344 in MED ON / STIM ON. The MED OFF / STIM OFF to MED ON / STIM ON paired contrast remained nominally significant but did not survive FDR correction ($n = 17$, mean difference $= -0.0073$, $p = 0.0150$, $q = 0.1800$, $d_z = -0.63$).

Theiler-excluded coupling endpoints showed corrected factorial stimulation effects. Natural-walking toe LLE coupling increased with stimulation ($n = 17$, mean effect $= 0.0706$, $p = 0.0005$, $q = 0.0091$, $d_z = 1.00$), and overall coupling showed a smaller corrected effect (mean effect $= 0.0561$, $p = 0.0038$, $q = 0.0346$, $d_z = 0.59$). However, the parameter-sensitivity analysis showed that the toe result was less stable than the primary paired findings. We therefore report these Theiler-excluded results as secondary evidence, not as the central conclusion. Table 3 places the primary and Theiler-excluded findings side by side.

Table 3: Key primary and Theiler-excluded natural-walking findings. Paired differences are condition B minus condition A. Factorial MED or STIM effects are within-subject 2×2 contrasts defined in Methods. The Δ/effect column therefore reports either a paired difference or a mean factorial effect. Asymmetry rows are in short-window LLE-derived slope units; coupling rows are unitless Pearson-correlation effects.

Variant	Endpoint	Contrast/effect	Δ/effect	p	q	d_z
Primary	Overall asym.	OFF/OFF to ON/ON	-0.0103	0.0038	0.0427	-0.83
Primary	Toe coupling	OFF/OFF to ON/ON	0.1270	0.0026	0.0427	0.86
Theiler	Overall asym.	OFF/OFF to ON/ON	-0.0073	0.0150	0.1800	-0.63
Theiler	Toe coupling	STIM effect	0.0706	0.0005	0.0091	1.00
Theiler	Overall coupling	STIM effect	0.0561	0.0038	0.0346	0.59

Stimulation Effects Were Strongest for Bilateral Coupling. Factorial analyses separated medication and stimulation effects among participants with complete observations in all four therapy conditions. In the primary LLE variant, FDR-significant natural-walking effects were concentrated in coupling endpoints: stimulation increased overall LLE coupling (mean effect $= 0.0656$, $p = 0.0093$, $q = 0.0420$, $d_z = 0.62$), heel LLE coupling (mean effect $= 0.0478$, $p = 0.0093$, $q = 0.0420$,

$d_z = 0.75$), and toe LLE coupling (mean effect $= 0.0696$, $p = 0.0056$, $q = 0.0420$, $d_z = 0.80$). Toe coupling also showed a significant medication main effect (mean effect $= 0.0574$, $p = 0.0067$, $q = 0.0420$, $d_z = 0.76$). Together, the paired and factorial analyses suggest a division of labor between endpoint types. Overall LLE asymmetry is the clearest paired clinical response marker, whereas bilateral LLE coupling is the strongest factorial marker of stimulation-related coordination changes.

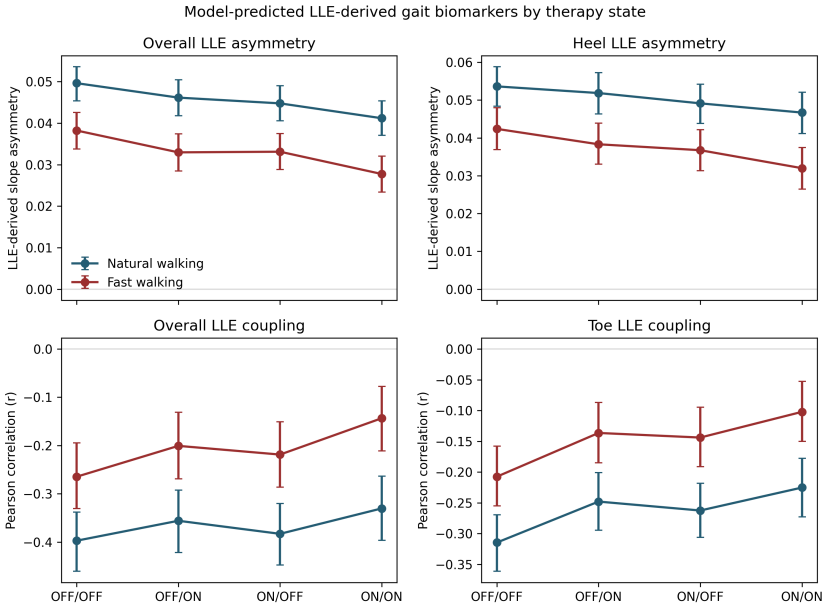

Figure 2: Bayesian predictions by MED/STIM state and walking task. Top: short-window LLE-derived asymmetry; bottom: Pearson coupling (r). Points are posterior means; bars show 95% credible intervals.

