COVERAGE ALLOCATION IN TWO-STAGE CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction (CP) provides marginal coverage guarantees under exchangeability and is generally evaluated through mean coverage and interval length. In clinical settings, mean coverage says nothing about whether coverage is balanced across patient groups. One way to address this is by adapting interval length to prediction difficulty, but stronger adaptation does not necessarily lead to better intervals, and can leave patient subpopulations under-covered. We introduce CASCADE-CRC, a framework for studying how adaptation strength affects coverage and interval length reallocation in two-stage conformal prediction. We evaluate on inpatient and outpatient Parkinson's disease cohorts, an external zero-inflated insurance dataset, and synthetic data with known heteroscedasticity. Our experiments reveal that adaptation strength reallocates interval length and, with it, coverage across groups, reducing the coverage gap (CovGap) by 31.7% on the outpatient cohort while keeping 80% marginal coverage, with benefits that vary by setting. Our findings emphasize the importance of adaptation strength as a modeling choice, motivating joint evaluation of group coverage and interval length.

1 INTRODUCTION

Parkinson's disease (PD) medication management is highly individualized and levodopa remains the gold standard medication (Armstrong & Okun, 2020). While there are different forms of dopamine replacement therapies, they are commonly standardized as a levodopa-equivalent daily dose (LEDD) (Tomlinson et al., 2010). However, prolonged exposure, alongside patient severity and risk profiles can pose diverse challenges: under-medication provides poor symptom control, with extended periods of OFF-time, while over-medication can cause debilitating side effects like levodopa-induced dyskinesia (LID) (Olanow et al., 2006; Jenner, 2008), orthostatic hypotension (Senard et al., 1997), and hallucinations (Taylor et al., 2016). The decision to adjust medications follows a two-stage setting: a neurologist first assesses if a patient warrants a medication change, then increases or decreases the dose as appropriate.

Recent advances in PD medication management resemble this approach, utilizing conformal prediction (CP) (Vovk et al., 2005) for uncertainty quantification of medication changes, and addressing uncertainty propagation between the two stages (Diaz-Rincon et al., 2025; 2026). CP provides marginal coverage guarantees for dosage changes. However, it can do so by concealing large allocation disparities, achieving its average by over-covering easy predictions and under-covering challenging ones across different patient strata.

Consider a neurologist using such a system to inform medication decisions. For each patient, the system estimates how hard the dose change is to predict, before it is observed, and groups them as low, medium or high difficulty. The neurologist observes that the model meets its specified marginal coverage, yet examining these groups reveals severe allocation disparities: at the 80% target ($\alpha = 0.2$), patients in the low-difficulty group are covered at 90%, while those in the high-difficulty group are covered at 75%. Marginal coverage and mean interval length alone do not reveal this disparity. Figure 1 illustrates this pattern with two schematic subpopulations.

In PD, such disparities can have direct clinical implications, as medication errors do not carry the same consequences across patients. Differences in clinical phenotype, disease progression, treatment response, and susceptibility to adverse effects mean that a dose adjustment tolerated by one patient may cause severe side effects in another.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

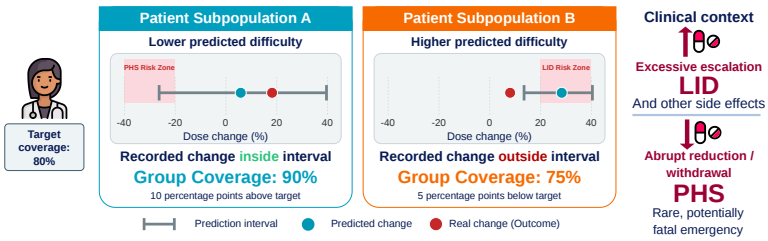

Figure 1: Marginal coverage can hide unequal coverage across patient groups. Intervals cover the recorded change in patient subpopulation A, but miss it in B. Clinical annotations (right side) illustrate potential consequences.

For a patient in the under-covered, high-difficulty group, miscoverage of a predicted dose escalation poses risk of over-medication, potentially contributing to LID and other complications. In the other direction, an under-estimated dose or an unsupported reduction, can lead to abrupt withdrawal of dopaminergic treatment, which may provide poor symptom control and precipitate parkinsonism–hyperpyrexia syndrome (PHS) (Newman et al., 2009), a potentially fatal emergency with high mortality rates, whose risk is greatest in patients on higher daily doses (Ghosh & Liddle, 2011; Beckers et al., 2025).

Thus, excess coverage in one patient group cannot be assumed to clinically compensate for under-coverage in another. Additionally, an interval wide enough to be deemed safe for one patient profile can be inadequate for another. A model can meet its marginal coverage and yet provide less reliable predictions for patients whose treatment decisions carry greater consequences. This motivates evaluating how coverage and interval length are distributed across patient groups.

Several approaches have addressed the limitations of marginal coverage. Exact distribution-free coverage conditional on every covariate is generally impossible to achieve without additional assumptions (Barber et al., 2021); however, guarantees over specified groups are possible (Gibbs et al., 2024), and coverage parity across groups has been studied as a notion of fairness (Romano et al., 2020). Group calibration struggles when groups are small, motivating approaches that pool similar classes (Ding et al., 2023). Other methods attempt to adapt interval length through conformalized quantile regression (CQR) (Romano et al., 2019) or by scaling residuals with an estimated difficulty, as in normalized conformal prediction (Papadopoulos et al., 2011; Lei et al., 2018). Normalized conformal prediction includes a sensitivity constant that moderates how the estimate scales interval length. Its effect on group coverage, however, has not been systematically examined. We show that this choice shifts group coverage by up to 20 percentage points at similar marginal coverage. Thus, we study adaptation strength while retaining pooled calibration, which lets us examine the tradeoffs between coverage and interval length without a separate calibration threshold per evaluation group, and report group coverage as an empirical outcome rather than a guarantee.

We introduce CASCADE-CRC, a framework for evaluating the strength of coverage allocation in two-stage conformal prediction. A single parameter controls how strongly interval length responds to a learned difficulty score, ranging from constant-length to fully scaled intervals, with each level calibrated separately via conformal risk control (CRC) (Angelopoulos et al., 2024). This allows moderate adaptation to reallocate interval length across patients, widening and shrinking intervals based on the difficulty score. Because marginal validity holds across strengths, adaptation serves as a policy lever rather than a constraint, enabling the assessment of tradeoffs between group coverage and interval length.

Our contributions are as follows: (1) We isolate adaptation strength through a controlled comparison holding the predictor, calibration protocol, and evaluation groups fixed, and benchmark against baselines (split conformal, Mondrian, and CQR) using group coverage and interval length. (2) We characterize settings favoring constant, moderate, or full adaptation across clinical, insurance, and synthetic data, emphasizing that the way coverage and interval length are reallocated depends on the dataset, target coverage, and evaluation partition. (3) We show how the preferred strength changes with the difficulty estimator, that classifier probabilities, Venn-Abers estimates, and interval widths add little beyond covariates alone at a given strength in our study, and that retrospectively favorable strengths are not always ones the data support selecting.

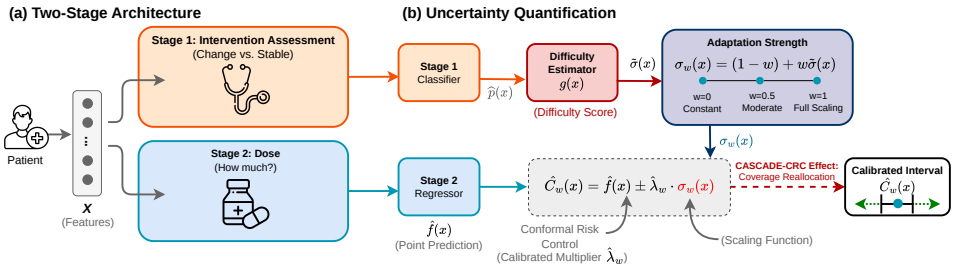

Figure 2: CASCADE-CRC overview. A difficulty estimate controls interval scaling through adaptation strength w. Each fixed strength is calibrated separately to retain marginal coverage for the evaluated population.

2 METHODS

We study how strongly prediction intervals should respond to estimated difficulty. For each fitted difficulty score, we vary adaptation strength while holding the point predictor, calibration observations, and evaluation groups fixed. Each fixed strength receives its own calibration multiplier. Figure 2 summarizes our framework and how adaptation strength controls interval scaling.

2.1 PREDICTION TARGET AND EVALUATION SCOPE

Let i index patients and k their eligible encounters. We predict the change in medication as measured by the LEDD,

$$Y_{ik} = \log\left(\frac{D_{ik}+1}{D_{i,k-1}+1}\right), \tag{1}$$

where D_{ik} is LEDD in mg/day. Positive and negative values indicate increases and decreases respectively. Adding 1 mg/day keeps Y_{ik} defined when a patient starts or stops treatment, which would otherwise divide by zero or take the log of zero. For typical doses, Y_{ik} is close to the usual log dose ratio. Covariates X contain only information available before the dosing decision.

Let $Z = \mathbf{1}\{Y \neq 0\}$ indicate a recorded medication change. For any measurable set $A \subseteq \mathbb{R}$,

$$\mathbb{P}(Y \in A \mid X) = \mathbb{P}(Z = 0 \mid X)\delta_0(A) + \mathbb{P}(Z = 1 \mid X)\mathbb{P}(Y \in A \mid X, Z = 1), \tag{2}$$

where δ_0 is the point mass at zero. The two terms on the right correspond to a classifier for whether a change occurs and a regressor for its size and direction. We evaluate the regressor (second stage) on encounters with a recorded change. The classifier can inform difficulty estimation but does not decide which encounters reach the regressor.

Sampling and data separation. For the clinical cohorts, each calibration and evaluation patient contributes one encounter with a recorded change, sampled uniformly. Let $P_1 = \mathcal{L}(X, Y \mid Z = 1)$ be the resulting distribution. Sampling one encounter per patient makes it reasonable to assume exchangeability. Using every encounter would weight patients by how often they were seen. For the insurance data, each policy is one unit.

To perform conformal calibration, we randomly partition units into a training (50%), calibration $\mathcal{D}_{\text{cal}}$ (30%), and evaluation set (20%). The training set is further divided by unit into a fitting subset (60%), used to train the point predictor $\hat{f}$, and a held-out subset $\mathcal{D}_{\text{hold}}$ (40%) used to fit and normalize the difficulty score. Calibration and evaluation units are used for neither. Below, $j = 1, \ldots, m$ indexes calibration units and (X_*, Y_*) is a future unit.

