CASCADE Conformal Prediction: Uncertainty-Adaptive Prediction Intervals for Two-Stage Clinical Decision Support

Ricardo Diaz-Rincon [1] **Muxuan Liang** [2] **Adolfo Ramirez-Zamora** [3] **Benjamin Shickel** [1]

Abstract

Effective medication management in Parkinson's Disease (PD) is challenging due to heterogeneous disease progression, variable patient response, and medication side effects. While AI models can forecast levodopa equivalent daily dose (LEDD) as a measure of medication needs, standard uncertainty quantification often fails to communicate the reliability of these predictions, treating high and low confidence clinical decisions identically. We introduce CASCADE (Calibrated Adaptive Scaling via Conformal And Distributional Estimation), a novel conformal prediction framework that propagates epistemic uncertainty from a screening classifier to adapt downstream predictions. Unlike standard conformal methods that rely on auxiliary residual regression, we leverage epistemic uncertainty from a primary classification task (identifying whether a medication change is needed) to dynamically scale the prediction intervals of a secondary regression task (predicting how much change). By mapping Venn-Abers multi-probabilistic uncertainty directly to non-conformity scores, our framework achieves continuous risk adaptation. We demonstrate that this "cascade effect" produces highly efficient intervals for confident patients (38.9% narrower than standard conformal baselines) while automatically expanding intervals to ensure robust coverage for uncertain cases, bridging the gap between discrete clinical decision-making and continuous dose forecasting in PD.

1. Introduction

Parkinson's Disease (PD) medication management is a high-stakes sequential decision problem: first determining whether a dosage adjustment is needed, then quantifying how much. Dosing errors carry direct patient risk: inadequate titration leaves patients with debilitating motor symptoms, while over-medication can trigger levodopa-induced dyskinesia (LID) and neuropsychiatric complications (Armstrong & Okun, 2020; Espay et al., 2016). The levodopa equivalent daily dose (LEDD) standardizes medication burden across drug classes (Tomlinson et al., 2010), but optimal titration remains a subjective, trial-and-error process heavily reliant on brief clinical assessments (Armstrong & Okun, 2020).

Recent advancements in artificial intelligence (AI) have accelerated the adoption of clinical decision support systems. Because most clinical decisions follow a hierarchical logic: first determining *if* an intervention is needed before deciding *how much*, these systems increasingly rely on two-stage architectures. In PD, such a system operates in two stages: first, a classifier identifies if a patient's medication regimen requires adjustment; second, a regression model predicts how much the dosage should change (Diaz-Rincon et al., 2025b). However, applying uncertainty quantification to these two-stage architectures presents a critical challenge: information loss at the decision boundary. While conformal prediction (CP) provides distribution-free coverage guarantees for such models (Vovk et al., 2005; Shafer & Vovk, 2008), it typically treats the regression stage in isolation, regardless of first-stage uncertainty.

Consider a neurologist using such a system for two patients: For patient A, the system is 99% certain a medication change is needed. For patient B, the system is ambivalent, predicting a change with 55% certainty (hovering near the 0.5 decision boundary). A standard regression model, blind to this upstream wavering, treats these two patients identically, resulting in static conservatism. It might predict a 20% medication increase for both, assigning similar confidence intervals. However, the clinical risk profile is vastly different: the prediction for patient A is actionable, while the prediction for patient B is uncertain, carrying a higher risk of errors that could potentially lead to levodopa-induced dyskinesia (Espay et al., 2016; Olanow et al., 2006; Jenner, 2008) and other complications (Newman et al., 2009; Taylor et al., 2016). This scenario, illustrated in Figure 2(a), exposes a critical gap that limits the adoption of uncertainty quantification in high-stakes healthcare settings.

[1]University of Florida [2]MD Anderson Cancer Center [3]University of Louisville. Correspondence to: Benjamin Shickel <shickelb@ufl.edu>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

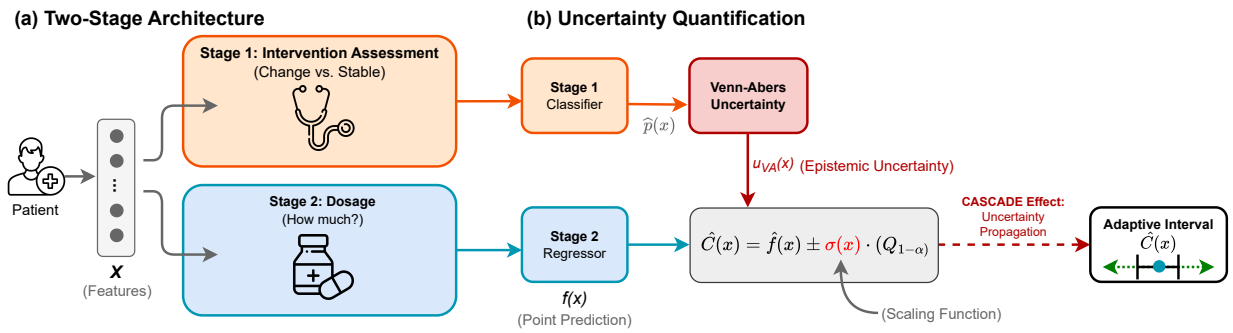

Figure 1. Overview of the CASCADE Framework. **(a) Two-Stage Architecture:** A Stage 1 classifier identifies patients requiring medication adjustment, followed by a Stage 2 regressor that predicts the dosage. **(b) Uncertainty Quantification:** CASCADE extracts the epistemic uncertainty score $u_{\text{VA}}(x)$ directly from the Stage 1 predicted probability $\hat{p}(x)$. This score flows into the scaling function $\sigma(x)$ (highlighted in red), which scales the prediction interval. This mathematical propagation drives the "cascade effect" (red dashed arrow), where the final adaptive interval $\hat{C}(x)$ expands or shrinks based on patient uncertainty.

To bridge this gap, we introduce CASCADE (Calibrated Adaptive Scaling via Conformal And Distributional Estimation), a framework that explicitly couples classification uncertainty with regression calibration. Instead of treating these stages independently, we propose that the epistemic uncertainty of the Stage 1 decision, measured through Venn-Abers (VA) predictors (Vovk & Petej, 2012), should serve as a dynamic scaling factor for the Stage 2 prediction intervals. This creates a "cascade effect": as the system's uncertainty increases, the prediction intervals automatically widen to reflect this. Clinically, this ensures that the model's reliability scales with the difficulty of the case, acting as a safeguard by preventing dangerous overconfidence in complex scenarios while maintaining high precision for straightforward patients.

Our contributions. (1) We introduce the CASCADE framework, creating a mechanism to map Venn-Abers uncertainty directly to conformal non-conformity scores. This enables cross-task uncertainty transfer, leveraging classification reliability to scale regression intervals without the computational overhead of training auxiliary error models. (2) We implement CASCADE using two distinct strategies: a discrete Mondrian approach and a Continuous Cascade approach using mean-centered scaling. Within the continuous implementation we introduce a tunable scaling parameter (β), allowing practitioners to explicitly calibrate the system's sensitivity to uncertainty. We demonstrate that this continuous parameterization removes the discretization artifacts and sample-size fragmentation inherent in traditional binning approaches. (3) We validate this approach on a decade of PD inpatient data from 631 patients at the University of Florida (UF) Health, showing that CASCADE maintains marginal coverage guarantees while reducing interval length by 38.9% for high-confidence patients compared to baselines.