Bayesian Hierarchical Machine Learning Identified Therapy-Dependent LLE Signatures. We next used Bayesian hierarchical regression to estimate medication and stimulation effects while accounting for repeated measurements within patients. The model structure is defined in Equations 3–6. The key posterior coefficients are summarized in Table 4; capital $P(\cdot)$ summarizes how strongly the Bayesian model supports the expected direction of the therapy effect. Model-predicted LLE biomarker trajectories are shown in Figure 2.

Table 4: Bayesian hierarchical regression summary for primary LLE-derived biomarkers. Coefficients are on the standardized outcome scale. CI denotes 95% credible interval. Posterior direction probability is the probability, under the fitted Bayesian model, that the therapy effect points in the expected clinical direction.

Endpoint	Effect	Mean β	95% CI	Post. dir. prob.
Overall LLE asymmetry	MED	-0.410	$[-0.664, -0.147]$	$P(\beta < 0) = 0.9985$
Overall LLE asymmetry	STIM	-0.352	$[-0.613, -0.097]$	$P(\beta < 0) = 0.9960$
Heel LLE asymmetry	MED	-0.350	$[-0.606, -0.093]$	$P(\beta < 0) = 0.9950$
Heel LLE asymmetry	STIM	-0.198	$[-0.458, 0.050]$	$P(\beta < 0) = 0.9347$
Toe LLE coupling	MED	0.436	$[0.143, 0.730]$	$P(\beta > 0) = 0.9985$
Toe LLE coupling	STIM	0.521	$[0.214, 0.824]$	$P(\beta > 0) = 1.0000$
Overall LLE coupling	STIM	0.298	$[0.009, 0.592]$	$P(\beta > 0) = 0.9772$

For overall LLE asymmetry, the posterior supported reductions associated with both medication and stimulation. In the primary LLE variant, the medication coefficient was negative ($\beta_{\text{MED}} = -0.410$, 95% credible interval $[-0.664, -0.147]$, $P(\beta_{\text{MED}} < 0) = 0.9985$), and the stimulation coefficient was also negative ($\beta_{\text{STIM}} = -0.352$, 95% credible interval $[-0.613, -0.097]$, $P(\beta_{\text{STIM}} <$

0.9960). The interaction coefficient was close to zero, suggesting an approximately additive reduction of overall asymmetry. Heel LLE asymmetry showed the clearest medication-sensitive asymmetry pattern. In the primary variant, medication reduced heel asymmetry ($\beta_{\mathrm{MED}} = -0.350$, 95% credible interval $[-0.606, -0.093]$, $P(\beta_{\mathrm{MED}} < 0) = 0.995$), whereas the stimulation effect was weaker ($\beta_{\mathrm{STIM}} = -0.198$, 95% credible interval $[-0.458, 0.050]$, $P(\beta_{\mathrm{STIM}} < 0) = 0.935$). This supports the clinical hypothesis that dopaminergic medication reduces lateralized gait-dynamics asymmetry.

Coupling endpoints showed a different pattern. Toe LLE coupling was strongly therapy-sensitive, with positive medication and stimulation coefficients in the primary variant ($\beta_{\mathrm{MED}} = 0.436$, 95% credible interval $[0.143, 0.730]$, $P(\beta_{\mathrm{MED}} > 0) = 0.9985$; $\beta_{\mathrm{STIM}} = 0.521$, 95% credible interval $[0.214, 0.824]$, $P(\beta_{\mathrm{STIM}} > 0) = 1.0000$). Thus, toe coupling captured therapy-dependent reorganization of bilateral gait coordination. Overall LLE coupling showed a weaker but directionally similar stimulation effect. The Theiler-excluded LLE variant preserved the same qualitative conclusions. Overall LLE asymmetry again showed negative medication and stimulation coefficients, and toe LLE coupling again showed positive medication and stimulation coefficients. The similar qualitative patterns across both variants reduce concern that the findings depend entirely on a single nearest-neighbour definition.

Across the 12 models, maximum $\hat{R}$ was 1.01, minimum bulk effective sample size was 976, and no divergences occurred. All observed means fell within 95% posterior predictive intervals; 13 of 372 spread checks fell outside, indicating adequate mean-effect fit but residual subgroup variance.

