When uncertainty quantification fails: structural vulnerabilities and corrections for selective prediction under clinical class imbalance

Walter Endrizzi[1,2*], Flavio Ragni[1], Raffaele Marchesi[1,3], Stefano Bovo[1], Monica Moroni[1], Giuseppe Jurman[1,4], Venet Osmani[5]

[1]Data Science for Health, Fondazione Bruno Kessler, Trento, Italy.
[2]Department of Cellular, Computational and Integrative Biology, University of Trento, Trento, Italy.
[3]Department of Mathematics, University of Pavia, Pavia, Italy.
[4]Department of Biomedical Sciences, Humanitas University, Milan, Italy.
[5]Digital Environment Research Institute, Queen Mary University of London, London, E1 1HH, United Kingdom.

*Corresponding author(s). E-mail(s): wendrizzi@fbk.eu;
Contributing authors: fragni@fbk.eu; rmarchesi@fbk.eu; sbovo@fbk.eu; mmoroni@fbk.eu; jurman@fbk.eu; v.osmani@qmul.ac.uk;

Abstract

Background. Selective prediction systems improve the safety of clinical artificial intelligence by deferring ambiguous cases to human experts while automating high-confidence decisions. However, this safety mechanism relies entirely on the premise that uncertainty scores accurately reflect clinical ambiguity. In prognostic tasks characterized by severe class imbalance, where the clinical outcome of interest is a minority, this assumption breaks down. Standard uncertainty measures cease to capture clinically aligned ambiguity; instead, they become a proxy for class membership, systematically penalizing minority patients regardless of their individual predictive clarity.

Methods. We characterize three structural failure modes driving this breakdown: the mismatch between standard entropy's mathematical peak and the clinical decision threshold; the asymmetric spatial penalty in margin-based uncertainty; and structural uncertainty divergence of minority predictions driven

by data scarcity. We propose three counter-mechanisms: an empirically validated *threshold-aware ambiguity score* that realigns maximum uncertainty with the clinical threshold, a *class-conditioned rejection* (CCR) mechanism ranking patients within their predicted trajectory rather than globally, and a class-stratified evaluation metric that prevents global risk scores from concealing automated minority exclusion.

Results. Across 56 synthetic conditions and three neurodegenerative cohorts spanning severe class imbalance to near-balanced conditions (multiple sclerosis, Parkinson's Disease, Alzheimer's Disease), our framework consistently maintained bounded, balanced risk across trajectories. In contrast, global metrics either collapsed minority Sensitivity or protected the minority at the expense of majority patients. An ablation study confirmed that applying CCR with standard entropy fails to protect the minority because it breaks within-class monotonicity when $\boldsymbol{\tau \neq 0.5}$. Combining CCR with threshold-aware metrics restores strict monotonicity, correctly identifies ambiguous patients, and safely mitigates minority exclusion.

Conclusions. Under clinical class imbalance, standard uncertainty quantification measures dataset sparsity rather than clinical ambiguity, creating a systemic illusion of safety. This bias is a correctable structural flaw. Clinical pipelines deploying selective prediction under imbalance should adopt threshold-aligned uncertainty measures and class-stratified evaluation. This reduces class-specific risk imbalance, supporting safer selective prediction under the evaluated conditions without silently excluding the most vulnerable patients.

Keywords: uncertainty, selective prediction, sample rejection, clinical AI, class imbalance, neurodegenerative disease, entropy, calibration, generalized risk-coverage curve

1 Introduction

Deployment of machine learning models in high-stakes clinical environments requires algorithmic systems that can reliably estimate their own limitations. In prognostic tasks, such as predicting the trajectories of neurodegenerative diseases [1–3], biological heterogeneity often makes definitive classification impossible. In these settings, the safety of an automated system depends on selective prediction: the mechanism by which an algorithm defers ambiguous cases to a human expert while automating high-confidence decisions [4].

Standard selective prediction relies on uncertainty quantification (UQ) metrics to rank patients for deferral. However, in real-world clinical prognosis, datasets are often highly imbalanced, with the class of highest clinical interest (such as rapid progressors) representing only a small minority. Under these conditions, standard UQ approaches may fail to provide reliable estimates of true predictive uncertainty; instead, they can become proxies for class membership. Because minority patients inherently lack dense feature representations, current uncertainty estimates become structurally biased [5, 6]. Therefore, algorithms learn to equate the minority class with high uncertainty, failing to disentangle triage-relevant ambiguity from dataset sparsity [7, 8].