2. Related Work

Machine Learning in PD Medication Management. To facilitate medication decisions, recent approaches have applied machine learning to forecast medication needs through the LEDD (Diaz-Rincon et al., 2024). These range from predicting raw levodopa doses (Salmanpour et al., 2023) and post-surgical adjustments (Shamir et al., 2015) to clustering medication regimes (Watts et al., 2021) and optimizing combinations via reinforcement learning (Kim et al., 2021). While these approaches achieve adequate accuracy, they predominantly rely on point predictions without accounting for prediction uncertainty. In high-stakes clinical settings, a single deterministic value without reliability bounds is insufficient, as it fails to distinguish between a robust forecast and a statistical guess. Without uncertainty bounds, models may encourage overconfident interventions in ambiguous cases, potentially leading to suboptimal patient outcomes (Belić et al., 2019). Consequently, there is a critical need for rigorous uncertainty quantification mechanisms that can translate model confidence into clinically interpretable safety bounds.

Conformal and Adaptive Conformal Prediction. Conformal prediction (Vovk et al., 2005; Shafer & Vovk, 2008) addresses this need by providing prediction intervals with distribution-free, finite-sample coverage guarantees. Methods include Naïve, Standard (Split), Cross-Validation (CV+) (Barber et al., 2021) and Jackknife+-After-Bootstrap (J+aB) (Kim et al., 2020). However, marginal validity can mask conditional failure: a model may achieve 80% marginal coverage by over-covering "easy" patients (samples) while systematically under-covering "difficult" ones. We implement these four baseline approaches for comparison. Comprehensive details about these methods are provided in Appendix B.2. To address heterogeneity, Mondrian (or Group-Balanced) CP (Vovk et al., 2005) achieves group-conditional validity by partitioning data into strata (e.g., by disease

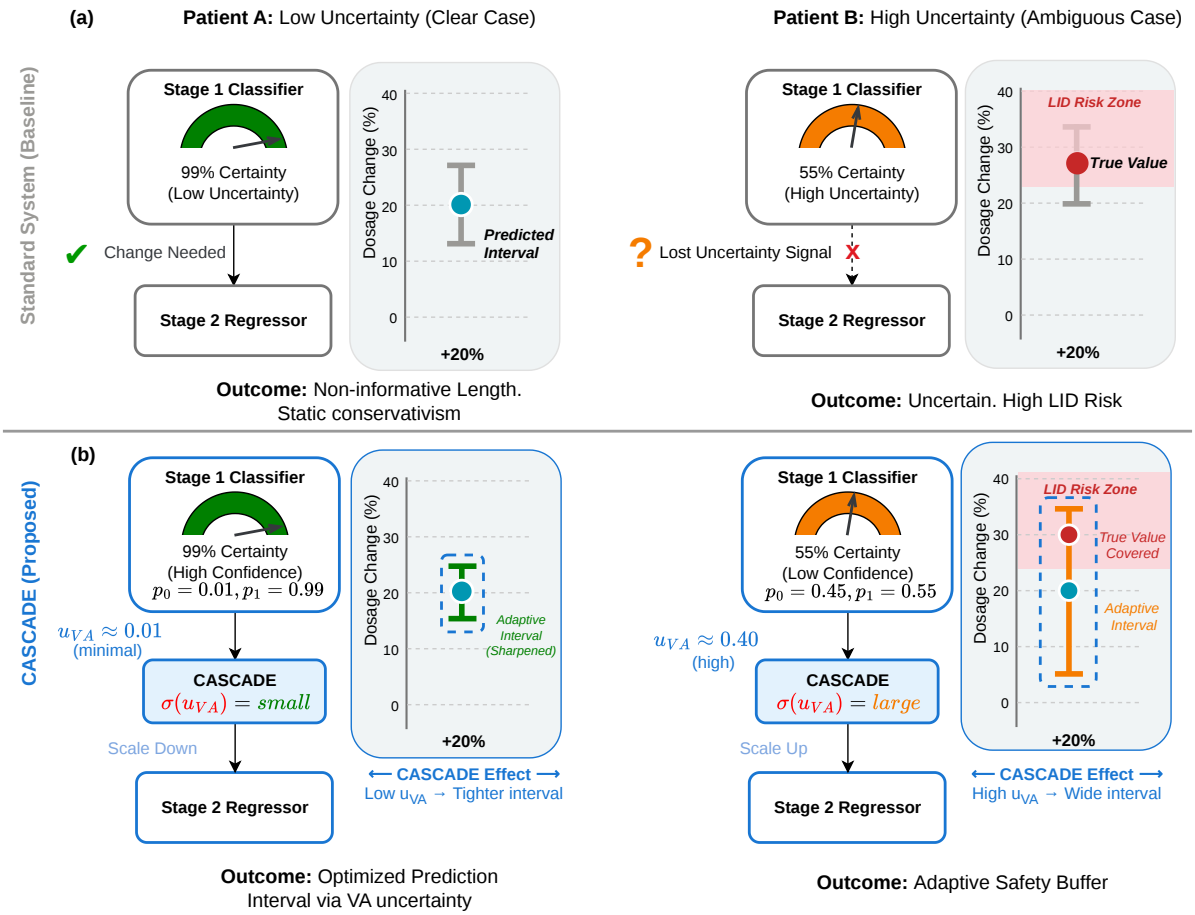

Figure 2. The CASCADE Mechanism. **(a) Standard System:** The regression model is blind to upstream confidence, producing non-informative, fixed-length prediction intervals for both the low uncertainty Patient A ($p = 0.99$) and the high uncertainty (ambiguous) Patient B ($p = 0.55$). While actionable for A, this rigidity is dangerous for B: the model confidently predicts a dosage increase despite high ambiguity. This erroneous precision carries a high risk of over-medication, potentially driving the patient into the levodopa-induced dyskinesia "(LID) Risk Zone". **(b) CASCADE (Proposed):** The framework solves this by explicitly coupling the two stages, translating the epistemic uncertainty from Stage 1 into Stage 2. Left (Patient A): High confidence results in a minimal Venn-Abers score (u_{VA}). The mechanism responds with compressive scaling, producing an optimized prediction, eliminating unnecessary conservatism and confirming the dosage target is safe. Right (Patient B): The increased uncertainty yields a high (u_{VA}). The mechanism responds with expansive scaling, creating an adaptive safety buffer that covers the ground truth and explicitly signals that a standard dosage increase carries significant risk.

stage) and calibrating intervals independently within each group. Complementarily, Normalized (or locally scaled) CP (Lei et al., 2018; Papadopoulos et al., 2011) focuses on adaptivity, scaling interval length based on a difficulty estimate $\sigma(x)$. However, both strategies primarily focus on single-stage prediction tasks without addressing the challenge of uncertainty propagation in cascaded systems, where an upstream classification decision inherently limits the reliability of downstream predictions.

Uncertainty Quantification in Two-Stage Systems. In clinical settings, prediction often follows a two-stage logic: first determining if an intervention is needed (classification), then assessing how much (regression). While multi-stage architectures effectively handle zero-inflated data by separating decision and prediction tasks (Diaz-Rincon et al., 2025a), they often introduce information loss at the decision boundary. Current frameworks typically decouple uncertainty quantification across these stages, utilizing the classification threshold as a hard boundary: once a patient crosses the decision threshold, the probabilistic ambiguity of that decision is discarded.