2.2 DIFFICULTY SCORE

The difficulty score g estimates the size of the prediction error: we regress the absolute residuals $|Y - \hat{f}(X)|$ on X over $\mathcal{D}_{\text{hold}}$, so $g(x)$ estimates their average size at x.

To turn it into a positive, unit-free scale, we floor it and divide by its mean,

$$\tilde{\sigma}(x) = \frac{\max\{g(x), \epsilon\}}{\bar{g}_\epsilon}, \qquad \bar{g}_\epsilon = \frac{1}{|\mathcal{D}_{\text{hold}}|} \sum_{(x,y)\in\mathcal{D}_{\text{hold}}} \max\{g(x), \epsilon\}, \tag{3}$$

where ϵ is the fifth percentile of g on $\mathcal{D}_{\text{hold}}$, but at least 10^{-6}. Flooring before averaging keeps $\tilde{\sigma}$ strictly positive and gives it mean one on $\mathcal{D}_{\text{hold}}$, so values above one mark encounters the model finds harder than average. Because scores can differ in spread even after this rescaling, the same strength can adapt more for one score than another. Section 2.5 considers alternative residual-quantile scores.

2.3 CASCADE-CRC

Adaptation strength. For a fixed $w \in [0, 1]$, the scale is

$$\sigma_w(x) = (1 - w) + w\tilde{\sigma}(x). \tag{4}$$

Equivalently $\sigma_w(x) = 1 + w\{\tilde{\sigma}(x) - 1\}$. Since $\tilde{\sigma}$ has mean one, the scale is centered at one and w controls how far it moves toward the difficulty score. $w = 0$ yields intervals of constant length on the modeled outcome scale, while $w = 1$ uses the full normalized difficulty score (Papadopoulos et al., 2011; Lei et al., 2018). Intermediate values temper the variation between patients, and are a reparametrization of the sensitivity constant in normalized conformal prediction, bounded here to $[0, 1]$. For the clinical target, constant log dose ratio length does not imply a constant length in mg/day because converting back through (1) is multiplicative in the previous dose. The endpoints bound the family we study, not every possible adaptive construction.

Calibration at each strength. We calibrate each strength w separately with conformal risk control (Angelopoulos et al., 2024), using the nested intervals and miscoverage loss

$$C_{w,\lambda}(x) = \big[\hat{f}(x) - \lambda\sigma_w(x),\ \hat{f}(x) + \lambda\sigma_w(x)\big], \qquad L_j^{(w)}(\lambda) = \mathbf{1}\big\{Y_j \notin C_{w,\lambda}(X_j)\big\}, \tag{5}$$

for $\lambda \in [0, \infty]$, with $C_{w,\infty}(x) = \mathbb{R}$. Because larger λ gives wider intervals and the loss is bounded by one, the CRC rule reduces to a quantile of normalized residuals:

$$S_j^{(w)} = \frac{|Y_j - \hat{f}(X_j)|}{\sigma_w(X_j)}, \qquad r = \lceil (m+1)(1-\alpha) \rceil, \qquad \hat{\lambda}_w = \begin{cases} S_{(r)}^{(w)}, & r \le m, \\ +\infty, & r > m, \end{cases} \tag{6}$$

where $S_{(r)}^{(w)}$ is the r-th smallest calibration score and α is the target miscoverage. Every strength uses the same calibration units but computes its own scores and multiplier. The interval at strength w is:

$$\widehat{C}_w(x) = C_{w,\hat{\lambda}_w}(x), \qquad \text{len}_w(x) = 2\hat{\lambda}_w\sigma_w(x). \tag{7}$$

Appendix B provides the general CRC conditions.

Guarantee and scope. For a strength fixed independently of calibration outcomes, exchangeability of the calibration and future units conditional on the fitted training components gives $\mathbb{P}\{Y_* \in \widehat{C}_w(X_*)\} \ge 1-\alpha$ under P_1. The probability averages over calibration data and the future observation, so it is not a guarantee conditional on the calibration sample. Under binary miscoverage, the resulting intervals are the normalized split-conformal construction. The guarantee provides neither coverage within every difficulty stratum nor end-to-end coverage after predicted routing. Selecting w using calibration or evaluation outcomes requires separate justification (Appendix C).

How strength changes allocation. Changing w changes both σ_w and $\hat{\lambda}_w$, so it does not simply widen or narrow every interval. What w controls directly is relative length: for $s_a = \tilde{\sigma}(x_a) > s_b = \tilde{\sigma}(x_b) > 0$ and finite, positive $\hat{\lambda}_w$,

$$\frac{\text{len}_w(x_a)}{\text{len}_w(x_b)} = \frac{1 + w(s_a - 1)}{1 + w(s_b - 1)}. \tag{8}$$

This ratio equals one at $w = 0$ and increases with w. Strength thus sets how length is shared across patients, while recalibration sets its overall level. Within each score we hold the predictor, calibration units and evaluation groups fixed, so differences in results across w reflect strength alone.

2.4 BASELINES

CASCADE. CASCADE (Diaz-Rincon et al., 2026) derives difficulty from the first stage. A Venn-Abers predictor (Vovk & Petej, 2012) returns a pair $[p_0(x), p_1(x)]$, whose width

$$u_{\mathrm{VA}}(x) = p_1(x) - p_0(x) \tag{9}$$

serves as an uncertainty score, scaled as

$$\sigma_{\mathrm{CASCADE}}(x) = 1 + \beta\left(\frac{u_{\mathrm{VA}}(x)}{\bar{u}_{\mathrm{VA}}} - 1\right), \qquad \beta \geq 0, \tag{10}$$

where $\bar{u}_{\mathrm{VA}}$ is the mean width on $\mathcal{D}_{\mathrm{hold}}$ and must be positive. The scale in (10) is strictly positive for $0 \leq \beta < 1$ and nonnegative at $\beta = 1$. We search $\beta \in \{0, 0.1, \ldots, 1\}$ and select it using training data only, so that no calibration or evaluation outcome informs the choice. CASCADE-CRC keeps this scale form but studies strength across a range, with the difficulty score and calibration protocol held fixed. Appendix B provides information regarding implementation bounds, the Venn-Abers guarantees, and selection details.

Other baselines. All methods share the same data splits and evaluation units. Split conformal (Papadopoulos et al., 2002) uses one radius for every patient; its intervals coincide with CASCADE-CRC at $w = 0$. Mondrian (Vovk et al., 2003) calibrates a separate quantile within each difficulty stratum used for evaluation, carrying group-conditional guarantees and returning infinite intervals when a stratum has too few calibration units. CQR (Romano et al., 2019) learns conditional quantile endpoints directly and allows asymmetric intervals.

2.5 ALTERNATIVE DIFFICULTY SCORES

The primary score of Section 2.2 predicts the average size of the residual $|Y - \hat{f}(X)|$. Alongside strength, the difficulty score is a second design choice, so we compare it with six alternatives that instead predict the $(1 - \alpha)$ quantile of the same residual. Because a quantile depends on the target, these six are refitted for each coverage level, whereas the primary score is fitted once. All seven are normalized by (3) and swept over the same strengths. Covariates use X alone; the others add first-stage features to it. Raw probability adds the classifier output $\hat{p}(X)$. VA probability replaces $\hat{p}(X)$ with the Venn-Abers midpoint $v(X) = [p_0(X) + p_1(X)]/2$. VA probability and width adds $u_{\mathrm{VA}}(X)$. VA upper adds only $p_1(X)$, and VA endpoints adds both $p_0(X)$ and $p_1(X)$.

Comparisons among the six vary only the features, whereas comparisons with the primary score also change what is being predicted. Every score and its normalization are computed on $\mathcal{D}_{\mathrm{hold}}$, with first-stage features cross-fitted within it, so no calibration or evaluation outcome enters any score (Appendix B).

3 DATA AND EXPERIMENTAL DESIGN

Datasets. We evaluate on two clinical cohorts, freMTPL2 (a zero-inflated external dataset) (Charpentier, 2014), and synthetic data. The clinical cohorts are the Parkinson's Progression Markers Initiative (PPMI) dataset (Marek et al., 2011), a longitudinal multi-site observational study, and XYZ Health, an inpatient cohort drawn from several hospitals within one health system. Both use the target of (1). freMTPL2, the French motor insurance portfolio, shares the same zero-inflated structure at a higher rate since 96.3% of policies record no claim, compared to 36% to 45% of clinical encounters with no recorded change (Table 10). The frequency of nonzero outcome varies across two-stage problems, with freMTPL2 at the extreme. Therefore, it tests whether our findings hold outside medicine. The synthetic data follow heteroscedastic t_5 noise at a scale we control, so each unit's true difficulty is known (Appendix E). Appendix D describes how each cohort was built and reports unit counts.

Metrics. We divide evaluation units into groups and report each group's coverage and mean interval length. Two summaries describe the whole partition: where cov_g is the coverage in group g of G,

$$\mathrm{CovGap} = \frac{1}{G}\sum_{g=1}^{G}\left|\mathrm{cov}_g - (1-\alpha)\right|, \qquad \mathrm{FSC} = \min_g \mathrm{cov}_g. \tag{11}$$

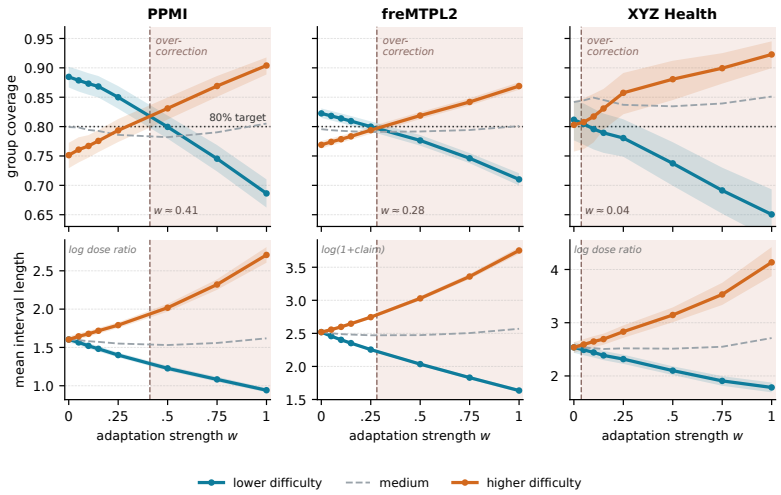

Figure 3: Coverage and length are reallocated across difficulty strata as w varies, at 80% coverage. Curves are equally weighted means over splits of group coverage (top) and mean interval length (bottom), with 95% bootstrap bands. Lengths differ by dataset. The low and high-difficulty strata cross where they receive equal coverage. Beyond the crossing (shaded) the ordering reverses. These curves are not CovGap, which is computed in Appendix Figure 4)

Lower CovGap (Ding et al., 2023) means smaller average absolute deviation across groups. Higher FSC (Angelopoulos & Bates, 2021) means the worst-covered group is better protected.