This measurement failure is not a consequence of poor model training, but rather the result of distinct structural vulnerabilities inherent to standard post-hoc UQ metrics as they are commonly deployed in applied heuristic pipelines. The first vulnerability, the threshold-entropy mismatch, occurs because the mathematical peak of standard entropy (i.e. Shannon entropy) is rigidly anchored to a perfectly balanced decision threshold (0.50). This structural misalignment with the naturally shifted thresholds (τ) required for imbalanced data raises concerns about its appropriateness as an uncertainty measure [9]: when predicted probabilities fall near the balanced decision point, standard entropy may produce biased uncertainty estimates that reflect class imbalance rather than clinically aligned ambiguity. Consequently, clinically relevant minority patients are either erroneously automated or unnecessarily deferred due to an uncertainty calibration structure that is poorly aligned with the imbalanced decision space.

Although probability rescaling and threshold moving are well-established techniques for optimizing classification performance under severe imbalance [10], their implications for realigning uncertainty estimation remain largely unexplored.

A second vulnerability concerns the spatial penalty introduced by margin-based uncertainty. The Classification Margin partially addresses the entropy-threshold mismatch by measuring uncertainty relative to the operative decision threshold rather than to 0.50. However, it still penalizes predictions strictly by their absolute distance from the threshold. When the decision boundary is tightly compressed against the majority-class space, this can disproportionately flag confident majority-class patients for rejection. Margin-based UQ may appear to protect minority predictions, but it does so by systematically deferring confident majority-class cases, yielding an undesirable and clinically fragile trade-off. Although previous works have identified structural failure modes in margin-based confidence [11] and learning with rejection [12], these analyses assume relatively balanced decision spaces and do not account for the asymmetric spatial penalty introduced by extreme class imbalance.

Finally, this ranking failure is compounded by structural uncertainty divergence. Because of their underlying sparsity, minority predictions inherently carry systematically higher ensemble disagreement than majority predictions, even when evaluated at the exact same distance from the decision boundary. Recent research in long-tailed medical imaging corroborates this phenomenon, observing that rare classes naturally exhibit wider confidence intervals strictly as a function of their frequency [13].

Taken together, these mechanisms show that **standard global ranking instruments** are intrinsically limited under severe class imbalance, as they can **conflate decision-level uncertainty with class prevalence, decision-boundary geometry, and minority-class sparsity**.

Standard evaluation frameworks commonly fail to detect these vulnerabilities. Although recent literature has begun to identify the structural limitations of standard selective evaluation paradigms [14], the specific masking effect caused by extreme class imbalance remains unaddressed. As systems defer patients, safety is commonly evaluated using multi-threshold metrics such as the Area Under the Risk-Coverage Curve (AURC) [15] or its recent refinement, the Area Under the Generalized Risk-Coverage Curve (AUGRC) [16]. However, because these metrics aggregate risk across

the entire global population, an algorithm can mathematically improve its global score by systematically excluding the harder-to-predict minority class. In this way, standard evaluation can reward automated minority exclusion, masking algorithmic bias under the guise of improved overall safety.

Contributions. Addressing this gap requires moving beyond global ranking and toward class-stratified uncertainty measurement and evaluation. In this paper, we characterize the structural vulnerabilities of standard UQ under clinical class imbalance, propose a framework of well-motivated, empirically validated measurement and evaluation instruments, and demonstrate their efficacy across synthetic and real-world neurodegenerative prognostic tasks. We validate the proposed framework across 56 controlled synthetic conditions and three real-world neurodegenerative patient cohorts comprising multiple sclerosis, Parkinson's Disease, and Alzheimer's Disease. Specifically, we make the following contributions:

- **Threshold-aware ambiguity score (H_τ):** We demonstrate that standard Shannon entropy fails under class-conditioned rejection (CCR) when $\tau \neq 0.5$ by breaking within-class monotonicity. To restore correct ordinal ranking, we introduce H_τ, an empirically validated geometric construction. By design, H_τ operates as a non-linear, information-theoretic alternative to the linear decision margin, satisfying three key structural properties: continuity across the probability space, strict monotonicity within each predicted class, and exact reduction to standard Shannon entropy at parity ($\tau = 0.5$).
- **CCR (Class-Conditioned Rejection):** A class-conditioned rejection mechanism that isolates intra-class ambiguity, establishing a structural, mechanical mechanism to prevent minority exclusion.
- **ccAUGRC (Class-Conditioned Area Under the Generalized Risk-Coverage Curve):** An evaluation decomposition that prevents risk-masking by quantifying the generalized risk-coverage trade-off independently for each clinical trajectory.