Consequently, the epistemic uncertainty of the Stage 1 decision ("Does this patient truly require a medication change?") is rarely propagated to Stage 2 ("How much change?"). This

creates a reliability mismatch, where predictions for borderline cases appear deceptively confident, failing to reflect the compounded risk of the hierarchical decision process. Our work bridges this gap by introducing a cascade mechanism that prevents this information loss, directly mapping classification reliability to downstream regression.

3. Methods

3.1. Problem Formulation

Consider a two-stage prediction setting characterized by a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. Where $x_i \in \mathbb{R}^d$ is a feature vector encoding patient characteristics (e.g., age, demographics, clinical variables), and the target y_i is governed by a latent decision process followed by a value estimation. While the framework applies to any such system, we demonstrate it on Parkinson's Disease medication management, where $y_i \in \mathbb{R}$ denotes the percentage change in LEDD ($y_i = 0$ indicates no medication adjustment). To perform conformal calibration, we randomly partition the data into a proper training set $\mathcal{D}_{\text{train}}$ and a hold-out calibration set $\mathcal{D}_{\text{cal}}$ (comprising 20% of $\mathcal{D}$). Let $\hat{f}$ denote the underlying regression model trained on $\mathcal{D}_{\text{train}}$, which outputs a prediction $\hat{f}(x)$ for a given input.

Our objective is to construct a prediction set $\hat{C}(x_{\text{new}})$ that satisfies marginal coverage:

$$\mathbb{P}(y_{\text{new}} \in \hat{C}(x_{\text{new}})) \geq 1 - \alpha \tag{1}$$

where $1 - \alpha$ is the confidence level (e.g., 80%). Crucially, beyond this marginal guarantee, we seek to improve local adaptivity, ensuring the interval length dynamically scales with the reliability of the upstream classification decision.

3.2. The CASCADE Framework

To address the information loss in standard two-stage systems, we introduce CASCADE (Calibrated Adaptive Scaling via Conformal And Distributional Estimation). The core principle of this framework is cross-task uncertainty transfer: we postulate that the epistemic uncertainty of Stage 1, which identifies whether a patient requires medication changes, is a critical signal for the reliability of the downstream regression task (Stage 2), which predicts the specific dosage.

Instead of treating the stages as statistically independent, CASCADE propagates the reliability of the classification step directly into the conformal calibration process, operating in two phases: (1) **Epistemic Uncertainty Quantification (Stage 1):** quantifying the ambiguity of the binary classifier using Venn-Abers calibration; and (2) **Adaptive Conformal Calibration (Stage 2):** scaling the regression prediction intervals based on the Stage 1 uncertainty.

3.3. Stage 1: Epistemic Uncertainty Estimation

We first train a classifier to estimate the predicted probability of the event of interest (e.g., medication change), denoted as $\hat{p}(x)$. To quantify the epistemic uncertainty of this decision, we employ Venn-Abers predictors (Vovk & Petej, 2012). Unlike standard probability estimates (e.g., softmax) which produce a single scalar and are often poorly calibrated in non-linear models (Guo et al., 2017), Venn-Abers provides a rigorous multi-probabilistic interval ($[p_0(x), p_1(x)]$). This interval explicitly captures epistemic uncertainty without requiring distributional assumptions.

We define the Venn-Abers Uncertainty Score $u_{\text{VA}}(x)$ as the length of this probability interval:

$$u_{\text{VA}}(x) = p_1(x) - p_0(x). \tag{2}$$

$u_{\text{VA}}(x)$ serves as a rigorous proxy for the stability of the clinical decision: a wide interval implies high ambiguity in whether the patient truly requires adjustment, regardless of the predicted class probability.

3.4. Stage 2: Adaptive Conformal Calibration

For the regression task (Stage 2), we focus on patients where medication adjustments are indicated. In our experiments, to isolate the performance of the conformal calibration from upstream classification errors, we perform our evaluation on the subset of patients requiring medication changes as indicated by their ground truth class ($y_i \neq 0$). Additional experiments for $\hat{y} \neq 0$ (using the predicted probability of medication changes) can be found in Appendix F.

We propose two distinct strategies for constructing uncertainty-aware intervals:

3.4.1. Mondrian (Discrete Stratified CP)

This implementation utilizes Mondrian Conformal Prediction (Vovk et al., 2005) as our discrete baseline. We a priori partition the calibration set $\mathcal{D}_{\text{cal}}$ into K disjoint strata $G_1, \ldots, G_K$ based on fixed quantiles of the uncertainty score $u_{\text{VA}}(x)$. In our experiments, we evaluate granularities of $K = 3$ (tertiles). Additional results for $K \in \{5, 7\}$ can be found in Appendix D Table 6. This method creates independent calibration pools, ensuring validity within each stratum although suffering from data fragmentation (reduced effective sample size).

For a new patient x_{new}, we identify the stratum k such that $u_{\text{VA}}(x_{\text{new}}) \in G_k$. The prediction interval is constructed using only the residuals of calibration examples in that same uncertainty stratum:

$$\hat{C}_{\text{strat}}(x_{\text{new}}) = \left[\hat{f}(x_{\text{new}}) - q_r^{(k)},\ \hat{f}(x_{\text{new}}) + q_r^{(k)}\right], \tag{3}$$

where $q_r^{(k)}$ is the $(1 - \alpha)$ quantile of the absolute residuals

$|y_i - \hat{f}(x_i)|$ computed solely from the sub-population in G_k. While valid, this method reduces the effective sample size for calibration to N_{cal}/K, a known limitation of Mondrian CP that can increase variance in the estimated quantiles (Angelopoulos & Bates, 2021).

3.4.2. Continuous CASCADE (Mean-Centered Scaling)

To avoid discretization artifacts and sample fragmentation, we propose a continuous scaling mechanism based on Normalized (Locally Scaled) Conformal Prediction (Lei et al., 2018). We define a scaling function $\sigma(x)$ that adapts the interval length relative to the population's average uncertainty, parameterized by a sensitivity factor $\beta \geq 0$:

$$\sigma(x) = 1 + \beta \left(\frac{u_{\text{VA}}(x)}{\bar{u}_{\text{VA}}} - 1 \right), \tag{4}$$

where $\bar{u}_{\text{VA}}$ is the mean VA uncertainty of the calibration set. This mean-centered scaling creates a logical pivot: when $u_{\text{VA}}(x) \approx \bar{u}_{\text{VA}}$, $\sigma(x) \approx 1$ (standard length); when $u_{\text{VA}}(x) > \bar{u}_{\text{VA}}$, $\sigma(x) > 1$ (interval expands); and when $u_{\text{VA}}(x) < \bar{u}_{\text{VA}}$, $\sigma(x) < 1$ (interval shrinks). We define the scaled non-conformity scores for each calibration point i as:

$$S_i = \frac{|y_i - \hat{f}(x_i)|}{\sigma(x_i)}. \tag{5}$$

We then compute $Q_{1-\alpha}$, which is the $(1-\alpha)$-th empirical quantile of the normalized scores $\{S_i\}$ across the full calibration set. The final prediction interval is:

$$\begin{aligned} \hat{C}_{\text{cont}}(x_{\text{new}}) = \Big[& \hat{f}(x_{\text{new}}) - Q_{1-\alpha} \cdot \sigma(x_{\text{new}}), \\ & \hat{f}(x_{\text{new}}) + Q_{1-\alpha} \cdot \sigma(x_{\text{new}}) \Big]. \end{aligned} \tag{6}$$

Continuous CASCADE uses the entire calibration set (n samples) to estimate the single non-conformity quantile $Q_{1-\alpha}$, whereas Mondrian CP splits the data into small bins. This global estimation ensures stable scaling factors and maintains robust statistical power, allowing the interval length to scale continuously with the difficulty of the clinical decision, as illustrated in Figure 2(b).