4 Discussion

Parkinson's disease is related to dopamine deficiency in basal-ganglia circuits that regulate movement initiation, scaling, and coordination with the environment. Dopaminergic medication, most commonly L-DOPA, acts as systemic pharmacological compensation, whereas DBS targets a small, deep-brain structure and acts as a focal neuromodulatory intervention whose effect depends on electrode location, stimulation field, stimulation parameters, and symptom lateralization. These mechanistic differences motivate the central question of whether medication and DBS leave different signatures in gait dynamics.

Our results suggest that medication and DBS leave complementary yet partially overlapping signatures in gait dynamics. Overall LLE asymmetry was the clearest paired clinical response marker: it decreased from MED OFF/STIM OFF to MED ON/STIM ON during natural walking, and also in the medication-sensitive MED OFF/STIM OFF to MED ON/STIM OFF contrast. This pattern is consistent with dopaminergic reduction of lateralized dysregulation and agrees with prior studies showing that lower-limb asymmetry, gait variability, and coordination measures are clinically meaningful and medication-sensitive in Parkinsonian walking [11–15]. Standard gait measures confirmed the expected therapy response, while focused UPDRS-III correlations provided initial convergent validity. Coupling endpoints showed a different pattern. Toe and overall LLE coupling showed therapy-related changes, while corrected Theiler stimulation effects were less stable under parameter sensitivity and remain secondary. Coupling is not a signed Lyapunov exponent or direct chaos or stability value, but a Pearson correlation between left- and right-sided LLE-derived profiles. Negative coupling reflects anti-correlated left-right local divergence profiles, compatible with the alternating phase structure of gait. A stimulation-related change in coupling therefore suggests that DBS may influence bilateral coordination or phase organization, consistent with prior DBS findings [16].

Taken together, the findings support complementary roles for the LLE-derived endpoints. Asymmetry is the more direct response marker, especially for medication and combined therapy, whereas coupling captures coordination and stimulation-related change. Short-window LLE features summarize continuous foot-marker dynamics beyond discrete step-level variables. The hierarchical model respects repeated measurements and estimates group-level therapy effects with patient-specific baselines, not patient-specific therapy slopes. The therapy-state-resolved design is also clinically valuable because each patient serves as their own control across medication and stimulation states. This within-patient cohort provides initial convergent evidence; larger external cohorts are needed to assess subgroup effects, additional clinical moderators, conventional-LLE equivalence, and clinical utility. Future studies should retain laboratory-grade motion capture in specialized gait facilities as a reference standard while evaluating scalable wearable and markerless sensing.

Acknowledgements and disclosures

The authors thank PJAIT Bytom for providing the research facilities and support necessary for this study. We also thank Mr. Tomasz Pytel for his assistance with data acquisition. Information about the institution is available at https://bytom.pja.edu.pl/.

Funding. The authors received no external grant funding or commercial support for this work during the 36 months before submission. PJAIT Bytom provided institutional research facilities and support as acknowledged above.

Competing interests. The authors declare no financial relationships or other competing interests during the 36 months before submission that could be perceived to influence this work.

References

[1] Rosenstein, M. T., Collins, J. J., & De Luca, C. J. (1993). A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1–2), 117–134.

[2] Bruijn, S. M., van Dieën, J. H., Meijer, O. G., & Beek, P. J. (2009). Is slow walking more stable? *Journal of Biomechanics*, 42(10), 1506–1512.

[3] Kang, H. G., & Dingwell, J. B. (2009). Dynamic stability of superior vs. inferior segments during walking in young and older adults. *Gait & Posture*, 30(2), 260–263.

[4] Amirpourabasi, A., Lamb, S. E., Chow, J. Y., & Williams, G. K. R. (2022). Nonlinear dynamic measures of walking in healthy older adults: a systematic scoping review. *Sensors*, 22(12), 4408.

[5] Mancini, M., Afshari, M., Almeida, Q., Amundsen-Huffmaster, S., Balfany, K., Camicioli, R., Christiansen, C., Dale, M. L., Dibble, L. E., Earhart, G. M., et al. (2025). Digital gait biomarkers in Parkinson's disease: susceptibility/risk, progression, response to exercise, and prognosis. *npj Parkinson's Disease*, 11(1), 51.

[6] Cury, R. G., Pavese, N., Aziz, T. Z., Krauss, J. K., Moro, E., & Neuromodulation of Gait Study Group from Movement Disorders Society. (2022). Gaps and roadmap of novel neuromodulation targets for treatment of gait in Parkinson's disease. *npj Parkinson's Disease*, 8(1), 8.