1.1 Related Work

Selective prediction foundations. The foundations of selective prediction have been extensively formalized to optimize the trade-off between coverage and accuracy [17–19]. While these seminal theoretical frameworks natively accommodated the decision threshold as a free parameter, modern applied pipelines often take a more rigid approach. Contemporary systems successfully leverage ensemble-based or Bayesian methods to generate predictive distributions for deferral [20], but they frequently default to unaligned ranking metrics such as standard Shannon entropy. Because this post-hoc ranking predominantly assumes balanced class priors, and because disentangling aleatoric from epistemic uncertainty remains an open challenge [21, 22], standard selective classification frequently magnifies performance disparities across distinct subgroups [23, 24]. This vulnerability underscores the critical need for class-aware triage strategies.

Entropy limitations and threshold-awareness. While base-classifier training frequently addresses imbalance via loss-weighting or resampling [25, 26], and recent

approaches integrate learning to reject directly with long-tail learning paradigms [27], the calibration of post-hoc uncertainty under imbalance remains critically underexplored. Existing solutions largely focus on probabilistic recalibration or multimodal data acquisition [28], but lack structural geometric corrections for the threshold-entropy mismatch and the asymmetric spatial penalties introduced by shifted decision boundaries, as identified in recent theoretical work [5, 9].

Evaluation corrections. The Generalized Risk-Coverage curve and AUGRC have become standard for evaluating selective models across dynamic operating points [16]. While methods exist to improve multi-threshold metrics, and concurrent work has recently introduced class-averaged risk-coverage measures using empirical quantiles for multiclass settings [29], these approaches primarily address triage routing rather than the underlying signal geometry. The literature requires a dedicated decomposition of global risk-coverage to account for the specific risk displacement artifacts caused by extreme clinical class imbalance, a gap this work addresses via ccAUGRC.

Conformal prediction and risk control. An alternative, prominent approach to managing uncertainty under imbalance is conformal prediction (CP) and distribution-free risk control [30]. Class-conditional conformal prediction (CCCP) explicitly addresses the problem of bounding risk within subpopulations by generating adaptive prediction sets that satisfy a user-defined coverage guarantee for each class independently [31, 32]. While CCCP and distribution-free risk control represent the standard for theoretical coverage guarantees, they solve a fundamentally different operational problem (bounding error rates via variable-sized prediction sets) than selective classification (managing a fixed clinical deferral bandwidth via fixed rejection quotas).

2 Methods

2.1 Problem formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a binary clinical prognostic dataset where $y_i \in \{0, 1\}$ indicates the absence or presence of disease progression. In imbalanced clinical settings, the class prior $\pi = P(y = 1)$ is substantially less than 0.5, reflecting the natural prevalence of the prognostic outcome in the population. A probabilistic classifier $f\colon \mathcal{X} \to [0, 1]$ produces a calibrated predicted probability $p_i = f(x_i)$. Clinical decisions are made using a decision threshold $\tau \in (0, 1)$ derived from the population prior [33, 34]: a patient is predicted as a progressor if $p_i \geq \tau$ and as stable otherwise. In standard practice, $\tau = \pi$, ensuring that the classifier's operating point is aligned with the clinical prevalence of the outcome rather than the mathematical center of the probability space.

A *selective prediction system* augments this classifier with a rejection function $g\colon \mathcal{X} \to \{0, 1\}$, where $g(x_i) = 0$ defers patient i to a clinician for manual review. Deferral is operationalized by thresholding an uncertainty score u_i: predictions exceeding a predefined limit are deferred to human experts. The retained cohort $\mathcal{A} = \{i : g(x_i) = 1\}$ is processed automatically; the deferred cohort $\mathcal{R} = \{i : g(x_i) = 0\}$ receives human review.

The central premise of selective prediction is that the uncertainty score u_i correctly identifies algorithmic ambiguity relative to the clinical decision threshold. In the

following subsections, we formalize three structural vulnerabilities where this premise breaks down under clinical class imbalance, and subsequently define the algorithmic framework proposed to address them.

2.2 Structural vulnerabilities of standard uncertainty quantification

We characterize three distinct mathematical failure modes that arise when standard uncertainty metrics are applied to imbalanced clinical prognostic tasks, which are conceptually illustrated in Figure 1.