4. Evaluation Metrics

4.1. Conformal Metrics

We assess model performance on the test set filtered on $y_i \neq 0$ (i.e., patients requiring medication changes as indicated by their ground truth class). This focuses the evaluation on actionable cases, measuring how well the prediction intervals cover the ground truth dosage change given that an intervention is necessary. We employ three metrics:

Marginal Coverage $(1-\alpha)$: Defined as the proportion of test samples where the true outcome lies within the predicted interval. We target a nominal coverage level 80% ($\alpha = 0.2$). This value aligns with standard benchmarks in conformal prediction literature (Angelopoulos & Bates, 2023; Vovk et al., 2005) and prioritizes the reliability-precision equilibrium required for clinical utility (Svensson et al., 2018; Papadopoulos et al., 2017). Ablation studies supporting our choice of α are in Appendix H.

Average Length (Efficiency): The mean size of the prediction interval. Shorter intervals are preferred, provided they satisfy coverage guarantees.

Cascade Ratio (Adaptivity): To quantify the framework's ability to differentiate between "safe" and "ambiguous" cases, we introduce the Cascade Ratio (CR). To ensure a consistent comparison across experiments with varying granularity K, we define CR as the ratio of the average interval length in the highest uncertainty partition (top $1/K$-th) to that of the lowest uncertainty partition (bottom $1/K$-th). For instance, $K = 3$ compares the top 33% against the bottom 33%. For continuous CASCADE, we evaluate the metric on the subset of patients falling into these same uncertainty quantiles to maintain comparability with the discrete baselines.

$$\text{CR} = \frac{\mathbb{E}[\text{len}(\hat{C}(x)) \mid u_{\text{VA}}(x) \geq Q_{1-1/K}]}{\mathbb{E}[\text{len}(\hat{C}(x)) \mid u_{\text{VA}}(x) \leq Q_{1/K}]}, \tag{7}$$

where Q_τ denotes the τ-th empirical quantile of the uncertainty scores in the test set. A higher CR indicates superior conditional adaptivity, implying the model correctly expands intervals for difficult decisions while narrowing them for clear ones.

4.2. Statistical Validation

To verify the observed adaptivity represents a genuine improvement rather than random variance, we perform a suite of normative statistical tests:

Kolmogorov-Smirnov (KS) Test: We utilize it to quantify the distributional divergence between the interval lengths of CASCADE and the Standard CP baseline. A significant statistic (D) confirms that our scaling mechanism successfully alters the interval structure, producing a statistically distinct distribution of interval lengths driven by uncertainty rather than the static profile of the baseline.

Spearman Correlation: To confirm that the interval length is driven by the uncertainty scores, we evaluate the rank correlation (ρ) between the interval length and the Venn-Abers score (u_{VA}). A strong positive correlation validates that prediction intervals scale with model uncertainty.

Winkler Score: Computed to evaluate the joint trade-off between coverage and interval length (lower is better). This proper scoring rule penalizes wide intervals and applies a penalty for non-coverage (true value is outside the interval),

providing a single summary statistic for model utility that balances safety with efficiency.

5. Results

5.1. Experimental Setup

For both Stage 1 classification and Stage 2 regression we employ XGBoost models (Chen & Guestrin, 2016) optimized via Bayesian Optimization (Wu et al., 2019). Code is available at `https://github.com/rdiazrincon/cascade_conformal_pd`. Detailed hyperparameters and training configurations are provided in Appendix B. We evaluate on a cohort of 631 inpatient admissions from UF Health. Further details about patient cohort and medications can be found in Appendix A. Additional experiments with different models (Logistic/Linear Regression) can be found in Appendix G.

5.2. Overall Performance Comparison

We evaluate the performance of CASCADE against four conformal benchmarks (Naïve, Standard CP, CV+, J+aB) and the discrete Mondrian baseline. Table 1 summarizes the comparative performance.

Baseline Selection. As shown in Table 1, both Naïve and J+aB fail to achieve nominal coverage (80%), rendering them unsuitable for reliable clinical decision-making. Among the valid methods, Standard CP and CV+ exhibit comparable performance. Given this parity, we select Standard CP as the primary baseline. This choice reflects its status as the established standard for split conformal prediction. By benchmarking against this fundamental method, we can most clearly isolate the impact of our adaptive scaling mechanism, evaluating its efficiency gains directly against the standard methods used in practice.

The Structural Limitation of Standard Baselines. By mathematical design, standard conformal methods assume homoscedasticity, applying a global non-conformity threshold across the entire feature space. Consequently, they structurally default to a Cascade Ratio near 1.0, as their interval lengths cannot respond to local epistemic risk. While this guarantees uniform interval sizes, the CR metric exposes the clinical danger of this rigidity: these standard models are forced to offer identical safety buffers to both straightforward and highly ambiguous patient profiles.

While mathematically valid, this results in over-coverage (84.0% [0.801–0.879]), revealing that standard baselines achieve safety through rigid, global conservatism rather than patient-level precision. CASCADE, by dynamically routing uncertainty, calibrates tightly to the 80.1% target. The variance in CASCADE's interval length ([0.135–0.162]) is not statistical noise but the intentional result of uncertainty-driven scaling, providing optimized precision for the patients that need it most.

5.3. Mechanism of Adaptivity: Group-wise Analysis

Global metrics often mask significant performance disparities between easy (low uncertainty) and difficult (high uncertainty) cases. While strict conditional coverage is theoretically ideal, recent work suggests it can yield overly conservative intervals in practice due to data scarcity in high-risk regions (Ding et al., 2023). To explicitly operationalize the trade-off between routine clinical precision and high-stakes safety, we adopted a stratified approach, grouping patients into tertiles based on their Venn-Abers uncertainty scores (u_{VA}). This distinction allows us to decouple performance for low and high uncertainty patients (bottom 33% and top 33% respectively).

Table 2 contrasts the behavior of Standard CP (fixed length 0.113) against the adaptive response of CASCADE ($\beta = 0.7$), revealing the core cascade effect. For **low-uncertainty** patients, CASCADE identifies the standard length as unnecessarily conservative and automatically tightens to 0.069 (-38.9%), offering clinicians more precise guidance for confident adjustments. For **high-uncertainty** patients (top 33%), CASCADE creates a dynamic safety net by expanding to 0.292 (+158.9%), simultaneously increasing coverage from 85.4% to 91.7% and signaling that ambiguous cases warrant additional caution before adjusting medication.

5.4. Quantitative Validation of Adaptivity

To validate the statistical significance of the proposed framework and verify the drivers of interval adaptivity, we employ the following validation tests:

Distributional Divergence (KS Test): The KS test yielded a statistic of $D = 0.62$ ($p < 10^{-54}$), rejecting the null hypothesis that CASCADE and the Standard CP baseline produce equivalent interval distributions. This confirms that CASCADE fundamentally alters the prediction structure, shifting from a static safety margin to a dynamic distribution.