[7] Fröhlich, H., Bontridder, N., Petrovska-Delacéta, D., Glaab, E., Kluge, F., Yacoubi, M. E., Marin Valero, M., Corvol, J.-C., Eskofier, B., Van Gyseghem, J.-M., et al. (2022). Leveraging the potential of digital technology for better individualized treatment of Parkinson's disease. *Frontiers in Neurology*, 13, 788427.

[8] Fekri Azgomi, H., Louie, K. H., Bath, J. E., Presbrey, K. N., Balakid, J. P., Marks, J. H., Wozny, T. A., Galifianakis, N. B., San Luciano, M., Little, S., et al. (2025). Modeling and optimizing deep brain stimulation to enhance gait in Parkinson's disease: personalized treatment with neurophysiological insights. *npj Parkinson's Disease*, 11(1), 173.

[9] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459.

[10] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

[11] Arippa, F., Leban, B., Monticone, M., Cossu, G., Casula, C., & Pau, M. (2022). A study on lower limb asymmetries in Parkinson's disease during gait assessed through kinematic-derived parameters. *Bioengineering*, 9(3), 120.

[12] Bryant, M. S., Rintala, D. H., Hou, J. G., Charness, A. L., Fernandez, A. L., Collins, R. L., Baker, J., Lai, E. C., & Protas, E. J. (2011). Gait variability in Parkinson's disease: influence of walking speed and dopaminergic treatment. *Neurological Research*, 33(9), 959–964.

[13] Agurto, C., Heisig, S., Abrami, A., Ho, B. K., & Caggiano, V. (2021). Parkinson's disease medication state and severity assessment based on coordination during walking. *PLoS One*, 16(2), e0244842.

[14] McNeely, M. E., Duncan, R. P., & Earhart, G. M. (2012). Medication improves balance and complex gait performance in Parkinson disease. *Gait & Posture*, 36(1), 144–148.

[15] Galna, B., Lord, S., Burn, D. J., & Rochester, L. (2015). Progression of gait dysfunction in incident Parkinson's disease: impact of medication and phenotype. *Movement Disorders*, 30(3), 359–367.

[16] Johnsen, E. L., Mogensen, P. H., Sunde, N. A., & Østergaard, K. (2009). Improved asymmetry of gait in Parkinson's disease with DBS: gait and postural instability in Parkinson's disease treated with bilateral deep brain stimulation in the subthalamic nucleus. *Movement Disorders*, 24(4), 590–597.

[17] Ihlen, E. A. F., Goihl, T., Wik, P. B., Sletvold, O., Helbostad, J., & Vereijken, B. (2012). Phase-dependent changes in local dynamic stability of human gait. *Journal of Biomechanics*, 45(13), 2208–2214.

[18] van Schooten, K. S., Rispens, S. M., Elders, P. J. M., van Dieën, J. H., & Pijnappels, M. A. G. M. (2014). Toward ambulatory balance assessment: estimating variability and stability from short bouts of gait. *Gait & Posture*, 39(2), 695–699.

[19] Terrier, P., & Reynard, F. (2018). Maximum Lyapunov exponent revisited: long-term attractor divergence of gait dynamics is highly sensitive to the noise structure of stride intervals. *Gait & Posture*, 66, 236–241.

A Validation and robustness details

Table 5: Natural-walking OFF/OFF to ON/ON changes in standard gait measures ($n = 17$). Differences are ON/ON minus OFF/OFF.

Measure	OFF/OFF	ON/ON	Difference	d_z
Walking speed (m/s)	0.5190	0.9515	0.4325	2.41
Step length (m)	0.3323	0.5381	0.2058	2.13
Stride length (m)	0.6670	1.1045	0.4375	2.02
Stride-length asymmetry (m)	0.0225	0.0706	0.0481	1.33

Table 6: Focused convergent-validity tests for natural-walking OFF/OFF to ON/ON changes ($n = 17$). Focused q values control FDR across the six planned comparisons.

LLE-derived endpoint	UPDRS-III change	ρ	p	focused q
Overall asymmetry	Item 30 (gait)	0.505	0.039	0.046
Toe coupling	Total	-0.567	0.018	0.046
Toe coupling	Postural stability	-0.512	0.036	0.046
Overall coupling	Total	-0.637	0.006	0.036
Overall coupling	Postural stability	-0.521	0.032	0.046
Theiler overall asymmetry	Item 30 (gait)	0.485	0.049	0.049

Table 7: Stability across the baseline configuration and seven one-factor perturbations. The range is the mean paired difference or factorial effect.