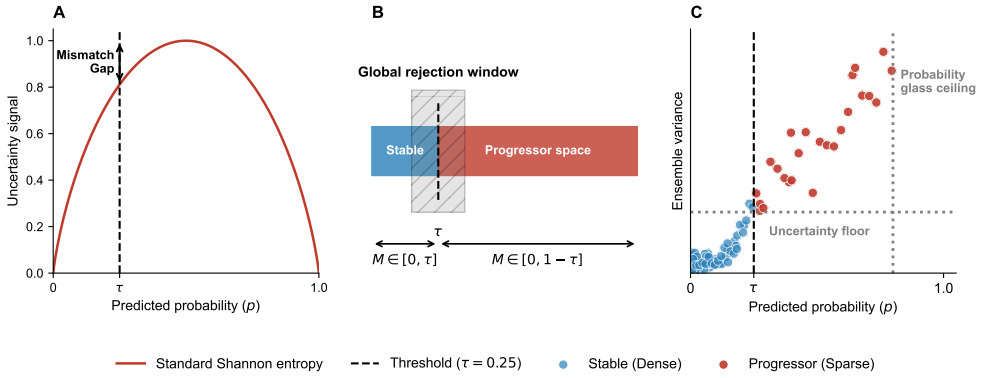


Fig. 1 Structural vulnerabilities of standard uncertainty quantification under clinical class imbalance. The three panels conceptualize the mathematical failure modes of global uncertainty metrics in a binary classification setting when the clinical decision threshold (τ) is shifted from 0.50. **(A) Threshold-entropy mismatch.** Standard Shannon entropy (red curve) is mathematically anchored to a perfectly balanced decision space. Evaluated at a naturally shifted clinical threshold (e.g., $\tau = 0.15$), a critical mismatch gap emerges. Maximum mathematical entropy fails to align with maximum clinical ambiguity ($p = \tau$), causing standard entropy to operate as a distance metric from 0.5 rather than a true measure of clinical uncertainty. **(B) The spatial penalty.** Anchoring uncertainty to the absolute decision margin $M(p) = |p - \tau|$ exposes the asymmetric geometry of the probability space. The majority class is structurally compressed into a narrow margin ($M \in [0, \tau]$), whereas the minority class lies in a vastly larger space ($M \in [0, 1 - \tau]$). Applying a single global rejection window systematically distorts the retained cohort by disproportionately penalizing or entirely excluding the compressed majority class. **(C) Structural uncertainty divergence.** An idealized illustration demonstrating the structural vulnerabilities introduced by extreme class imbalance. A combination of aleatoric noise and epistemic sparsity compresses minority predictions (red) into a bounded space, imposing a theoretical probability glass ceiling ($g < 1$, dotted line). Forced into this compressed regime, the minority class theoretically exhibits a structural uncertainty floor (dashed line), meaning their ensemble disagreement remains artificially elevated above majority patients (blue) even at identical mathematical distances from the decision boundary (τ, solid line). We empirically validate the real-world manifestation of this structural asymmetry in Section 3.

2.2.1 Vulnerability 1. Threshold-entropy mismatch

The most widely used uncertainty metric in selective prediction is predictive entropy, which in binary classification reduces to Shannon binary entropy [35, 36]:

$$H(p) \;=\; -p\log_2 p \;-\; (1-p)\log_2(1-p). \tag{1}$$

A fundamental limitation of $H(p)$ is that it achieves its global maximum exclusively at $p = 0.5$, independently of the clinical decision threshold τ. This follows directly from setting the derivative to zero: $\frac{dH}{dp} = -\log_2 p + \log_2(1-p) = 0$ only when $p = 0.5$, confirming a unique global maximum at the mathematical center of the probability space.

Consequently, a patient predicted exactly at the decision boundary $p = \tau$, the point of maximum clinical ambiguity, achieves an entropy $H(\tau) < H(0.5) = 1$ for any $\tau \neq 0.5$. Conversely, a patient confidently predicted at $p = 0.5$ reaches maximum entropy regardless of τ. For example, in a disease with prevalence $\pi = 0.10$, confident progressor predictions near $p = 0.50$ yield maximal entropy. Under these conditions, standard entropy operates as a distance metric from $p = 0.5$ rather than a measure of triage-relevant ambiguity, disproportionately flagging minority patients for deferral. This vulnerability is naturally instantiated by standard global Shannon entropy (H_{Total}).

2.2.2 Vulnerability 2. Spatial penalty

A natural correction for Vulnerability 1 is to replace entropy with the absolute decision margin $M(p) = |p - \tau|$, which anchors uncertainty to the clinical threshold τ rather than to 0.5. This eliminates the entropy mismatch, but introduces a distinct structural anomaly arising from the asymmetric geometry of the probability space around τ.