Driver of Adaptivity (Spearman Correlation): To verify that this distinct behavior is driven by the uncertainty scores, we evaluate the correlation between interval length and the Venn-Abers score (u_{VA}). CASCADE demonstrates a near-perfect Spearman correlation of $\rho = 0.999$ ($p < 0.001$), validating that prediction intervals are directly and monotonically responsive to the detected epistemic risk.

Global Utility (Winkler Score): The adaptivity of CASCADE introduces an expected trade-off in global efficiency. The framework achieved a Winkler score of 0.392, compared to the baseline's 0.282. This increase in the aggregate penalty reflects CASCADE's interval length adaptivity to

Table 1. Overall Performance Comparison. Continuous CASCADE achieves the optimal balance between coverage and adaptivity (CR 4.23). Standard baselines produce fixed-length intervals, resulting in no conditional adaptivity (CR 1.00).

Method	Marg. Cov.[‡]	Length[‡]	CR[†‡]
Standard Baselines			
Naïve	52.5% [0.469–0.583]	0.031 [0.031–0.031]	1.00 [1.00–1.00]
Standard CP	84.0% [0.801–0.879]	0.113 [0.113–0.113]	1.00 [1.00–1.00]
CV+	83.5% [0.792–0.876]	0.100 [0.096–0.105]	1.06 [0.94–1.20]
J+aB	60.6% [0.547–0.658]	0.132 [0.128–0.137]	0.97 [0.89–1.04]
CASCADE Framework			
Mondrian ($K = 3$)	86.5% [0.824–0.902]	0.118 [0.113–0.123]	2.02 [2.02–2.02]
Cont. CASCADE ($\beta = 0.7$)	**80.1%** [0.756–0.844]	**0.148** [0.135–0.162]	**4.23** [3.82–4.70]

[†]CR: avg length (Top 33% Unc.) / (Bottom 33% Unc.) [‡]95% bootstrap confidence intervals

Table 2. Stratified Performance Analysis. Comparisons are made against the Standard CP baseline length (0.113). CASCADE dynamically reallocates this budget: it tightens intervals for confident cases (Low) by 38.9% while creating a safety buffer for the high-risk patients (Top 33%).

Uncertainty	Method	Cov.	Length	Δ Length[†]
Low	Standard CP	81.1%	0.113	—
	Cascade	**69.7%**	**0.069**	**−38.9%**
Medium	Standard CP	86.5%	0.113	—
	Cascade	**82.0%**	**0.100**	**−10.9%**
High	Standard CP	85.4%	0.113	—
	Cascade	**91.7%**	**0.292**	**+158.9%**

[†]Relative to Standard CP Baseline Length (0.113).

high-risk, uncertain predictions. While the baseline maintains tighter average intervals by ignoring local uncertainty, CASCADE prioritizes clinical safety by explicitly inflating bounds where epistemic risk is high, providing the essential adaptive signaling required for safe downstream clinical intervention.

A detailed analysis of Mondrian sensitivity across $K \in \{3, 5, 7\}$ stratification bins, and the corresponding robustness of Continuous CASCADE, is provided in Appendix D.

5.5. Beta Ablation

We perform a comprehensive ablation study on the scaling parameter to characterize the trade-off between coverage stability and interval length. Figure 3 visualizes this dynamic, while quantitative results are detailed in Table 8 in Appendix E.

Our analysis identifies $\beta = 0.7$ as the optimal value. At this threshold, the model maximizes adaptivity (Cascade Ratio 4.23) while strictly maintaining valid marginal coverage (80.1%).

Lower values ($\beta \leq 0.5$) provide valid coverage but reduced adaptive resolution (CR < 3.0). Conversely, aggressive scaling ($\beta \in [0.9, 1.0]$) violates the coverage guarantee.

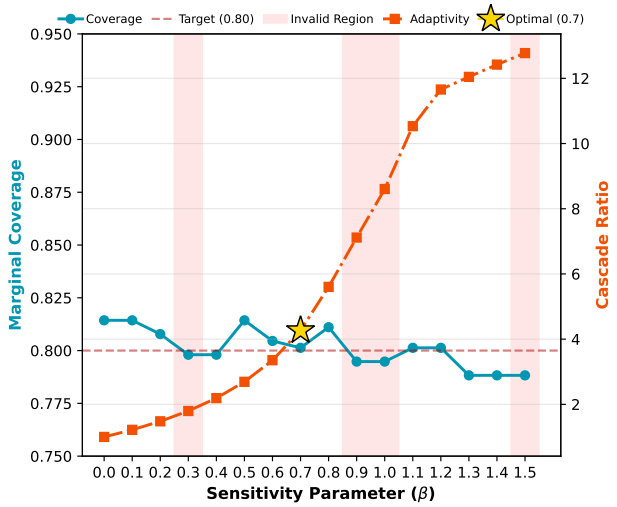

Figure 3. β Ablation Study. The trade-off between safety and adaptivity. The dark red line denotes nominal coverage (target 80%), while the orange line denotes the Cascade Ratio (adaptivity). The optimal parameter $\beta = 0.7$ (gold star) maximizes adaptivity (CR 4.23) while strictly maintaining adequate coverage (80.1%). Lower values are inefficient while higher values violate coverage.

While extreme scaling ($\beta = 1.2$) restores validity, it results in an average interval length of 0.267 (more than double the baseline length) rendering the predictions inefficient for clinical utility. Thus, $\beta = 0.7$ represents the maximal adaptivity achievable within the safety constraints of the task.

6. Discussion

Uncertainty Propagation in Two-Stage Systems. Two-stage clinical frameworks successfully decouple intervention assessment from dosage prediction, but they typically treat the subsequent regression task as statistically independent. Our results challenge this compartmentalization: the epistemic uncertainty generated during the assessment phase is a vital signal that should not be discarded. CASCADE

enables the downstream predictor to dynamically adapt its intervals based on upstream ambiguity, ensuring that the final recommendation reflects the cumulative uncertainty of the entire pipeline.

The Stability/Adaptivity Tradeoff. A common critique of adaptive conformal methods is the loss of coverage stability across subgroups. While Standard CP achieves uniform coverage (84%) across all risk profiles, our analysis suggests this "stability" results in unnecessarily wide intervals. As Ding et al. (2023) warns, strict conditional coverage can lead to overly conservative predictions in data-scarce regions. Instead, CASCADE adopts local adaptivity. By allowing the interval length to fluctuate, we accept a marginal coverage adjustment in the low-risk majority (69.7%) to secure a significant reduction in interval length (38.9%). While this targeted local adaptivity inherently results in a higher global Winkler score compared to baseline, this trade-off prioritizes high-stakes clinical utility over theoretical uniformity.

Limits of Discrete Stratification. Our analysis reveals a fundamental tension between granularity and statistical power in discrete frameworks. While Mondrian CP offers intuitive group-conditional guarantees, it suffers from sample size fragmentation. As shown in Appendix D, increasing the bin count to $K = 7$ achieved a higher Cascade Ratio (4.63) but caused Mondrian's average interval length to inflate to 0.170 (a 44% increase compared to 0.118 at $K = 3$) due to reduced effective sample size per bin (N_{cal}/K). In contrast, Continuous CASCADE decouples uncertainty scaling from quantile estimation, maintaining strong adaptivity (CR 6.83 at $K = 7$) without the fragmentation penalty.