Result	Direction agreement	Nominal $p < 0.05$	Effect range
Overall asymmetry, combined therapy	8/8	8/8	$[-0.0111, -0.0061]$
Overall asymmetry, medication only	8/8	7/8	$[-0.0076, -0.0056]$
Toe coupling, combined therapy	8/8	7/8	$[0.0518, 0.2341]$
Theiler toe coupling, STIM effect	6/8	5/8	$[-0.0531, 0.0876]$

The supplementary release provides the complete standard-metric table, all change-to-change correlations including null results, speed-adjusted clinical correlations, leave-one-patient-out outputs, phase-randomized surrogate results, parameter configurations, and posterior predictive diagnostics.

B Code and Data Availability

The code, derived datasets, and supplementary analysis results supporting this study are publicly archived on Zenodo (version 1.0.0): https://doi.org/10.5281/zenodo.23962355.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction frame the work as a Use-Inspired therapy-response analysis. They call the endpoints candidate biomarkers, distinguish the short-window features from conventional whole-trial LLE, and limit the claims to group-level effects in this cohort.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper qualifies the coupling and Theiler-sensitivity findings, distinguishes the short-window features from conventional whole-trial LLE, limits the clinical correlations to initial convergent evidence, reports residual subgroup variance, and states that larger external cohorts are needed to evaluate subgroup effects, additional clinical moderators, conventional-LLE equivalence, and clinical utility.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical theorems or proofs. The mathematical content consists of methodological definitions for delay embedding, local divergence estimation, and Bayesian hierarchical regression.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The Methods specify the cohort, signals, preprocessing, feature parameters, endpoint definitions, statistics, Bayesian model and priors, standard gait metrics, focused clinical validation, parameter perturbations, surrogate analysis, leave-one-patient-out checks, and posterior predictive checks.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The licensed Zenodo archive provides de-identified participant-session-task features, standard gait metrics, analysis code, reported outputs, and run instructions under a version-specific DOI. Raw trajectories and linked clinical source records are excluded because unrestricted release is not authorized.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper describes the therapy design, tasks, signals, feature parameters, endpoint aggregation, statistical corrections, Bayesian predictors and priors, validation analyses, robustness settings, and posterior predictive procedure. Complete machine-readable outputs are supplied separately.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The Results report paired-sample tests, sample sizes, mean differences, Wilcoxon p values, Benjamini–Hochberg FDR-adjusted q values, paired effect sizes d_z, and Bayesian posterior means with 95% credible intervals and direction probabilities. Figure error bars are defined as 95% credible intervals.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The Bayesian model script writes a compute report and per-model runtime table. In the current run, 12 Bayesian models with 4 chains, 1000 tuning iterations, and 1000 posterior draws completed on a CPU workstation with a 13th Gen Intel Core i9-13900KF, 32 logical CPU cores, and 31.8 GB RAM in 90.5 seconds, approximately 1.5 minutes. No GPU was used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work analyzes anonymized clinical motion-capture and UPDRS-derived features for therapy-response estimation. The paper does not propose deployment for autonomous clinical decision-making and frames the biomarkers as research tools requiring further validation.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive clinical-research impacts through interpretable therapy-response biomarkers and acknowledges limitations that may affect clinical translation, including small cohort size and missing stimulation/electrode metadata. Potential harms are mitigated by presenting the method as exploratory and not as a standalone diagnostic or treatment-adjustment system.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk dual-use model or scraped dataset. The release is limited to de-identified derived features and analysis code; raw biometric and linked clinical records remain restricted.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the methodological work used for LLE-derived analysis, gait measurement, and Bayesian modeling, and the supplementary package records software dependencies. The release applies the MIT License to author-written code and CC BY 4.0 to author-produced derived data and documentation. Raw participant data and third-party materials are excluded from these licenses and from the release.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper and supplementary documentation describe the de-identified participant-session-task feature dataset, therapy states, tasks, marker features, standard gait metrics, endpoint definitions, aggregation, limitations, and permitted scope of release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

Justification: The paper does not involve crowdsourcing. It analyzes clinical motion-capture and UPDRS data from consenting participants; the Methods section reports ethics approval, informed consent, cohort structure, therapy states, walking tasks, and clinical scoring. Screenshots, crowd-worker instructions, and compensation details are not applicable to this clinical feature-analysis submission.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The Methods section states that data collection was approved by the relevant institutional ethics committee and that all participants provided informed consent. The study used non-invasive motion capture and clinical assessments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a core methodological component of the research. The core methods are motion-capture preprocessing, short-window LLE-derived feature extraction, statistical testing, and Bayesian hierarchical regression.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.