Groups and protocol. The main groups are tertiles of the difficulty score ($G = 3$), with the thresholds determined on held-out training data and applied unchanged for each method. Thus, the group membership is consistent across methods within a split. Model predicted difficulty is a property on itself, not a measure of clinical severity. In addition we introduce two further partitions: tertiles of the true noise scale in the synthetic data and previously recorded dose on PPMI. For each metric, we compute it within a split and then average it. We repeat the random partition of Section 2 30 times at three coverage targets, $\alpha \in \{0.20, 0.10, 0.05\}$. Confidence intervals are paired percentile bootstraps over splits and describe how much results move across different splits of the same cohorts, not across independent samples. No adjustment for multiplicity is made. We report size stratified coverage violation (SSCV) (Angelopoulos & Bates, 2021) in Appendix A.

4 Results

4.1 Adaptation reallocates coverage and interval length

Adjusting adaptation strength reallocates coverage and interval length across difficulty strata while maintaining coverage close to target. The clinical scenario in Figure 1 shows unequal subgroup coverage even when marginal coverage target is met. Table 1 further shows that, at 80%, coverage varies very little across strengths: it goes from 0.804 to 0.809 on PPMI, and from 0.814 to 0.823 on XYZ Health.

In contrast, there are noticeable changes in group coverage and interval length. Figure 3 shows that allocation changes with w. As it increases, coverage decreases in the lower-difficulty stratum and rises in the higher-difficulty stratum. On PPMI at 80%, the lower-difficulty stratum goes from 0.885 at $w = 0$ to 0.686 at $w = 1$, nearly 20 percentage points (Appendix Table 5). Interval lengths also diverge: they are identical across strata at $w = 0$, but become shorter for predictions the model finds easier and longer for those it finds harder. Eventually the low and high-difficulty coverage curves cross. When they do, the lower-difficulty stratum becomes less covered than the higher-difficulty one. The crossings occur at approximately $w = 0.41$ on PPMI, 0.28 on freMTPL2, and 0.04 on XYZ Health. This crossing does not need to coincide with the lowest observed CovGap, reported in Appendix Figure 4 over $\alpha \in \{0.20, 0.10, 0.05\}$.

Table 1: Clinical cohorts: CovGap and FSC use predicted-difficulty strata. Split conformal and CASCADE-CRC (w=0) give identical intervals by construction, hence the shared row. Mondrian's mean length is ∞ on XYZ Health at $\alpha \in \{0.10, 0.05\}$ due to few calibration records per group. Bold values show the lowest CovGap and highest FSC per dataset and target while shaded rows keep at least one. Paired differences can be found in Appendix Table 4.

	$\alpha = 0.20$ (80% nominal)				$\alpha = 0.10$ (90% nominal)				$\alpha = 0.05$ (95% nominal)			
Method	Cov.	Len.	CovGap ↓	FSC ↑	Cov.	Len.	CovGap ↓	FSC ↑	Cov.	Len.	CovGap ↓	FSC ↑
PPMI (outpatient, multi-site)												
Split / CASCADE-CRC (w=0)	0.804	1.60	0.060	0.742	0.906	2.20	0.037	0.866	0.956	2.78	0.024	0.931
Mondrian	0.806	1.61	0.045	0.760	0.911	2.23	0.032	0.874	0.959	3.16	0.025	0.932
CQR	0.811	1.55	0.043	**0.768**	0.900	2.30	**0.030**	0.870	0.951	3.19	**0.020**	0.930
CASCADE	0.805	1.89	0.060	0.744	0.907	2.56	0.038	0.866	0.954	3.82	0.023	0.930
CASCADE-CRC (w=0.15)	0.806	1.60	0.050	0.760	0.909	2.20	0.032	0.877	0.958	2.75	0.022	**0.936**
CASCADE-CRC (w=0.25)	0.805	1.60	0.045	0.766	0.911	2.20	0.032	**0.881**	0.959	2.76	0.022	**0.936**
CASCADE-CRC (w=0.50)	0.804	1.63	**0.041**	0.758	0.910	2.25	0.031	0.875	0.956	2.85	0.023	0.931
CASCADE-CRC (w=0.75)	0.807	1.71	0.058	0.733	0.910	2.37	0.042	0.856	0.955	3.09	0.027	0.919
CASCADE-CRC (w=1)	0.809	1.83	0.083	0.684	0.910	2.62	0.063	0.812	0.957	3.66	0.035	0.902
XYZ Health (inpatient, multi-hospital)												
Split / CASCADE-CRC (w=0)	0.816	2.54	0.099	0.726	0.909	3.47	**0.059**	0.852	0.960	4.97	0.044	0.924
Mondrian	0.823	2.65	0.106	0.718	0.920	∞	0.070	0.851	0.967	∞	0.048	**0.929**
CQR	0.798	2.40	0.100	0.691	0.894	3.39	0.072	0.815	0.964	5.06	0.045	0.927
CASCADE	0.807	2.90	0.092	**0.728**	0.910	3.94	0.061	0.851	0.961	5.67	0.043	0.924
CASCADE-CRC (w=0.15)	0.818	2.54	**0.091**	0.726	0.908	3.48	0.061	0.851	0.960	4.94	**0.042**	0.926
CASCADE-CRC (w=0.25)	0.823	2.58	0.095	0.726	0.910	3.53	0.062	**0.853**	0.962	4.97	0.044	0.924
CASCADE-CRC (w=0.50)	0.821	2.63	0.102	0.699	0.911	3.64	0.067	0.847	0.960	5.20	0.049	0.909
CASCADE-CRC (w=0.75)	0.817	2.74	0.115	0.666	0.917	3.95	0.073	0.839	0.963	5.73	0.051	0.905
CASCADE-CRC (w=1)	0.814	2.98	0.130	0.637	0.918	4.47	0.083	0.819	0.961	6.61	0.054	0.896

Average interval length can fail to capture this reallocation. On PPMI at 80%, the switch from $w = 0$ to $w = 0.50$ reduces mean length in the lower-difficulty group from 1.60 to 1.23, whereas for the high-difficulty group it increases it from 1.60 to 2.02. Mean length increases by 0.023, while 56% of patients receive a shorter interval (Appendix Table 6).

4.2 The choice of strength depends on the target coverage and dataset

Reallocation results in better balanced coverage in some scenarios. Nevertheless, the most appropriate strength differs between datasets. At 80% coverage, the lowest observed CovGap is achieved by $w = 0.50$ on PPMI, and $w = 0.25$ on freMTPL2 (Tables 1 and 2). On XYZ Health, certain strengths produce lower CovGap compared to constant-length intervals ($w = 0$). None, however, showed improvement supported by the paired differences (Appendix Table 4).

Choosing the right coverage target also matters. On PPMI, the gap between low and high-difficulty groups under constant-length intervals narrows from 0.133 (0.885 versus 0.751) at 80% coverage to 0.040 at the 95% coverage (Appendix Table 5). The gains from moderate adaptation shown at 80% fade away at tighter targets, reversing on freMTPL2 at 95% (Tables 1 and 2). Thus, selecting an appropriate strength depends on the target coverage as well as the dataset.

The synthetic data shows a different story: The imbalance under constant length is large, CovGap = 0.118 against 0.060 on PPMI, and coverage balance keeps improving until reaching full adaptation (CovGap falls to 0.030 at $w = 1$), at a cost in length of about 20% (2.76 to 3.30, Table 2). CQR reaches CovGap = 0.044 at length 2.54, and CASCADE 0.106 at length 3.48, close to $w = 0.15$ on balance but with much wider intervals. Full adaptation attains the best CovGap and FSC across every method and target on synthetic data, so practitioners should examine the range rather than assume either endpoint.

4.3 Coverage balance depends on the evaluation groups

The comparisons carried out on real data shown above evaluate coverage balance along the primary difficulty score. In order to test whether the ranking persists under a different setting, we regroup PPMI by previous recorded dose. This allow us to separate patients already on dopamine replacement therapies from those with no recorded dose. This partition is fixed by the data and separates two clinically distinct situations: adjusting an existing regimen, where medication changes can be made

Table 2: External and synthetic settings: The synthetic data are grouped by the generated noise scale for evaluation. Mondrian calibrates on predicted-difficulty groups, so unlike the clinical settings its calibration and evaluation partitions differ. Layout and conventions as Table 1.

	$\alpha = 0.20$ (80% nominal)				$\alpha = 0.10$ (90% nominal)				$\alpha = 0.05$ (95% nominal)			
Method	Cov.	Len.	CovGap ↓	FSC ↑	Cov.	Len.	CovGap ↓	FSC ↑	Cov.	Len.	CovGap ↓	FSC ↑
freMTPL2 (French motor insurance)												
Split / CASCADE-CRC (w=0)	0.794	2.52	0.025	0.765	0.894	3.99	0.017	0.880	0.948	5.09	**0.010**	0.939
Mondrian	0.796	2.52	0.021	0.772	0.896	4.00	**0.015**	0.881	0.949	5.11	**0.010**	0.938
CQR	0.796	2.62	0.024	0.766	0.897	3.90	0.017	0.877	0.948	4.91	0.011	0.936
CASCADE	0.796	3.37	0.029	0.761	0.894	5.23	0.018	0.878	0.949	6.72	0.011	**0.940**
CASCADE-CRC (w=0.15)	0.794	2.50	0.019	0.774	0.895	3.99	**0.015**	**0.883**	0.948	5.08	0.012	0.934
CASCADE-CRC (w=0.25)	0.795	2.50	**0.018**	**0.778**	0.896	3.97	0.016	0.880	0.947	5.10	0.016	0.928
CASCADE-CRC (w=0.50)	0.797	2.54	0.023	0.769	0.898	3.99	0.026	0.867	0.947	5.20	0.025	0.912
CASCADE-CRC (w=0.75)	0.796	2.60	0.037	0.744	0.898	4.06	0.039	0.845	0.948	5.44	0.032	0.898
CASCADE-CRC (w=1)	0.797	2.70	0.058	0.710	0.898	4.25	0.052	0.821	0.949	5.88	0.042	0.882
Synthetic (κ=0.6; PPMI-sized budget)												
Split / CASCADE-CRC (w=0)	0.803	2.76	0.118	0.630	0.903	4.03	0.080	0.785	0.953	5.59	0.046	0.886
Mondrian	0.808	2.89	0.069	0.715	0.906	4.11	0.048	0.842	0.957	5.88	0.029	0.921
CQR	0.801	2.54	0.044	0.749	0.902	3.47	0.037	0.856	0.952	4.69	0.029	0.914
CASCADE	0.802	3.48	0.106	0.649	0.902	4.97	0.074	0.795	0.952	6.59	0.044	0.888
CASCADE-CRC (w=0.15)	0.803	2.72	0.105	0.652	0.902	3.87	0.072	0.798	0.952	5.23	0.042	0.891
CASCADE-CRC (w=0.25)	0.803	2.72	0.095	0.667	0.902	3.81	0.066	0.806	0.952	5.10	0.039	0.896
CASCADE-CRC (w=0.50)	0.801	2.77	0.071	0.705	0.901	3.82	0.051	0.831	0.952	5.05	0.031	0.909
CASCADE-CRC (w=0.75)	0.801	2.94	0.046	0.747	0.901	4.09	0.033	0.861	0.951	5.41	0.023	0.925
CASCADE-CRC (w=1)	0.800	3.30	**0.030**	**0.777**	0.900	4.79	**0.023**	**0.882**	0.950	6.68	**0.016**	**0.938**