For patients predicted stable ($p < \tau$), the margin is strictly bounded within $[0, \tau]$: the most certain stable patient achieves at most $M = \tau$. For patients predicted as progressors ($p > \tau$), the margin spans $[0, 1-\tau]$: the most certain progressor can achieve $M = 1-\tau$. When $\tau < 0.5$ the two classes occupy structurally different certainty spaces. With $\tau = 0.10$, for instance, the entire stable class is compressed within $M \in [0, 0.10]$, while the progressor class has access to $M \in [0, 0.90]$.

While unequal margin supports do not strictly guarantee unequal retention, as this ultimately depends on the density of the class-conditional distributions—under adequate class separability, this asymmetry makes it impossible to find a single global threshold δ that fairly partitions both classes. If $\delta > \tau$, no stable patient can achieve $M > \delta$, because their entire margin space lies below δ, leading to deferral of the entire stable class (majority patients). If $\delta \leq \tau$, some stable patients can be retained, but the threshold simultaneously retains every progressor whose margin exceeds δ, since their margin space is far larger. In either regime, the retained cohort is systematically distorted: the stable class either disappears entirely or is underrepresented relative to its clinical prevalence. This pathology is naturally instantiated by global margin thresholding.

2.2.3 Vulnerability 3. Structural uncertainty divergence

Even if an uncertainty metric is perfectly anchored to the decision threshold τ, a third structural vulnerability persists: structural uncertainty divergence. This phenomenon causes global uncertainty rankings to systematically defer minority patients through two compounding effects.

First, minority predicted probabilities are compressed into a restricted range $[\tau, g]$ with a structural upper bound $g < 1$, which we term the probability glass ceiling. Two distinct factors drive this ceiling. Biological heterogeneity in the minority class introduces genuine, irreducible aleatoric uncertainty, which inherently limits prediction confidence. Concurrently, data sparsity prevents the model from resolving confident minority-class boundaries, driving up ensemble disagreement (often used as a proxy for epistemic uncertainty). This sparsity-driven uncertainty is further amplified by standard machine learning pipeline mechanics. For instance, ensemble aggregation inherently smooths extreme probabilities. Furthermore, post-hoc probability calibration (such as Platt scaling [37]) fitted on imbalanced data heavily weights the majority prior; consequently, the calibrator acts conservatively, systematically penalizing and pulling down high-confidence minority predictions. Together, these factors artificially compress the minority probability range into a glass ceiling, preventing the model from assigning high-confidence minority predictions regardless of how clear the patient's actual prognosis might be.

Second, because minority predictions are constrained to a compressed range $[\tau, g]$ rather than covering the entire interval $[\tau, 1]$, they are functionally compressed relative to majority predictions, which the model can confidently assign to the stable domain near $p \approx 0$. This creates an uncertainty floor: even the most certain minority prediction carries more absolute uncertainty than a typical majority prediction. This elevation is a modeling artifact rather than a reflection of the patient's individual clinical ambiguity: it arises from the model's data limitations and is shared by all minority patients regardless of their true prognostic clarity. Ultimately, due to this floor, the global ranking by any absolute uncertainty score preferentially rejects minority patients, as their absolute scores remain structurally elevated above those of majority patients at equivalent clinical distances from the decision threshold.

2.3 Proposed framework

Because the three vulnerabilities stem from distinct layers of the selective prediction pipeline, i.e. the metric formulation, the probability geometry, and the learning dynamics, they cannot be resolved by a single algorithmic adjustment. Instead, we propose a cohesive framework that intervenes sequentially: first at the signal layer to correct the metric formulation, second at the triage layer to correct the sorting geometry, and finally at the evaluation layer to accurately measure clinical risk.

2.3.1 Threshold-aware ambiguity score

To correct the entropy mismatch (Vulnerability 1) at the signal layer, we propose H_τ, constructed by applying a piecewise linear rescaling of the predicted probability that

maps the clinical decision boundary to the entropy peak:

$$\tilde{p} \;=\; \begin{cases} \dfrac{p}{2\tau} & \text{if } 0 \le p \le \tau, \\[2ex] 0.5 + \dfrac{p-\tau}{2(1-\tau)} & \text{if } \tau < p \le 1. \end{cases} \tag{2}$$

The final threshold-aware ambiguity score is then computed as:

$$H_\tau(p) \;=\; -\tilde{p}\,\log_2 \tilde{p} \;-\; (1-\tilde{p})\,\log_2(1-\tilde{p}). \tag{3}$$

The transformation $p \mapsto \tilde{p}$ maps $[0, \tau]$ to $[0, 0.5]$ and $(\tau, 1]$ to $(0.5, 1]$ monotonically, ensuring $\tilde{p}(\tau) = 0.5$. Consequently, $H_\tau(p)$ attains its maximum strictly at the clinical boundary $p = \tau$. The metric possesses two core properties: first, it reduces to standard Shannon entropy when $\tau = 0.5$ ($H_\tau \equiv H$), acting as a strict generalization. Second, the rescaling is monotone-preserving within each predicted class, ensuring patients closer to τ receive higher H_τ scores than patients further from τ within the same trajectory. The mapping is continuous at $p = \tau$ but not differentiable there unless $\tau = 0.5$: the left-hand derivative is $1/(2\tau)$ and the right-hand derivative is $1/(2(1-\tau))$, producing a sharp peak at the decision boundary for any $\tau \neq 0.5$. Because the framework uses ordinal ranking of H_τ scores this non-smoothness does not affect the algorithm's operation.