Clinical Implications. The lack of adaptivity in Standard CP (CR 1.00) poses a latent safety risk by presenting identical interval lengths for both clear-cut and ambiguous cases. This false sense of certainty can lead to dangerous over-reliance on model outputs and unfortunate patient outcomes. The CASCADE framework addresses this by acting as an automated safeguard: for high-uncertainty patients, it expands the prediction interval and explicitly signals to the neurologist that the case warrants caution, prompting investigation of confounding factors (e.g., protein intake or sleep quality) before adjusting medication. This aligns model behavior with the principle of *primum non nocere* (first, do no harm). Crucially, CASCADE does not achieve this safety by inflating intervals globally. Instead, it employs a targeted mechanism, utilizing optimized prediction intervals to preserve high precision for the majority of confident cases while reserving adaptive safety buffers for highly ambiguous cases.

Beyond Parkinson's Disease. While validated on PD medication adjustment, CASCADE is fundamentally domain-agnostic. It addresses the reliability gap inherent to any two-stage pipeline, where a discrete intervention assessment precedes a continuous quantification task. Examples include Deep Brain Stimulation, where discrete selection of the surgical target (globus pallidus interna vs. subthalamic nucleus) precedes fine-tuning of voltage and frequency (Ramirez-Zamora & Ostrem, 2018). Similarly, in dystonia, identifying affected muscle groups is a prerequisite for calculating botulinum toxin dosage (Dressler et al., 2021). By propagating the epistemic uncertainty from the initial assessment to the downstream prediction, CASCADE offers a scalable mechanism for building trustworthy, uncertainty-aware AI systems across high-stakes medical domains.

7. Limitations and Future Work

Our framework employs symmetric scaling, reflecting the clinical reality that both under and over-medication are detrimental. We aim to explore asymmetric scaling for profiles where one risk strictly outweighs the other, and to add a formal abstention mechanism enabling the model to defer highly ambiguous cases to human experts. Finally, we aim to validate CASCADE across the PD continuum, ensuring robustness against the longitudinal physiological shifts inherent to disease progression.

8. Conclusion

We present CASCADE, a novel uncertainty quantification framework that resolves the tension between classification precision and regression reliability. By leveraging Venn-Abers uncertainty to continuously scale prediction intervals, CASCADE helps overcome the "one-size-fits-all" nature of PD medication management. By dynamically calibrating the prediction interval to the upstream confidence signal, CASCADE achieves a four-fold improvement in adaptivity (CR 4.23 vs 1.00) while maintaining coverage guarantees. This ensures that AI-driven dosage recommendations are not only accurate, but contextually aware: providing optimized prediction intervals for straightforward patients and a necessary adaptive safety buffer for complex ones. Ultimately, CASCADE demonstrates that adaptivity is not just a technical enhancement but a clinical necessity.

Impact Statement

CASCADE is designed to surface uncertainty rather than suppress it, ensuring clinicians receive explicit warning signals before acting on ambiguous predictions. As with all clinical AI, prospective validation and regulatory review are necessary before deployment. There are many potential societal consequences of our work, none which we feel must be specifically highlighted beyond this note.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint*, 7 2021. URL http://arxiv.org/abs/2107.07511.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16:494–591, 2023. ISSN 1935-8237. doi: 10.1561/2200000101.

Armstrong, M. J. and Okun, M. S. Diagnosis and treatment of parkinson disease: A review. *JAMA - Journal of the American Medical Association*, 323:548–560, 2 2020. ISSN 15383598. doi: 10.1001/jama.2019.22360.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49:486–507, 2021.

Belić, M., Bobić, V., Badža, M., Šolaja, N., Đurić Jovičić, M., and Kostić, V. S. Artificial intelligence for assisting diagnostics and assessment of parkinson's disease—a review. *Clinical Neurology and Neurosurgery*, 184, 9 2019. ISSN 18726968. doi: 10.1016/j.clineuro.2019.105442.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016, pp. 785–794. Association for Computing Machinery, 8 2016. ISBN 9781450342322. doi: 10.1145/2939672.2939785.

Diaz-Rincon, R., Khoshbouei, H., and Shickel, B. Ai for personalized medication management in parkinson's disease (pd). In *MOVEMENT DISORDERS*, volume 39, pp. S48–S49. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2024.

Diaz-Rincon, R., Liang, M., Ramirez-Zamora, A., and Shickel, B. Uncertainty-aware prediction of parkinson's disease medication needs: A two-stage conformal prediction approach. In *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298 of *Proceedings of Machine Learning Research*. PMLR, 15–16 Aug 2025a. URL https://proceedings.mlr.press/v298/diaz-rincon25a.html.

Diaz-Rincon, R., Liang, M., Ramirez-Zamora, A., and Shickel, B. Improved decision-making for in-hospital medication management in parkinson's disease. In *MOVEMENT DISORDERS*, volume 40, pp. S23–S24. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2025b.

Ding, T., Angelopoulos, A., Bates, S., Jordan, M., and Tibshirani, R. J. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Dressler, D., Altavista, M. C., Altenmueller, E., Bhidayasiri, R., Bohlega, S., Chana, P., Chung, T. M., Colosimo, C., Fheodoroff, K., Garcia-Ruiz, P. J., et al. Consensus guidelines for botulinum toxin therapy: general algorithms and dosing tables for dystonia and spasticity. *Journal of Neural Transmission*, 128(3):321–335, 2021.

Espay, A. J., Bonato, P., Nahab, F. B., Maetzler, W., Dean, J. M., Klucken, J., Eskofier, B. M., Merola, A., Horak, F., Lang, A. E., Reilmann, R., Giuffrida, J., Nieuwboer, A., Horne, M., Little, M. A., Litvan, I., Simuni, T., Dorsey, E. R., Burack, M. A., Kubota, K., Kamondi, A., Godinho, C., Daneault, J. F., Mitsi, G., Krinke, L., Hausdorff, J. M., Bloem, B. R., and Papapetropoulos, S. Technology in parkinson's disease: Challenges and opportunities. *Movement Disorders*, 31:1272–1282, 9 2016. ISSN 15318257. doi: 10.1002/mds.26642.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Jenner, P. Molecular mechanisms of l-dopa-induced dyskinesia, 9 2008. ISSN 1471003X.

Kim, B., Xu, C., and Barber, R. Predictive inference is free with the jackknife+-after-bootstrap. *Advances in Neural Information Processing Systems*, 33:4138–4149, 2020.

Kim, Y., Suescun, J., Schiess, M. C., and Jiang, X. Computational medication regimen for parkinson's disease using reinforcement learning. *Scientific Reports*, 11, 12 2021. ISSN 20452322. doi: 10.1038/s41598-021-88619-4.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Newman, E. J., Grosset, D. G., and Kennedy, P. G. The parkinsonism-hyperpyrexia syndrome. *Neurocritical Care*, 10(1):136–140, 2009.

Ng, A. Y. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. Technical report, 2004.

Olanow, C. W., Obeso, J. A., and Stocchi, F. Continuous dopamine-receptor treatment of parkinson's disease: scientific rationale and clinical implications. *The Lancet Neurology*, 5:677–687, 2006.

Papadopoulos, H., Vovk, V., and Gammerman, A. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

Papadopoulos, H., Kyriacou, E., and Nicolaides, A. Unbiased confidence measures for stroke risk estimation based on ultrasound carotid image analysis. *Neural Computing and Applications*, 28:1209–1223, 6 2017. ISSN 09410643. doi: 10.1007/s00521-016-2590-3.

Ramirez-Zamora, A. and Ostrem, J. L. Globus pallidus interna or subthalamic nucleus deep brain stimulation for parkinson disease: a review. *JAMA neurology*, 75(3): 367–372, 2018.