in both directions (increase or decrease), and prescribing from no recorded dose, where every recorded change is an increase. At 80% coverage, full adaptation achieves the largest CovGap, 0.083, compared to 0.060 for constant length (Table 1), and the smallest in the previous-dose task, 0.028 versus 0.050 (Table 7). Moreover, it produces the most balanced coverage at 90% coverage. Hence, the usefulness of adaptation depends on the group whose coverage is being assessed. Thus, a method cannot be called balanced without saying across which groups.

4.4 Coverage balance and directional information differ

Group coverage shows how often intervals contain the recorded outcome, yet says nothing about how well those intervals distinguish a dose increase from a decrease. Thus, we examine how often intervals exclude zero. For our symmetric intervals, this happens when the absolute predicted change exceeds the half-length. On PPMI at 80% coverage, it ranges from 51.5% to 52.8% across the evaluated configurations and baselines (Appendix Table 8), despite larger differences within groups. From $w = 0$ to $w = 1$, zero-exclusion increases from 26.4% to 36.7% in the low-difficulty group and decreases from 56.4% to 48.1% in the high-difficulty group. The former also loses coverage as its intervals shorten, showing that subgroup coverage and directional information can move in opposite directions. These patterns, however, do not establish a within-patient trade-off. On XYZ Health, none of the evaluated methods produced intervals excluding zero for more than 1.92% of encounters at any of the three coverage targets (Appendix Table 8). Almost all intervals remain compatible with changes in either direction. Therefore, reporting CovGap alone would miss these differences. This motivates separate evaluation of coverage balance and directional information.

4.5 The interplay between strength and difficulty score

All above comparisons use a constant, primary difficulty score. We then examine whether changing its inputs or prediction target changes the preferred strength (Appendix Table 3). First, changing the first-stage features leads to small observed differences in CovGap: at any fixed strength, the differences between the six quantile scores stay within 0.007 of one another regardless of setting and target, while moving along rows changes CovGap by up to 0.042 on PPMI. Comparing each score at its best evaluated strength and adding Venn-Abers probabilities, their widths or their endpoints lowered CovGap below covariates alone only on XYZ Health, and the difference does not exceed 0.004. Second, what the score predicts shifts which strength gives the lowest observed CovGap. On PPMI at 80% coverage, the mean-residual score is best at $w = 0.50$, while the covariates-only quantile score performs better at $w = 0.75$; at $w = 1$, their CovGap values differ by 0.0387.

In the synthetic data, every score achieves its lowest CovGap at $w = 1$ (Appendix Table 9). Compared at their best evaluated configurations at the 80% target, however, the two families' CovGap values differ by at most 0.0036. These comparisons are retrospective. The lowest observed CovGap on PPMI at 80% is 0.0388, compared with 0.0432 for CQR.

5 Discussion

We show that adaptation strength facilitates reallocation of interval length, and consequently coverage across patient groups. This matters because the intervals carry the range of plausible dose changes, while coverage describes how reliably that range contains the recorded outcome. In PD, both under and over medication carry direct clinical consequences. This motivates careful assessment and further review for patients with wider intervals, and more precise information to guide dosing decisions for shorter ones, provided their coverage is adequate, as length alone does not establish clinical safety.

The caveat, however, is that marginal coverage does not show how reliably patient subpopulations are covered. In PPMI, the patients with the lowest predicted difficulty were covered by constant-length intervals at 88.5%, and the highest predicted difficulty at 75.1%, against an 80% target coverage. A clinician looking at the marginal coverage sees 80%, and has no idea which patients fall on either side of that gap. CASCADE-CRC provides a way to change this allocation explicitly via adaptation strength: at $w = 0.25$, the same gap reduces from 0.133 to 0.056, with the highest-difficulty group covered at 79.4% and the lowest at 85.0%, and a slightly lower observed mean interval length.

While evaluating the reallocation of coverage and interval length, it is necessary to consider both the clinical purpose and the statistical properties. Similar average interval lengths can conceal substantial differences in the precision and reliability of predictions for different groups. Therefore, making adaptation strength explicit helps examine how much a difficulty estimate should influence interval length, alongside its influence on coverage across patient groups. The choice of the difficulty estimator matters too: a score that yields poor balance under full scaling may provide better balance when its influence is moderated.

Selecting the strength a priori is more difficult than its evaluation. A strategy here could be selecting the strength that minimizes CovGap on held-out calibration samples. In our study, it improved coverage balance on three out of four datasets. On the other hand, a rule requiring statistical evidence of improvement did not certify improvement at the sample sizes tested (Appendix C). Group choice is an important criterion as well, as balancing coverage across predicted-difficulty strata does not necessarily imply balance across clinically defined groups.

Finally, we provide the following suggestions for practitioners: report group coverage and interval length next to the marginal rate; sweep across a range of adaptation strengths rather than settling on one; and evaluate at least one grouping the scaling score had no part in defining, ideally patient characteristics that matter to the decision. Together, these can help a clinician understand which patients the model really serves well, and at what cost in interval length.

6 Conclusion and limitations

CASCADE-CRC varies interval length depending on the expected difficulty. It reallocates coverage across patient groups, with the marginal guarantee preserved because each strength is calibrated separately. For instance, at moderate strength ($w = 0.5$), it decreased CovGap from 0.060 to 0.041 while slightly increasing mean interval length from 1.60 to 1.63 at 80% coverage on the PPMI dataset. Thus, adaptation strength affects how reliably different patient groups are covered and how wide their intervals are. We note that no single strength was the most effective across all scenarios, and that the preferred strength depends on the dataset, coverage target and evaluation grouping.

We observe that grouping patients by their previous recorded dose changed which adaptation strength performed best. Our primary evaluation assesses coverage balance across difficulty strata, and clinically defined risk groups can provide a further test. The confidence intervals reported describe sensitivity to data splitting within the cohorts. Further evaluation would be needed to determine whether the allocation patterns hold for other datasets. Overall, CASCADE-CRC shows that meeting marginal coverage is necessary but not sufficient, and that adaptation strength shapes how coverage and interval length are distributed across patient groups.

AI USE STATEMENT

Large Language Models did not play a significant role and were only used for proof-reading.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

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

Melissa J. Armstrong and Michael S. Okun. 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.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Milan Beckers, Danny Hommel, Herma Lennaerts, Clementine Stuijt, Paul Smit, and Bastiaan R Bloem. Discontinuation or acute unplanned cessation of oral dopaminergic medications in persons with parkinson's disease: A practice review. *Journal of Parkinson's Disease*, 15(6):1062–1077, 2025.

Arthur Charpentier. *Computational actuarial science with R*. CRC press, 2014.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Ricardo Diaz-Rincon, Muxuan Liang, Adolfo Ramirez-Zamora, and Benjamin Shickel. 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*, pp. 1–20, Rochester, MN, 15–16 Aug 2025. PMLR. URL https://proceedings.mlr.press/v298/diaz-rincon25a.html.

Ricardo Diaz-Rincon, Muxuan Liang, Adolfo Ramirez-Zamora, and Benjamin Shickel. Cascade conformal prediction: Uncertainty-adaptive prediction intervals for two-stage clinical decision support. *arXiv preprint arXiv:2605.20468*, 2026.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

R Ghosh and BJ Liddle. Emergency presentations of parkinson's disease: early recognition and treatment are crucial for optimum outcome. *Postgraduate medical journal*, 87(1024):125–131, 2011.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arxiv preprint*, 2024.

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

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, Werner Poewe, Brit Mollenhauer, Todd Sherer, Mark Frasier, Claire Meunier, Alice Rudolph, Cindy Casaceli, John Seibyl, Susan Mendick, Norbert Schuff, Ying Zhang, Arthur Toga, Karen Crawford, Alison Ansbach, Pasquale de Blasio, Michele Piovella, John Trojanowski, Les Shaw, Andrew Singleton, Keith Hawkins, Jamie Eberling, David Russell, Laura Leary, Stewart Factor, Barbara Sommerfeld, Penelope Hogarth, Emily Pighetti, Karen Williams, David Standaert, Stephanie Guthrie, Robert Hauser, Holly Delgado, Joseph Jankovic, Christine Hunter, Matthew Stern, Baochan Tran, Jim Leverenz,

Marne Baca, Sam Frank, Cathi Ann Thomas, Irene Richard, Cheryl Deeley, Linda Rees, Fabienne Sprenger, Elisabeth Lang, Holly Shill, Sanja Obradov, Hubert Fernandez, Adrienna Winters, Daniela Berg, Katharina Gauss, Douglas Galasko, Deborah Fontaine, Zoltan Mari, Melissa Gerstenhaber, David Brooks, Sophie Malloy, Paolo Barone, Katia Longo, Tom Comery, Bernard Ravina, Igor Grachev, Kim Gallagher, Michelle Collins, Katherine L. Widnell, Suzanne Ostrowizki, Paulo Fontoura, F. Hoffmann La-Roche, Tony Ho, Johan Luthman, Marcel van der Brug, Alastair D. Reith, and Peggy Taylor. The parkinson progression marker initiative (ppmi), 12 2011. ISSN 03010082.

Edward J Newman, Donald G Grosset, and Peter GE Kennedy. The parkinsonism-hyperpyrexia syndrome. *Neurocritical Care*, 10(1):136–140, 2009.

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

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pp. 345–356, 2002.

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

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In H Wallach, H Larochelle, A Beygelzimer, F d Alché-Buc, E Fox, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2), 2020.