Due to this strict within-class monotonicity, any rejection mechanism applied independently to each side of the threshold will induce the exact same ordinal ranking as a simple linear margin ($|p - \tau|$). Despite this operational equivalence for triage, we utilize H_τ as our primary signal for its non-linear Sensitivity and information-theoretic interpretability. Unlike a linear margin which evaluates probability shifts linearly across the entire probability space, the logarithmic nature of H_τ penalizes predictions more severely when they fall near the operational threshold τ. This provides a probabilistically bounded measure of missing information that aligns with standard information-theoretic definitions of uncertainty.

H_τ resolves Vulnerability 1 by anchoring maximum ambiguity to the clinical decision boundary. Applied globally, however, it does not resolve Vulnerability 3: because epistemic sparsity restricts minority predictions to a compressed probability range, their absolute H_τ scores are subject to an artificially elevated uncertainty floor. Consequently, even at comparable clinical distances from τ, minority patients remain structurally penalized under a global ranking, preferentially triggering their deferral. This motivates the class-conditioned rejection mechanism described in the following section.

2.3.2 Class-Conditioned Rejection

To address the spatial penalty (Vulnerability 2) and structural uncertainty divergence (Vulnerability 3), we propose CCR (Class-Conditioned Rejection). CCR partitions the cohort by predicted class and ranks patients by uncertainty within each partition independently (Algorithm 1).

Algorithm 1 Class-Conditioned Rejection (CCR)

Require: Cohort $\mathcal{D}$, predicted probabilities p, uncertainty signal u, rejection rate $r \in [0, 1)$, decision threshold τ

Ensure: Retained cohort $\mathcal{A}$

$\mathcal{D}_0 \leftarrow \{i \in \mathcal{D} \mid p_i < \tau\}$ ▷ Partition predicted stable

$\mathcal{D}_1 \leftarrow \{i \in \mathcal{D} \mid p_i \geq \tau\}$ ▷ Partition predicted progressor

Sort $\mathcal{D}_0, \mathcal{D}_1$ in descending order by uncertainty signal u

$k_0 \leftarrow \lfloor r \times |\mathcal{D}_0| \rfloor$ ▷ Rejection quota for stable class

$k_1 \leftarrow \lfloor r \times |\mathcal{D}_1| \rfloor$ ▷ Rejection quota for progressor class

$\mathcal{R}_0 \leftarrow$ top k_0 patients from $\mathcal{D}_0$

$\mathcal{R}_1 \leftarrow$ top k_1 patients from $\mathcal{D}_1$

$\mathcal{A} \leftarrow \mathcal{D} \setminus (\mathcal{R}_0 \cup \mathcal{R}_1)$ ▷ Merge remaining patients

return $\mathcal{A}$

By rejecting fraction r from each predicted class independently, CCR ensures the preservation of the predicted class proportions (up to a finite-sample rounding error) in the retained cohort. While this structural class-preservation prevents the systematic depletion of a predicted trajectory, it is equally important to examine how this mechanism filters misclassifications (false negatives and false positives) within these isolated queues.

In the $\mathcal{D}_0$ queue (predicted stable), false negatives typically lie near the decision boundary ($p \approx \tau$). Because all probabilities in $\mathcal{D}_0$ are strictly bounded below τ, even standard entropy increases monotonically toward this boundary. Consequently, simply isolating the $\mathcal{D}_0$ queue forces both standard and threshold-aware metrics to correctly assign the highest uncertainty to boundary false negatives, pushing them to the front of the rejection list.