Salmanpour, M. R., Hosseinzadeh, M., Bakhtiyari, M., Maghsudi, M., and Rahmim, A. Prediction of drug amount in parkinson's disease using hybrid machine learning systems and radiomics features. *International Journal of Imaging Systems and Technology*, 33:1437–1449, 7 2023. ISSN 10981098. doi: 10.1002/ima.22868.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. Technical report, 2008.

Shamir, R. R., Dolber, T., Noecker, A. M., Walter, B. L., and McIntyre, C. C. Machine learning approach to optimizing combined stimulation and medication therapies for parkinson's disease. *Brain Stimulation*, 8:1025–1032, 11 2015. ISSN 18764754. doi: 10.1016/j.brs.2015.06.003.

Svensson, F., Aniceto, N., Norinder, U., Cortes-Ciriano, I., Spjuth, O., Carlsson, L., and Bender, A. Conformal regression for quantitative structure-activity relationship modeling - quantifying prediction uncertainty. *Journal of Chemical Information and Modeling*, 58:1132–1140, 5 2018. ISSN 1549960X. doi: 10.1021/acs.jcim.8b00054.

Taylor, J., Anderson, W. S., Brandt, J., Mari, Z., and Pontone, G. M. Neuropsychiatric complications of parkinson disease treatments: importance of multidisciplinary care. *The American Journal of Geriatric Psychiatry*, 24(12): 1171–1180, 2016.

Tomlinson, C. L., Stowe, R., Patel, S., Rick, C., Gray, R., and Clarke, C. E. Systematic review of levodopa dose equivalency reporting in parkinson's disease. *Movement Disorders*, 25:2649–2653, 11 2010. ISSN 08853185. doi: 10.1002/mds.23429.

Vovk, V. and Petej, I. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 2005.

Watts, J., Khojandi, A., Vasudevan, R., Nahab, F. B., and Ramdhani, R. A. Improving medication regimen recommendation for parkinson's disease using sensor technology. *Sensors*, 21, 5 2021. ISSN 14248220. doi: 10.3390/s21103553.

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., and Deng, S. H. Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, 17:26–40, 3 2019. ISSN 2666223X. doi: 10.11989/JEST.1674-862X.80904120.

Ying, X. An overview of overfitting and its solutions. In *Journal of Physics: Conference Series*, volume 1168. Institute of Physics Publishing, 3 2019. doi: 10.1088/1742-6596/1168/2/022022.

A. Patient Cohort and Medication Details

A.1. Demographics

Table 3 details the demographic characteristics of the study cohort ($N = 631$). The population is predominantly elderly (Mean Age 78.0) and male (64.5%), consistent with Parkinson's Disease epidemiology.

Table 3. Demographic Characteristics of the Study Cohort.

Characteristic	Category	Number	Percentage (%)
Total Patients		631	100.00
Mean Age, Years (SD)		77.98 (10.59)	
Gender	Male	407	64.50
	Female	224	35.50
Race	White	526	83.62
	Black	63	10.01
	Other	34	5.41
	Asian	5	0.79
	Multiracial	1	0.16
Ethnicity	Not Hispanic	598	95.07
	Hispanic	31	4.93

A.2. Medications

Table 4 lists the antiparkinsonian medications administered in our patient cohort.

Table 4. Classification of antiparkinsonian medications administered during the study.

Drug Class	Medications
Levodopa Formulations	Carbidopa-Levodopa, Carbidopa-Levodopa ER, Carbidopa-Levodopa-Entacapone, Duopa
Dopamine Agonists	Pramipexole, Pramipexole ER, Ropinirole, Rotigotine, Apomorphine, Bromocriptine, Cabergoline
MAO-B Inhibitors	Rasagiline, Selegiline, Safinamide
COMT Inhibitors	Entacapone, Tolcapone
Anticholinergics	Benztropine, Trihexyphenidyl
NMDA Antagonists	Amantadine, Amantadine ER
Adenosine Antagonists	Istradefylline

B. Model Implementation and Conformal Framework

B.1. Underlying Predictors (XGBoost)

To ensure that the uncertainty quantification results reflect the framework's performance rather than model instability, we utilized strong underlying predictors. For both classification and regression, we employed XGBoost models (Chen & Guestrin, 2016) with hyperparameters optimized via GridSearchCV and Bayesian Optimization (Wu et al., 2019).

Key hyperparameters included a learning rate of 0.1, max depth of 7, and 700 boosting rounds. We applied L1/L2 regularization (Ng, 2004) (alpha: 0.1, lambda: 1) and early stopping to prevent overfitting (Ying, 2019). This configuration demonstrated strong predictive capabilities, achieving an AUC of 0.96 for the classification task and high precision for the dosage prediction ($R^2 = 0.9801$, RMSE$= 0.0995$, MAE$= 0.0380$).

B.2. Conformal Prediction Baselines

We compared the proposed method against four standard benchmarks.

Standard Conformal Prediction: We employ the Standard Conformal Prediction framework. The available training data is split into a proper training set $\mathcal{D}_{\text{train}}$ and a calibration set $\mathcal{D}_{\text{cal}}$. The model is fitted on $\mathcal{D}_{\text{train}}$, and non-conformity scores $s_i = |y_i - \hat{\mu}(x_i)|$ are computed on $\mathcal{D}_{\text{cal}}$. The prediction interval is constructed as $\hat{\mu}(x_{\text{new}}) \pm \hat{q}_{1-\alpha}$, where $\hat{q}_{1-\alpha}$ is the $(1-\alpha)$ quantile of the calibration scores.

Naïve: This method constructs fixed-length intervals based solely on the empirical distribution of residuals. The prediction interval is defined as $\hat{\mu}(x_{\text{new}}) \pm q_{1-\alpha}$, where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of the absolute errors on the calibration set. Unlike conformal prediction, the Naïve approach lacks finite-sample coverage guarantees and does not account for the statistical variability introduced by the calibration step.

Cross Validation Conformal (CV+): This method employs K-fold cross-validation ($K = 10$) to leverage multiple calibration sets. While CV+ provides a theoretical coverage guarantee of $1 - 2\alpha$, it is computationally expensive (K model fits) and often produces conservative intervals (Barber et al., 2021).

Jackknife+ After Bootstrap (J+aB): Designed specifically for ensemble methods. It reuses the bootstrap samples from the ensemble training to create leave-one-out predictions. J+aB provides intervals "for free" with ensemble methods but may exhibit instability in small datasets (Kim et al., 2020).

C. Extended Mondrian Results

We provide detailed performance metrics for the discrete stratification approach (Mondrian Conformal Prediction) discussed in the main text. Table 5 presents the coverage, interval length, and adaptivity signal for Mondrian ($K = 3$).

Table 5. Performance Metrics for Mondrian Conformal Prediction ($K = 3$). The method stratifies patients into three discrete bins based on uncertainty. While coverage is valid, the adaptivity signal is coarse compared to the continuous approach.

Uncertainty	Coverage	Length	Adaptivity
Low	80.3%	0.090	Moderate
Medium	88.8%	0.090	Moderate
High	92.7%	0.181	High
Overall Cascade Ratio (CR)			2.02

D. Impact of Stratification Granularity

To ensure our conclusions are not artifacts of the chosen discretization level ($K = 3$), we evaluate the impact of increasing the number of uncertainty strata ($K \in \{5, 7\}$). Table 6 presents the performance of Mondrian as the number of bins increases. Moving from $K = 3$ to $K = 5$, the model maintains a stable Cascade Ratio (CR) while holding coverage relatively close to the target guarantee. However, as K is pushed to 7, the framework exhibits a sharp spike in adaptivity (CR 4.63). This demonstrates the framework's underlying capacity to scale intervals for highly uncertain cohorts when given higher resolution.