JM Senard, S Rai, M Lapeyre-Mestre, C Brefel, O Rascol, A Rascol, and JL Montastruc. Prevalence of orthostatic hypotension in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 63(5):584–589, 1997.

Jacob Taylor, William S Anderson, Jason Brandt, Zoltan Mari, and Gregory M Pontone. Neuropsychiatric complications of parkinson disease treatments: importance of multidisciplinary care. *The American Journal of Geriatric Psychiatry*, 24(12):1171–1180, 2016.

Claire L. Tomlinson, Rebecca Stowe, Smitaa Patel, Caroline Rick, Richard Gray, and Carl E. Clarke. 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.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, Alex Gammerman, et al. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005.

Table 3: CovGap at the 80% target for seven difficulty scores crossed with six adaptation strengths, on the clinical cohorts (predicted-difficulty strata). The shaded row is the primary score, a mean absolute residual; the others predict a $(1-\alpha)$ quantile from different feature sets. Bold marks the best strength in each row. freMTPL2 and the synthetic data are in Appendix Table 9.

Difficulty score	w=0	0.15	0.25	0.50	0.75	1
PPMI						
mean absolute residual	0.0601	0.0495	0.0451	**0.0412**	0.0575	0.0831
covariates	0.0601	0.0560	0.0509	0.0432	**0.0388**	0.0444
raw probability	0.0601	0.0543	0.0497	0.0438	**0.0411**	0.0454
VA probability	0.0601	0.0558	0.0514	0.0435	**0.0421**	0.0469
VA probability + width	0.0601	0.0561	0.0491	0.0432	**0.0395**	0.0448
VA upper	0.0601	0.0565	0.0509	0.0439	**0.0414**	0.0461
VA endpoints	0.0601	0.0545	0.0503	0.0440	**0.0412**	0.0455
baselines: constant 0.0601, Mondrian 0.0450, CQR 0.0432						
XYZ Health						
mean absolute residual	0.0985	**0.0915**	0.0952	0.1017	0.1153	0.1304
covariates	0.0985	0.0993	0.0971	0.0981	**0.0947**	0.1023
raw probability	0.0985	0.0972	0.0963	0.0962	**0.0921**	0.0957
VA probability	0.0985	0.0972	0.0960	0.0939	**0.0907**	0.1014
VA probability + width	0.0985	0.0988	0.0990	**0.0923**	0.0924	0.0985
VA upper	0.0985	0.0976	0.0948	**0.0939**	0.0951	0.1023
VA endpoints	0.0985	0.0986	0.0961	**0.0913**	0.0929	0.0962
baselines: constant 0.0985, Mondrian 0.1055, CQR 0.1004						

Table 4: Paired differences between each adaptation strength and constant-length intervals (w=0), computed within split and averaged over splits, or over the 240 independent replications of Table 2. Brackets are 95% percentile bootstrap intervals from 4,000 paired resamples, shown for ΔCovGap. Bold numbers show interval excluding zero, for every column. Negative ΔCovGap and Δlength and positive ΔFSC favor adaptation. Shaded rows improve CovGap in at least one target. Intervals describe sensitivity to the split on reused cohorts, not independent replication, and are not adjusted for multiple comparisons.

	$\alpha = 0.20$			$\alpha = 0.10$			$\alpha = 0.05$		
Method	ΔCovGap	ΔFSC	ΔLen.	ΔCovGap	ΔFSC	ΔLen.	ΔCovGap	ΔFSC	ΔLen.
PPMI (outpatient, multi-site)									
CASCADE-CRC (w=0.15)	**-0.0106**[-0.0140,-0.0072]	**+0.0183**	-0.002	**-0.0047**[-0.0078,-0.0014]	**+0.0109**	-0.002	**-0.0023**[-0.0045,-0.0000]	+0.0051	**-0.023**
CASCADE-CRC (w=0.25)	**-0.0150**[-0.0207,-0.0094]	**+0.0241**	-0.006	-0.0047[-0.0096,+0.0002]	**+0.0149**	-0.002	-0.0024[-0.0048,+0.0000]	+0.0051	-0.013
CASCADE-CRC (w=0.5)	**-0.0189**[-0.0287,-0.0098]	**+0.0157**	+0.023	-0.0059[-0.0126,+0.0009]	+0.0091	**+0.047**	-0.0016[-0.0054,+0.0020]	+0.0004	**+0.075**
CASCADE-CRC (w=0.75)	-0.0025[-0.0147,+0.0089]	-0.0094	**+0.104**	+0.0053[-0.0033,+0.0134]	-0.0104	**+0.170**	+0.0028[-0.0030,+0.0080]	-0.0124	**+0.314**
CASCADE-CRC (w=1)	**+0.0230**[+0.0072,+0.0380]	**-0.0582**	**+0.228**	**+0.0260**[+0.0138,+0.0378]	**-0.0541**	**+0.425**	**+0.0109**[+0.0036,+0.0181]	**-0.0285**	**+0.880**
XYZ Health (inpatient, multi-hospital)									
CASCADE-CRC (w=0.15)	-0.0070[-0.0157,+0.0009]	+0.0002	+0.005	+0.0019[-0.0031,+0.0067]	-0.0015	+0.003	-0.0015[-0.0049,+0.0017]	+0.0018	-0.027
CASCADE-CRC (w=0.25)	-0.0033[-0.0133,+0.0065]	-0.0000	+0.042	+0.0027[-0.0028,+0.0081]	+0.0009	+0.052	+0.0003[-0.0039,+0.0041]	-0.0001	-0.005
CASCADE-CRC (w=0.5)	+0.0032[-0.0121,+0.0188]	-0.0275	**+0.097**	**+0.0075**[+0.0010,+0.0135]	-0.0053	**+0.166**	+0.0050[-0.0022,+0.0124]	-0.0145	+0.226
CASCADE-CRC (w=0.75)	+0.0167[-0.0021,+0.0338]	**-0.0599**	**+0.198**	**+0.0131**[+0.0005,+0.0242]	-0.0127	**+0.478**	**+0.0068**[+0.0018,+0.0128]	**-0.0184**	**+0.755**
CASCADE-CRC (w=1)	**+0.0319**[+0.0096,+0.0511]	**-0.0890**	**+0.445**	**+0.0236**[+0.0101,+0.0366]	**-0.0332**	**+1.001**	**+0.0105**[+0.0046,+0.0166]	**-0.0279**	**+1.641**
freMTPL2 (French motor insurance)									
CASCADE-CRC (w=0.15)	**-0.0060**[-0.0080,-0.0041]	**+0.0090**	**-0.018**	-0.0014[-0.0032,+0.0005]	+0.0025	-0.005	**+0.0017**[+0.0002,+0.0031]	**-0.0043**	-0.014
CASCADE-CRC (w=0.25)	**-0.0065**[-0.0095,-0.0036]	**+0.0128**	**-0.017**	-0.0008[-0.0034,+0.0019]	-0.0001	-0.023	**+0.0055**[+0.0035,+0.0074]	**-0.0110**	+0.004
CASCADE-CRC (w=0.5)	-0.0024[-0.0077,+0.0025]	+0.0035	+0.017	**+0.0088**[+0.0047,+0.0128]	**-0.0132**	+0.003	**+0.0141**[+0.0117,+0.0164]	**-0.0265**	**+0.110**
CASCADE-CRC (w=0.75)	**+0.0122**[+0.0064,+0.0177]	**-0.0209**	**+0.082**	**+0.0218**[+0.0174,+0.0261]	**-0.0349**	**+0.075**	**+0.0213**[+0.0186,+0.0240]	**-0.0402**	**+0.347**
CASCADE-CRC (w=1)	**+0.0331**[+0.0268,+0.0392]	**-0.0552**	**+0.183**	**+0.0348**[+0.0306,+0.0388]	**-0.0590**	**+0.263**	**+0.0317**[+0.0288,+0.0345]	**-0.0568**	**+0.786**
Synthetic (κ=0.6)									
CASCADE-CRC (w=0.15)	**-0.0136**[-0.0143,-0.0128]	**+0.0219**	**-0.034**	**-0.0082**[-0.0089,-0.0076]	**+0.0124**	**-0.160**	**-0.0040**[-0.0045,-0.0036]	**+0.0057**	**-0.360**
CASCADE-CRC (w=0.25)	**-0.0231**[-0.0241,-0.0220]	**+0.0372**	**-0.040**	**-0.0139**[-0.0148,-0.0130]	**+0.0211**	**-0.217**	**-0.0065**[-0.0072,-0.0059]	**+0.0100**	**-0.491**
CASCADE-CRC (w=0.5)	**-0.0472**[-0.0490,-0.0452]	**+0.0751**	+0.009	**-0.0294**[-0.0307,-0.0280]	**+0.0456**	**-0.207**	**-0.0146**[-0.0156,-0.0137]	**+0.0236**	**-0.533**
CASCADE-CRC (w=0.75)	**-0.0727**[-0.0756,-0.0697]	**+0.1173**	**+0.182**	**-0.0466**[-0.0487,-0.0446]	**+0.0752**	+0.057	**-0.0232**[-0.0246,-0.0219]	**+0.0390**	**-0.179**
CASCADE-CRC (w=1)	**-0.0880**[-0.0918,-0.0840]	**+0.1474**	**+0.547**	**-0.0573**[-0.0600,-0.0545]	**+0.0971**	**+0.761**	**-0.0303**[-0.0320,-0.0286]	**+0.0528**	**+1.091**

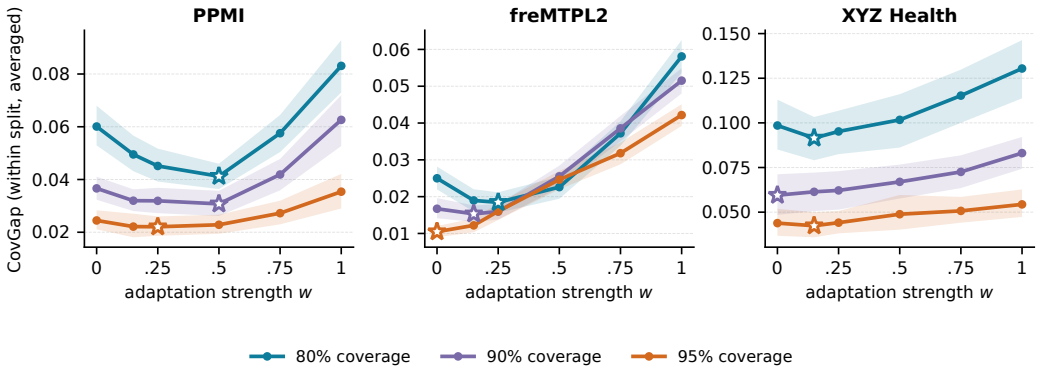

Figure 4: CovGap against adaptation strength at the three coverage targets, computed within each split and averaged. Stars show the lowest observed CovGap among the evaluated strengths. Bands are 95% bootstrap intervals over splits. On PPMI at the 80% target, moving from $w = 0$ to $w = 0.50$ reduces CovGap by 31.7%, while that mean absolute deviation falls by 63.1%; on XYZ Health at 90%, constant-length intervals give split-averaged coverages of 0.902, 0.915 and 0.910, a mean deviation of 0.009, against a CovGap of 0.059.