Conversely, reliable ranking in the $\mathcal{D}_1$ queue (predicted progressor) mathematically necessitates a threshold-aware metric. As established, when $\tau < 0.5$, the mathematical peak of standard entropy (0.5) falls inside the $\mathcal{D}_1$ space. Applying standard entropy here breaks within-class monotonicity, mistakenly assigning maximum uncertainty to confident predictions near $p \approx 0.5$ rather than to boundary false positives. Threshold-aware metrics, such as H_τ or the decision margin, resolve this by anchoring peak uncertainty exactly at τ. Thus, while CCR provides the macro-level mechanism for class preservation, it must be paired with a τ-centered metric to correctly filter boundary misclassifications across both ordinal queues. Furthermore, by ranking ordinally within the predicted trajectory, CCR bypasses the uncertainty floor: minority patients are compared only to other minority patients, neutralizing their structurally elevated absolute uncertainty.

2.3.3 Class-conditioned Area Under the Generalized Risk-Coverage Curve (ccAUGRC)

Standard evaluation metrics for selective prediction compute performance globally across all patients. Under class imbalance, this aggregation introduces a structural

evaluation artifact caused by inter-class substitution: a triage mechanism can artificially minimize its overall risk by systematically replacing minority patients (whose epistemic compression, Vulnerability 3, produces structurally higher error rates) with lower-risk majority patients in the accepted cohort. Global evaluation is therefore blind to whether safety is actually maintained across distinct patient trajectories.

To expose and resolve this substitution artifact, we adapt the recently formalized Area Under the Generalized Risk-Coverage curve (AUGRC; [16]). While standard AUGRC defines risk at a global coverage c as the fraction of total errors in the accepted cohort over the original population N, we compute this trade-off strictly within each class subpopulation $k \in \{0, 1\}$. Let $c_k \in [0, 1]$ denote the fraction of class-k patients retained at a given rejection rate r, and E_k the corresponding number of misclassified class-k patients. The class-conditioned AUGRC (ccAUGRC) is defined as:

$$\text{ccAUGRC}_k \;=\; \int_{c_k(r_{\max})}^{1} \frac{E_k}{N_k}\, \mathrm{d}c_k, \tag{4}$$

where N_k is the total number of class-k patients. The formulation is structurally identical to AUGRC but substitutes the global population denominator N with the class-specific denominator N_k, and the global coverage c with the class-specific coverage c_k. The integral is approximated by the trapezoidal rule over the discrete pairs $\{(c_k(r),\, E_k(r)/N_k)\}_{r\in[0,\, r_{\max}]}$, sorted by increasing c_k; in all experiments $r_{\max} = 0.80$.

ccAUGRC is not a new metric but a class-stratified application of AUGRC using the same integration machinery. Reporting ccAUGRC for both classes alongside global AUGRC allows a system that achieves low global AUGRC through automated minority exclusion to be distinguished from one that maintains bounded risk across both trajectories simultaneously.

2.4 Datasets

2.4.1 Real-World clinical cohorts

We validate our framework on three neurodegenerative disease cohorts spanning distinct imbalance regimes. Each cohort represents an independent prognostic task with disease-specific features, clinical endpoints, and population characteristics. The intentional heterogeneity across feature spaces and clinical endpoints is a design choice: we are not benchmarking a classifier but validating an evaluation framework, which should generalize across task characteristics and vary predictably with the class prior π. In all three cohorts the decision threshold τ is set equal to the observed class prior π, consistent with standard clinical practice and with the problem formulation in Section 2.1.

- **Multiple sclerosis (MS, severe imbalance, $\pi = 0.13$).** $N = 719$ patients from a multicentric Italian cohort described in full detail in Poretto et al. [1]. Briefly, patients were recruited from three Italian MS centers (Careggi University Hospital, Florence; IRCCS Ospedale Policlinico San Martino, Genoa; ASUIT, Trento) and diagnosed according to the 2010 McDonald criteria. The outcome variable is Progression Independent of Relapse Activity (PIRA) at year 3, defined as EDSS

worsening in the absence of relapse activity [38]. Minority class: progressors ($N = 92$, 12.8%). Predictive features cover demographic, clinical, and MRI-derived measures collected at baseline and two-year follow-up. Ethics approval and full preprocessing details are reported in [1].