Table 6. Mondrian Sensitivity. As the number of strata (K) increases, the framework demonstrates an increased capacity for adaptivity, peaking at a CR of 4.63 at $K = 7$.

Configuration	Cov. (%)	Length	CR
Mondrian ($K = 3$)	86.6%	0.118	2.02
Mondrian ($K = 5$)	84.0%	0.121	1.98
Mondrian ($K = 7$)	85.3%	0.170	4.63

D.1. Continuous CASCADE (Evaluation Robustness)

To verify that our efficiency gains are not sensitive to the choice of evaluation bins, we analyze the Cascade Ratio across varying granularities (from tertiles to septiles). As shown in Table 7, the performance remains stable across all resolutions. While we report results for $K = 3$ in the main text to align with standard clinical risk strata (Low, Medium, High), CASCADE maintains high adaptivity even at finer resolutions. Notably, even at the finest granularity ($K = 7$), the model

maintains a Cascade Ratio of 6.83, confirming that the uncertainty scaling is a fundamental property of the signal rather than an artifact of discretization.

Table 7. Continuous CASCADE Robustness. The adaptivity signal remains strong across different evaluation resolutions, consistently outperforming the discrete baseline.

Evaluation Split	Comparison	CR
Tertiles ($K = 3$)	Top 33% vs Bottom 33%	4.23
Quintiles ($K = 5$)	Top 20% vs Bottom 20%	5.85
Septiles ($K = 7$)	Top 14% vs Bottom 14%	6.83

E. Extended Beta Ablation

To validate the stability of the CASCADE framework, we performed a fine-grained ablation study on the scaling parameter β. The range $\beta \in [0.0, 1.5]$ was chosen to encompass the full theoretical spectrum: from $\beta = 0$ (homoscedastic assumption, effectively Standard CP) through $\beta = 1$ (linear scaling proportional to uncertainty), up to $\beta = 1.5$ (aggressive supra-linear penalization). Table 8 shows these results.

Table 8. Full β Ablation. We vary the scaling factor $\beta \in [0.0, 1.5]$. The selected value $\beta = 0.7$ (bold) is chosen by the Cascade Ratio (CR), which quantifies the adaptivity gain relative to the base interval length. $\beta = 0.7$ achieves the optimal CR (4.23) while maintaining valid marginal coverage. Values of $\beta \geq 0.9$ result in coverage collapse for the low-uncertainty cohort.

	Low Unc.		Med. Unc.		High Unc.			
β	Cov.	Len.	Cov.	Len.	Cov.	Len.	CR	Status
0.0	80.3%	0.128	83.1%	0.128	81.2%	0.128	1.00	No Adaptivity
0.1	77.9%	0.112	83.1%	0.116	84.4%	0.137	1.22	Minimal
0.2	77.0%	0.099	82.0%	0.106	84.4%	0.147	1.49	Mild
0.3	74.6%	0.087	82.0%	0.097	84.4%	0.158	1.81	Conservative
0.4	73.0%	0.086	82.0%	0.100	86.5%	0.189	2.21	Conservative
0.5	73.8%	0.084	82.0%	0.105	90.6%	0.228	2.71	Valid
0.6	71.3%	0.077	82.0%	0.103	90.6%	0.258	3.36	Valid
0.7	**69.7%**	**0.069**	**82.0%**	**0.100**	**91.7%**	**0.292**	**4.23**	**Optimal**
0.8	70.5%	0.067	82.0%	0.110	93.8%	0.369	5.49	Borderline
0.9	65.6%	0.060	83.1%	0.108	93.8%	0.416	6.93	Unsafe
1.0	63.1%	0.056	83.1%	0.105	96.9%	0.469	8.44	Unsafe
1.1	63.9%	0.057	83.1%	0.115	97.9%	0.591	10.45	Unsafe
1.2	63.9%	0.056	83.1%	0.113	97.9%	0.679	12.01	Unsafe
1.3	63.9%	0.058	78.7%	0.106	97.9%	0.733	12.56	Unsafe
1.4	63.9%	0.058	78.7%	0.098	97.9%	0.757	12.96	Unsafe
1.5	63.9%	0.058	78.7%	0.092	97.9%	0.779	13.34	Unsafe

F. Model Performance on Predicted Patients

To demonstrate the robustness of our approach, we perform additional tests for patients where $\hat{y} \neq 0$, meaning patients where the classifier predicts a change in medications. Optimal threshold was calculated using Youden's J. Results are summarized in Table 9.

Table 9. While evaluating on predicted patients, Continuous CASCADE achieves perfect marginal coverage while delivering the highest Cascade Ratio (1.43) compared to Baseline and Mondrian.

Method	Cov.	Len.	CR	β
Baseline	79.4%	0.662	1.00	—
Mondrian	82.1%	0.625	0.83	—
Cont. CASCADE	80%	0.6836	1.43	0.2

G. Alternative Underlying Models

We perform additional experiments using Logistic Regression (Stage 1) and Linear Regression for Stage 2 ($R^2 = 0.08$). As expected, the high underlying model error caused global interval lengths to expand. While Mondrian's binning mechanism failed to calibrate under this noise (72.3% coverage), CASCADE maintained strict marginal validity (80.1% coverage) at the cost of longer intervals. Furthermore, CASCADE successfully achieved a higher CR (2.56), compared to baseline.

Table 10. Although with expanded intervals due to model errors, CASCADE achieves the highest CR (2.56).

Method	Cov.	Len.	CR	β
Baseline	78.5%	1.649	1.00	—
Mondrian	72.3%	1.721	1.07	—
Cont. CASCADE	80.13%	2.267	2.56	0.4

H. Alpha Ablation

We perform alpha ablation ($\alpha \in \{0.3, 0.1, 0.05\}$) to choose the most appropriate value for clinical utility. Cont. CASCADE consistently satisfies marginal coverage across all thresholds (e.g., 90.9% coverage at $\alpha = 0.1$; 95.4% at $\alpha = 0.05$), maintaining tighter calibration and narrower intervals than Mondrian and baselines. We focus on $\alpha = 0.2$ as it provides optimal clinical utility in PD medication management. While ($\alpha = 0.05$) preserves adaptivity (CR = 4.23), its intervals become too wide to guide precise LEDD titration.

Table 11. α Ablation.

α	Method	Cov	Len	CR	β
0.30	Baseline	70.4%	0.060	1.00	—
	Mondrian	75.9%	0.058	1.39	—
	Cont. CASCADE	70.0%	0.068	1.81	0.3
0.20	Baseline	84.0%	0.112	1.00	—
	Mondrian	86.6%	0.118	2.02	—
	Cont. CASCADE	80.1%	0.148	4.23	0.7
0.10	Baseline	93.5%	0.258	1.00	—
	Mondrian	96.1%	0.376	2.47	—
	Cont. CASCADE	90.9%	0.260	2.71	0.5
0.05	Baseline	97.7%	0.492	1.00	—
	Mondrian	98.0%	0.922	2.47	—
	Cont. CASCADE	95.4%	0.444	4.23	0.7