Table 5: Group coverage and mean interval length by predicted-difficulty stratum at all three targets, for the primary mean-residual score and the baselines. Each entry averages the within-split stratum value over splits; groups are identical across methods within a split. Strata are not equal in size, since thresholds are fixed on held-out data: mean evaluated units per split are 64.6, 87.5 and 87.1 on PPMI, 19.5, 16.9 and 24.3 on XYZ Health, and about 379 each on freMTPL2.

	$\alpha = 0.20$ (80% nominal)						$\alpha = 0.10$ (90% nominal)						$\alpha = 0.05$ (95% nominal)					
	coverage			length			coverage			length			coverage			length		
Method	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
PPMI (outpatient, multi-site)																		
Split / CASCADE-CRC (w=0)	0.885	0.799	0.751	1.60	1.60	1.60	0.943	0.913	0.873	2.20	2.20	2.20	0.975	0.963	0.935	2.78	2.78	2.78
Mondrian	0.822	0.799	0.801	1.32	1.63	1.80	0.915	0.914	0.904	1.88	2.23	2.47	0.963	0.960	0.953	2.86	2.72	3.83
CQR	0.787	0.818	0.823	1.16	1.59	1.78	0.889	0.903	0.906	1.71	2.20	2.84	0.946	0.954	0.951	2.44	2.97	3.95
CASCADE	0.886	0.798	0.755	1.99	1.80	1.91	0.948	0.911	0.875	2.62	2.49	2.59	0.973	0.959	0.936	4.00	3.67	3.89
CASCADE-CRC (w=0.15)	0.868	0.793	0.776	1.48	1.57	1.72	0.930	0.910	0.893	2.03	2.16	2.36	0.968	0.962	0.946	2.55	2.71	2.95
CASCADE-CRC (w=0.25)	0.850	0.786	0.794	1.40	1.55	1.79	0.923	0.907	0.907	1.93	2.13	2.46	0.962	0.961	0.953	2.42	2.68	3.10
CASCADE-CRC (w=0.50)	0.800	0.782	0.831	1.23	1.53	2.02	0.895	0.903	0.930	1.70	2.11	2.79	0.943	0.960	0.961	2.15	2.69	3.54
CASCADE-CRC (w=0.75)	0.745	0.790	0.869	1.08	1.56	2.32	0.862	0.909	0.947	1.50	2.16	3.22	0.923	0.964	0.969	1.96	2.82	4.21
CASCADE-CRC (w=1)	0.686	0.805	0.904	0.94	1.62	2.71	0.813	0.928	0.964	1.35	2.32	3.88	0.903	0.975	0.977	1.88	3.24	5.43
XYZ Health (inpatient, multi-hospital)																		
Split / CASCADE-CRC (w=0)	0.812	0.842	0.803	2.54	2.54	2.54	0.902	0.915	0.910	3.47	3.47	3.47	0.953	0.967	0.964	4.97	4.97	4.97
Mondrian	0.824	0.843	0.812	2.72	2.73	2.60	0.912	0.926	0.928	3.94	∞	3.89	0.968	0.967	0.974	∞	∞	∞
CQR	0.765	0.765	0.852	2.10	2.29	2.71	0.879	0.867	0.923	2.99	3.26	3.80	0.952	0.960	0.979	4.66	4.93	5.49
CASCADE	0.809	0.825	0.793	2.88	2.79	2.93	0.906	0.908	0.917	3.94	3.73	4.09	0.955	0.963	0.968	5.58	5.44	5.82
CASCADE-CRC (w=0.15)	0.789	0.845	0.831	2.39	2.51	2.69	0.893	0.912	0.920	3.26	3.43	3.68	0.946	0.965	0.973	4.64	4.88	5.23
CASCADE-CRC (w=0.25)	0.781	0.837	0.857	2.32	2.52	2.83	0.889	0.914	0.926	3.16	3.44	3.87	0.942	0.964	0.982	4.46	4.85	5.45
CASCADE-CRC (w=0.50)	0.737	0.835	0.881	2.10	2.51	3.14	0.871	0.919	0.941	2.90	3.47	4.34	0.928	0.961	0.987	4.14	4.95	6.19
CASCADE-CRC (w=0.75)	0.691	0.839	0.899	1.91	2.55	3.53	0.855	0.932	0.964	2.75	3.68	5.10	0.919	0.969	0.995	3.98	5.31	7.36
CASCADE-CRC (w=1)	0.651	0.851	0.923	1.78	2.71	4.13	0.834	0.940	0.976	2.67	4.06	6.20	0.908	0.970	0.996	3.93	5.97	9.12
freMTPL2 (French motor insurance)																		
Split / CASCADE-CRC (w=0)	0.822	0.796	0.769	2.52	2.52	2.52	0.903	0.896	0.883	3.99	3.99	3.99	0.949	0.949	0.946	5.09	5.09	5.09
Mondrian	0.799	0.795	0.793	2.25	2.51	2.76	0.892	0.899	0.895	3.72	4.04	4.21	0.947	0.950	0.949	5.04	5.12	5.15
CQR	0.770	0.798	0.817	2.13	2.61	3.07	0.882	0.899	0.909	3.46	3.89	4.29	0.940	0.949	0.956	4.57	4.88	5.22
CASCADE	0.831	0.798	0.763	3.76	3.30	3.10	0.908	0.896	0.880	5.85	5.12	4.79	0.951	0.950	0.946	7.43	6.53	6.28
CASCADE-CRC (w=0.15)	0.809	0.792	0.783	2.35	2.48	2.65	0.894	0.895	0.897	3.75	3.96	4.22	0.938	0.947	0.957	4.78	5.04	5.37
CASCADE-CRC (w=0.25)	0.800	0.791	0.794	2.25	2.47	2.75	0.888	0.893	0.905	3.58	3.92	4.35	0.929	0.947	0.964	4.59	5.03	5.59
CASCADE-CRC (w=0.50)	0.776	0.792	0.819	2.03	2.47	3.03	0.869	0.892	0.927	3.21	3.90	4.77	0.912	0.948	0.977	4.17	5.07	6.21
CASCADE-CRC (w=0.75)	0.746	0.794	0.842	1.83	2.50	3.36	0.847	0.893	0.946	2.86	3.91	5.25	0.898	0.954	0.984	3.83	5.24	7.03
CASCADE-CRC (w=1)	0.710	0.801	0.869	1.64	2.57	3.75	0.821	0.899	0.963	2.57	4.04	5.91	0.882	0.965	0.992	3.55	5.59	8.17

A Additional results

A.1 Coverage allocation across strengths and targets

A.2 Further tables

SSCV. Width bins are four equal-count quantile bins of the full set length, ties collapsed at relative tolerance 10^{-9} of the median width, minimum 20 observations per bin. Constant-width methods

Table 6: Mean interval length by predicted-difficulty stratum at all three targets. The shaded column is the fraction of evaluated records whose interval is shorter than under constant length (w=0), paired by record; it is reported for the clinical cohorts, where record-level outputs are retained. Conventions as Table 5.

	$\alpha = 0.20$ (80% nominal)				$\alpha = 0.10$ (90% nominal)				$\alpha = 0.05$ (95% nominal)			
Method	Low	Med	High	Frac. shorter	Low	Med	High	Frac. shorter	Low	Med	High	Frac. shorter
PPMI (outpatient, multi-site)												
Split / CASCADE-CRC (w=0)	1.60	1.60	1.60	—	2.20	2.20	2.20	—	2.78	2.78	2.78	—
Mondrian	1.32	1.63	1.80	0.40	1.88	2.23	2.47	0.43	2.86	2.72	3.83	0.42
CQR	1.16	1.59	1.78	0.53	1.71	2.20	2.84	0.50	2.44	2.97	3.95	0.42
CASCADE	1.99	1.80	1.91	0.25	2.62	2.49	2.59	0.34	4.00	3.67	3.89	0.32
CASCADE-CRC (w=0.15)	1.48	1.57	1.72	0.59	2.03	2.16	2.36	0.58	2.55	2.71	2.95	0.62
CASCADE-CRC (w=0.25)	1.40	1.55	1.79	0.59	1.93	2.13	2.46	0.59	2.42	2.68	3.10	0.60
CASCADE-CRC (w=0.50)	1.23	1.53	2.02	0.56	1.70	2.11	2.79	0.54	2.15	2.69	3.54	0.53
CASCADE-CRC (w=0.75)	1.08	1.56	2.32	0.50	1.50	2.16	3.22	0.48	1.96	2.82	4.21	0.43
CASCADE-CRC (w=1)	0.94	1.62	2.71	0.45	1.35	2.32	3.88	0.40	1.88	3.24	5.43	0.33
XYZ Health (inpatient, multi-hospital)												
Split / CASCADE-CRC (w=0)	2.54	2.54	2.54	—	3.47	3.47	3.47	—	4.97	4.97	4.97	—
Mondrian	2.72	2.73	2.60	0.34	3.94	∞	3.89	0.38	∞	∞	∞	0.27
CQR	2.10	2.29	2.71	0.63	2.99	3.26	3.80	0.58	4.66	4.93	5.49	0.49
CASCADE	2.88	2.79	2.93	0.33	3.94	3.73	4.09	0.28	5.58	5.44	5.82	0.28
CASCADE-CRC (w=0.15)	2.39	2.51	2.69	0.52	3.26	3.43	3.68	0.52	4.64	4.88	5.23	0.53
CASCADE-CRC (w=0.25)	2.32	2.52	2.83	0.48	3.16	3.44	3.87	0.49	4.46	4.85	5.45	0.53
CASCADE-CRC (w=0.50)	2.10	2.51	3.14	0.49	2.90	3.47	4.34	0.46	4.14	4.95	6.19	0.49
CASCADE-CRC (w=0.75)	1.91	2.55	3.53	0.47	2.75	3.68	5.10	0.40	3.98	5.31	7.36	0.42
CASCADE-CRC (w=1)	1.78	2.71	4.13	0.41	2.67	4.06	6.20	0.34	3.93	5.97	9.12	0.34

Table 7: PPMI on the previous recorded dose partition (previous LEDD positive, 128.6 evaluated patients per split, versus zero, 110.6). The partition is undefined on XYZ Health, which has no previous-dose-zero records. Conventions as Table 1.

	$\alpha = 0.20$		$\alpha = 0.10$		$\alpha = 0.05$	
Method	CovGap ↓	FSC ↑	CovGap ↓	FSC ↑	CovGap ↓	FSC ↑
Split / CASCADE-CRC (w=0)	0.050	0.757	0.026	0.887	0.020	0.943
Mondrian	0.043	0.766	0.027	0.891	0.020	**0.948**
CQR	0.037	**0.793**	0.024	0.883	0.021	0.935
CASCADE	0.053	0.758	0.029	0.885	**0.017**	0.944
CASCADE-CRC (w=0.25)	0.048	0.758	0.027	0.892	0.019	**0.948**
CASCADE-CRC (w=0.50)	0.043	0.764	0.027	0.891	0.020	0.945
CASCADE-CRC (w=0.75)	0.035	0.775	0.025	0.891	0.019	0.942
CASCADE-CRC (w=1)	**0.028**	0.785	**0.023**	**0.894**	0.023	0.939

occupy a single bin, so their SSCV reduces to $|\text{coverage} - (1-\alpha)|$ and is degenerate. PPMI at 80%: CASCADE-CRC (w=0) 0.032 (one bin), Mondrian 0.057, w=0.5 0.096, CQR 0.110. These are not comparable across arm types.

B Calibration conditions and comparator details

Conformal risk control conditions. For a nested family $\{C_\lambda\}$ and a loss $L(\lambda)$ bounded above by B, conformal risk control selects

$$\hat{\lambda} = \inf\Big\{\lambda : \tfrac{m}{m+1}\widehat{R}_m(\lambda) + \tfrac{B}{m+1} \le \alpha\Big\}, \qquad \widehat{R}_m(\lambda) = \tfrac{1}{m}\textstyle\sum_{i=1}^{m} L_i(\lambda), \tag{12}$$

with $\hat{\lambda} = \lambda_{\max}$ when no λ satisfies the condition, and guarantees $\mathbb{E}[L_{m+1}(\hat{\lambda})] \le \alpha$ (Angelopoulos et al., 2024). The statement requires a nested family, $\lambda \le \lambda' \Rightarrow C_\lambda(x) \subseteq C_{\lambda'}(x)$; a loss $L(\lambda) = \ell(Y, C_\lambda(X))$ that is non-increasing in λ, right-continuous, and bounded above by B; and $L(\lambda_{\max}) \le \alpha$ almost surely. We use the miscoverage loss $L(\lambda) = \mathbf{1}\{Y \notin C_{w,\lambda}(X)\}$ with $B = 1$ and $\Lambda = [0, \infty]$, so $L(\lambda_{\max}) = 0$ and the condition holds. The guarantee is stated for a strength w fixed before the calibration set is used.

Venn-Abers guarantee. A Venn-Abers predictor outputs $p_0(x)$ and $p_1(x)$ obtained by labelling the test object with each class in turn and isotonically calibrating. Under the Venn assumptions one of

Table 8: Zero-exclusion rate: the fraction of evaluated records whose interval excludes zero and therefore indicates a direction of change. Computed from the actual endpoints, so it applies to the asymmetric CQR intervals as well as the symmetric ones; empty intervals are not counted as informative and their frequency is reported separately. Rates are averaged within split and then across splits; subgroup columns are pooled over splits within a stratum. Zero-exclusion concerns the recorded dose changes.

				By difficulty stratum		
Method	Target	Zero-excl.	Empty	Low	Med	High
PPMI (outpatient, multi-site)						
w=0	80%	0.5217	0.00000	0.264	0.669	0.564
w=0.25	80%	0.5168	0.00000	0.280	0.675	0.533
w=0.50	80%	0.5152	0.00000	0.295	0.677	0.515
w=1	80%	0.5204	0.00000	0.367	0.672	0.481
CQR	80%	0.5278	0.00082	0.309	0.691	0.526
Mondrian	80%	0.5215	0.00000	0.290	0.669	0.544
w=0	90%	0.4839	0.00000	0.245	0.644	0.500
w=0.25	90%	0.4793	0.00000	0.251	0.645	0.481
w=0.50	90%	0.4780	0.00000	0.259	0.645	0.471
w=1	90%	0.4836	0.00000	0.295	0.643	0.463
CQR	90%	0.4815	0.00068	0.259	0.652	0.475
Mondrian	90%	0.4807	0.00000	0.253	0.644	0.485
w=0	95%	0.4683	0.00000	0.238	0.634	0.471
w=0.25	95%	0.4681	0.00000	0.242	0.636	0.467
w=0.50	95%	0.4683	0.00000	0.244	0.637	0.465
w=1	95%	0.4667	0.00000	0.254	0.634	0.456
CQR	95%	0.4717	0.00070	0.245	0.643	0.467
Mondrian	95%	0.4595	0.00000	0.231	0.636	0.451
XYZ Health (inpatient, multi-hospital)						
w=0	80%	0.0055	0.00000	0.005	0.004	0.007
w=0.25	80%	0.0060	0.00000	0.012	0.004	0.003
w=0.50	80%	0.0076	0.00000	0.019	0.004	0.001
w=1	80%	0.0192	0.00000	0.056	0.002	0.001
CQR	80%	0.0000	0.00000	0.000	0.000	0.000
Mondrian	80%	0.0082	0.00000	0.003	0.004	0.015
w=0	90%	0.0017	0.00000	0.003	0.000	0.001
w=0.25	90%	0.0017	0.00000	0.003	0.000	0.001
w=0.50	90%	0.0017	0.00000	0.003	0.000	0.001
w=1	90%	0.0049	0.00000	0.015	0.000	0.000
CQR	90%	0.0000	0.00000	0.000	0.000	0.000
Mondrian	90%	0.0006	0.00000	0.000	0.000	0.001
w=0	95%	0.0000	0.00000	0.000	0.000	0.000
w=0.25	95%	0.0005	0.00000	0.002	0.000	0.000
w=0.50	95%	0.0011	0.00000	0.003	0.000	0.000
w=1	95%	0.0022	0.00000	0.007	0.000	0.000
CQR	95%	0.0000	0.00000	0.000	0.000	0.000
Mondrian	95%	0.0000	0.00000	0.000	0.000	0.000

the two is the probability produced by a valid multiprobabilistic predictor, so the guarantee attaches to the pair through a selector rather than to a fixed endpoint or to the midpoint (Vovk & Petej, 2012). CASCADE uses the width $u_{\mathrm{VA}} = p_1 - p_0$ as its uncertainty signal; the alternative difficulty models of Section 2.5 also use the midpoint and the endpoints as features. These constructions do not require either endpoint or the midpoint to be a calibrated probability.

Alternative difficulty scores. The six quantile scores of Section 2.5 are fitted by gradient-boosted regression under pinball loss at the $(1-\alpha)$ quantile of $|Y - \hat{f}(X)|$, refitted per coverage target, and normalized by equation 3 exactly as the primary score is: floored at $\epsilon = \max\{q_{0.05}(g), 10^{-6}\}$ on $\mathcal{D}_{\mathrm{hold}}$, then divided by the mean of the floored score on the same subset. Each is then swept over $w \in \{0, 0.15, 0.25, 0.50, 0.75, 1\}$ through equation 4, giving 36 quantile configurations; with the primary mean-residual score at the same six strengths this is 42 score by strength configurations per dataset and target. Normalization is what makes the sweep meaningful: at w=1 a common positive factor cancels, since $\sigma'(x) = \sigma(x)/c$ implies $\hat{\lambda}' = c\hat{\lambda}$ and $\hat{\lambda}'\sigma'(x) = \hat{\lambda}\sigma(x)$, but at intermediate w the constant in equation 4 fixes the units and an unnormalized score would dominate the interpolation.

CASCADE. The scale is equation 10 with $\bar{u}_{\mathrm{VA}}$ the mean Venn-Abers width on $\mathcal{D}_{\mathrm{hold}}$. We search over $\beta \in \{0, 0.1, \ldots, 1.0\}$ and select, among values attaining the nominal level on an internal validation half of the training data, the one minimizing absolute coverage error, breaking ties toward shorter mean length. The original CASCADE selects β by comparing candidate coverage on held-out

Table 9: As Table 3, for freMTPL2 (predicted-difficulty strata) and the synthetic data (grouped by the generated noise scale; κ=0.6, PPMI-sized budget, 150 independent replications). These are a separate run from the 240 of Table 2, so the constant-width value differs slightly (0.1208 against 0.118). At the 90% and 95% targets, not shown, the six quantile scores remain within 0.006 of one another at every strength in every dataset. In the synthetic data the first-stage features are functions of X by construction, so the comparison among feature sets is uninformative there; the comparison between the mean and quantile scores is not.

Difficulty score	w=0	0.15	0.25	0.50	0.75	1
freMTPL2						
mean absolute residual	0.0250	0.0190	**0.0185**	0.0226	0.0372	0.0581
covariates	0.0250	0.0209	0.0189	**0.0185**	0.0227	0.0332
raw probability	0.0250	0.0214	0.0201	**0.0196**	0.0213	0.0297
VA probability	0.0250	0.0219	0.0197	**0.0197**	0.0217	0.0313
VA probability + width	0.0250	0.0220	0.0199	**0.0190**	0.0212	0.0294
VA upper	0.0250	0.0218	0.0197	**0.0196**	0.0216	0.0316
VA endpoints	0.0250	0.0219	0.0198	**0.0197**	0.0218	0.0313
baselines: constant 0.0250, Mondrian 0.0214, CQR 0.0237						
Synthetic (κ=0.6; PPMI-sized budget)						
mean absolute residual	0.1208	0.1074	0.0977	0.0739	0.0482	**0.0321**
covariates	0.1208	0.1091	0.1008	0.0795	0.0564	**0.0358**
raw probability	0.1208	0.1093	0.1011	0.0789	0.0565	**0.0367**
VA probability	0.1208	0.1090	0.1007	0.0784	0.0557	**0.0377**
VA probability + width	0.1208	0.1090	0.1008	0.0790	0.0568	**0.0365**
VA upper	0.1208	0.1090	0.1008	0.0789	0.0562	**0.0372**
VA endpoints	0.1208	0.1090	0.1007	0.0785	0.0557	**0.0377**
baselines: constant 0.1208, Mondrian 0.0715, CQR 0.0446						

data; we select it on training data instead, matching the protocol applied to all other methods here. The implementation bounds the scale below at 0.2 and above at 5. We do not claim numerical reproduction of the published results.

Information boundaries. Training units are split once into a fitting subset and the held-out subset $\mathcal{D}_{\text{hold}}$. The point predictor $\hat{f}$ and the first-stage classifier are both fitted on the fitting subset. The Venn-Abers calibration maps are cross-fitted *within* $\mathcal{D}_{\text{hold}}$: its patients are partitioned into two halves, the map applied to each half is fitted on the other, and a map fitted on all of $\mathcal{D}_{\text{hold}}$ is applied to calibration and evaluation units. Every difficulty score g, the floor ϵ, the normalizing constant $\bar{g}_\epsilon$, and the evaluation stratum boundaries are computed on the eligible rows of $\mathcal{D}_{\text{hold}}$. No calibration or evaluation unit contributes to any of these steps.

Models. Every regressor, including the point predictor, the difficulty scores and the CQR quantile models, is a histogram gradient-boosted tree (scikit-learn's HistGradientBoostingRegressor). The first-stage classifier is HistGradientBoostingClassifier.

Splits and sampling. Within each repetition, units are permuted and assigned $50/30/20$ to training, calibration and evaluation, and training units are shuffled again and assigned $60/40$ to the fitting subset and $\mathcal{D}_{\text{hold}}$, all by fixed seeds. Training uses every eligible encounter of its patients; each calibration and evaluation patient contributes one encounter with a recorded change, sampled uniformly. Repetitions with fewer than 30 calibration or 15 evaluation units are skipped.

Groups and scales. Group thresholds are the 33.3rd and 66.7th percentiles of a mean-residual difficulty model on $\mathcal{D}_{\text{hold}}$; a value equal to a cut point joins the higher group. The same thresholds apply to every method and every score, including all configurations of Table 3. The freMTPL2 outcome is $\log(1 + \text{claim amount})$, and its interval lengths are on that scale. Every reported run produced finite interval lengths except for Mondrian, which returns infinite intervals for 0.60% and 5.66% of evaluated XYZ Health records at the 90% and 95% targets, and for none elsewhere.

C STRENGTH SELECTION

We swept w and inspected the outcome. We split the calibration units, freeze the candidate intervals equation 7 on one half, and select on the other.

Two rules are compared. An *empirical* rule takes the strength with the smallest observed CovGap. A *certifying* rule demands evidence that the choice is an improvement. Using simultaneous Clopper–Pearson intervals (Clopper & Pearson, 1934) $[l_g(w), u_g(w)]$ for each candidate's group coverages at total error $\delta/15$, with $\delta = 0.05$, across the five candidates $w \in \{0, 0.25, 0.50, 0.75, 1\}$ and three groups, it defines

$$L(w) = \tfrac{1}{|G|} \textstyle\sum_g \max\{l_g(w) - (1{-}\alpha),\ (1{-}\alpha) - u_g(w),\ 0\}, \tag{13}$$

$$U(w) = \tfrac{1}{|G|} \textstyle\sum_g \max\{|l_g(w) - (1{-}\alpha)|,\ |u_g(w) - (1{-}\alpha)|\}, \tag{14}$$

the smallest and largest values of CovGap consistent with the confidence intervals. It certifies an adaptive strength only when $U(w) < L(0)$: the candidate's worst plausible CovGap must beat constant width's best plausible one. Among qualifying candidates it takes the smallest $U(w)$, and both rules break ties toward smaller w.

Three conditions bound what the certificate means. The calibration units are split at patient level (policy level for freMTPL2) by a fixed, outcome-independent rule before eligibility filtering, and the evaluation set is not used. The intervals assume that selection units are conditionally independent and identically distributed within each group given the frozen candidates. And the certificate concerns CovGap on the selection distribution only: it does not extend the marginal guarantee of Section 2.3 to the selected strength.

D COHORTS AND DATA PREPARATION

Construction. Both clinical cohorts use the target of equation 1, with covariates restricted to information available before the dosing decision. Table 10 gives unit counts for every dataset.

Table 10: Datasets, units and matched-benchmark sample sizes (per split, means). Clinical benchmarks sample one eligible encounter per patient.

Dataset	Unit	Total units	Calibration	Evaluation	Nonzero rate
PPMI	patient	1,359	361	239	0.554
XYZ Health	patient	369	92	61	0.636
freMTPL2	policy	150,000	1,692	1,137	0.037
Synthetic	simulated unit	—	180 / 47	2,003	≈0.50

PPMI. The outpatient cohort is drawn from a longitudinal multi-site observational study of Parkinson's disease, with encounters sequenced per patient so each dose change is measured against that patient's previous recorded dose. Features are age, education, encounter index, days since the previous visit, the previous dose and its log, and indicators for whether the previous encounter changed the dose and whether the patient was on treatment. All are fixed or recorded before the dosing decision, so no same-visit measurement enters X. Splits are by patient and each calibration or evaluation patient contributes one sampled encounter, so evaluated units are one per patient. Encounters with a recorded change make up 55.4% of the cohort. PPMI reports LEDD using the conversion factors of (Tomlinson et al., 2010), which we also apply to XYZ Health, so doses are comparable across cohorts.

XYZ Health. The inpatient cohort is drawn from several hospitals in one health system and uses the same target as PPMI. Features are the recorded demographic and encounter-level fields available before the dosing decision. Splits are by patient on the same protocol, and doses are converted to LEDD with the same conversion factors. First encounters are dropped rather than treated as no change, and any rescaling of the outcome is fitted on training data alone.

freMTPL2. The frequency and severity tables are joined at the policy level. Policies with a positive claim count but no matched severity record are excluded as unresolved rather than zero-filled, since no severity row carries a zero amount. Each policy contributes one unit.

E SYNTHETIC DATA GENERATION

Each independent replication fixes three weight vectors $w_z, w_\mu, w_d \sim \mathcal{N}(0, I_6)/\sqrt{6}$. Units draw $X \sim \mathcal{N}(0, I_6)$, and Z is Bernoulli with $\mathbb{P}(Z{=}1 \mid X) = \{1 + \exp(-X^\top w_z)\}^{-1}$. When $Z = 1$,

$$Y = X^\top w_\mu + \sigma(X)\,\varepsilon, \qquad \sigma(X) = \exp\{\kappa\, d(X)\}, \qquad d(X) = \rho\,(X^\top w_z) + \sqrt{1-\rho^2}\,(X^\top w_d), \tag{15}$$

with ε a t_5 variable rescaled to unit variance. Here κ sets how strongly the noise scale varies across units and ρ how far that variation aligns with the classifier's linear predictor, so $\rho = 0.8$ makes first-stage signal informative about second-stage difficulty, $\rho = 0$ makes it irrelevant and $\rho = -0.8$ misleading. Outcomes are either signed or taken in absolute value plus 0.1. True difficulty is $\sigma(X)$, which is known, so the synthetic strata are tertiles of the true noise scale rather than of an estimate.

The frozen grid crosses $\kappa \in \{0, 0.6\}$, $\rho \in \{-0.8, 0, 0.8\}$, both sign structures and two calibration budgets, with 40 replicates each. Table 2 uses $\kappa = 0.6$ at the PPMI-sized budget, so 3 values of ρ times 2 sign structures times 40 replicates give 240 independent replications. Each independent replication draws 361 calibration units, of which 179.6 on average have a recorded change, and $4{,}000$ evaluation units, of which $2{,}001$ do; the XYZ-sized budget draws 92 calibration units with 46.5 eligible. Table 10 reports the eligible counts. Development data are drawn separately for fitting and for score training, so no calibration or evaluation unit informs either.

F DIRECTIONAL INFORMATION AND DOSE-SCALE INTERVALS

Coverage balance and length describe statistical behavior, but a clinician reads milligrams and asks which way the dose should move. This appendix examines two further properties of the same intervals, both of which qualify the allocation results.

Whether an interval indicates a direction. An interval containing zero is consistent with both an increase and a decrease. We measure the complement, the *zero-exclusion rate*, from the actual endpoints, so it applies to the asymmetric CQR intervals as well as the symmetric ones; empty intervals are not counted as informative and are reported separately. For a symmetric interval, $0 \in \widehat{C}_w(x)$ exactly when $|\hat{f}(x)| \le \hat{\lambda}_w \sigma_w(x)$. The point prediction stays the same as w varies, while both $\hat{\lambda}_w$ and σ_w change, so the rate can move with w and what we report below is measured.

On PPMI at the 80% target the overall rate changes little across configurations, ranging from 0.515 at $w{=}0.50$ to 0.528 for CQR, with 0.522 at $w{=}0$ and 0.520 at $w{=}1$. The subgroup rates do change. Moving from $w{=}0$ to $w{=}1$ raises the rate in the low-difficulty stratum from 0.264 to 0.367 and lowers it in the high-difficulty stratum from 0.564 to 0.481, so adaptation reallocates directional information between strata in the opposite direction to coverage. On XYZ Health the rates are far lower and are not constant across methods: at the 80% target they run from 0.000 for CQR to 0.019 at $w{=}1$, with 0.006 at $w{=}0$. Whatever the coverage properties on that cohort, few of its intervals indicate a direction. Table 8 in Appendix A reports every configuration and target. Zero-exclusion refers to the direction of the recorded dose change; it is not evidence that an indicated direction is clinically correct, safe, or actionable.

What the intervals look like in dose units. Applying the inverse of equation 1, $D = (D_{\text{prev}} + 1)e^y - 1$ with the lower endpoint floored at zero, converts each interval into a range of possible next doses. Under $w{=}0$ every record has the same length on the log scale, yet the implied dose range widens with the previous dose, from a median of 497 mg/day for records with no previously recorded dose to 560, 898 and 1551 mg/day for previous doses of 1–300, 300–600 and above 600 mg/day. Because the transformation is multiplicative in the previous dose, a fixed log-scale width becomes a wider range in milligrams for patients already on higher doses, before any adaptation is applied.

Adaptation is visible on this scale as well. The median implied dose range in the high-difficulty stratum grows from 564 mg/day at $w{=}0$ to 726 mg/day at $w{=}0.50$, while the low-difficulty stratum

falls from 1099 to 784 mg/day; the overall median rises only from 610 to 639 mg/day. The ordering across strata reflects the dose distribution within them rather than difficulty alone, since the low-difficulty stratum contains many high-dose records.

We report these as descriptive properties of the predictions. We do not define a clinical acceptability threshold, and improvements in coverage balance neither imply nor preclude improvements in directional usefulness, which depends on interval length relative to the predicted change.