- **Parkinson's Disease (PD, moderate imbalance, $\pi = 0.39$).** $N = 247$ patients from a multicentric Italian cohort (NeuroArtP3 project) described in full detail in Endrizzi et al. [39]. Briefly, patients were recruited from two Northern Italian sites (APSS, Trento; IRCCS San Martino Hospital, Genoa) and diagnosed according to standard clinical criteria. The outcome variable is the presence of Levodopa-induced dyskinesia (LID) within three years from the baseline visit. For this analysis, the predictive task focuses on forecasting future dyskinesia status independently of the patients' baseline motor complications. Minority class: patients with future LID ($N = 96$, 39.0%). Predictive features encompass baseline clinical-demographic data, core motor and non-motor assessments (e.g., MDS-UPDRS III, Hoehn & Yahr), and detailed medication history. Ethics approval and full preprocessing details are reported in [39]. Note that while the baseline cohort comprises 247 patients, the internal stratification logic of the cross-validation pipeline drops one patient due to uneven group splits; consequently, all empirical evaluations accurately reflect the 246 patients evaluated in the test folds.
- **Alzheimer's Disease (AD, near-balanced, $\pi = 0.54$).** $N = 314$ patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI; adni.loni.usc.edu). The predictive task is the conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease within 36 months of baseline assessment. Class 1: MCI-to-AD converters ($N = 170$, 54.1%). Predictive features encompass baseline cognitive assessments (e.g., ADAS-Cog, MMSE) and demographic information. Complete details regarding cohort selection, inclusion/exclusion criteria, and data preprocessing for this specific analytic sample are provided in Appendix A.

The three cohorts span $\pi \in \{0.13, 0.39, 0.54\}$, covering the severe, moderate, and near-balanced imbalance regimes. This range was selected to provide systematic coverage of the operating conditions under which the framework predicts distinct effects.

2.4.2 Synthetic data

Synthetic data are used to establish a controlled phase diagram of standard UQ failure as a function of class prior π and class separability d, providing a reference frame against which the real-world results are interpreted. While these synthetic distributions assume Gaussian normality to isolate mathematical vulnerabilities, the subsequent evaluation on real-world clinical cohorts serves to explicitly validate these failure modes under highly non-Gaussian, non-parametric conditions.

Each synthetic condition is defined by two parameters. The class prior $\pi \in \{0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50\}$ determines the proportion of minority patients and simultaneously serves as the decision threshold (τ). The separability $d \in \{0.5, 1.0, 2.0, 3.0\}$ is the mean separation between classes in standard deviation units (Cohen's d), ranging from heavily overlapping to well-separated class distributions.

Features ($M = 20$ per patient) are generated from class-conditional Gaussian distributions with an exponentially decaying signal across features:

$$x^{(j)} \sim \begin{cases} \mathcal{N}(0,1) & \text{class 0} \\ \mathcal{N}(d \cdot e^{-0.5j}, 1) & \text{class 1} \end{cases} \tag{5}$$

where $j \in \{0, 1, \ldots, M-1\}$ is the feature index. The exponential decay ensures that the classification signal is concentrated in a smaller subset of features, avoiding trivially separable tasks at high d. Two sample sizes are used ($N \in \{300, 1000\}$). The complete grid of $7 \times 4 \times 2 = 56$ conditions is processed through the same ensemble calibration pipeline described in the next sections.

2.5 Experimental setup

2.5.1 Asynchronous ensemble pipeline

All experiments use a consistent asynchronous ensemble pipeline (Figure 2). The pipeline is identical across all three cohorts and all synthetic conditions, ensuring that performance differences reflect the uncertainty signal and rejection mechanism rather than modeling choices. This workflow is used to produce a single calibrated mean predicted probability for each patient who falls in the test set during model testing and can be decomposed into four macro stages.

Stochastic partitioning. For each of 30 bootstrap iterations, the dataset is split into 80% training and 20% test sets using stratified random sampling with a unique random seed per iteration, preserving the natural class distribution.

Independent hyperparameter optimization. Five heterogeneous base classifiers are trained independently per iteration: Gradient Boosted Trees (XGBoost), Random Forest, Extra Trees Classifier, Logistic Regression, and Support Vector Classifier with RBF kernel. Each classifier undergoes independent randomized grid search cross-validation ($n_{\text{iter}} = 20$, 3-fold CV, 5 repeats) optimizing the Area Under the Precision-Recall Curve (AUPRC), chosen as the optimization target because it is sensitive to minority class performance and appropriate for imbalanced datasets.

Post-hoc calibration. To ensure predicted probabilities reflect reliable estimates of class membership, each trained classifier is independently calibrated using Platt scaling. To prevent data leakage, the scaling model is fitted on a held-out calibration fold within the same training iteration. This calibration is applied prior to ensemble aggregation to avoid score compression artifacts.

Patient-specific prediction pool. For each patient appearing in a test set, calibrated predictions from all models across all iterations in which that patient was in the test set are aggregated. The ensemble consensus μ_i (mean) and σ_i (standard deviation) provide patient-specific estimates of the predicted probability and its variability across the ensemble, respectively. The ensemble disagreement σ_i^2 provides a measure of predictive uncertainty, which serves as a proxy for epistemic uncertainty and is used in the asymmetry test (Section 3.2.1).

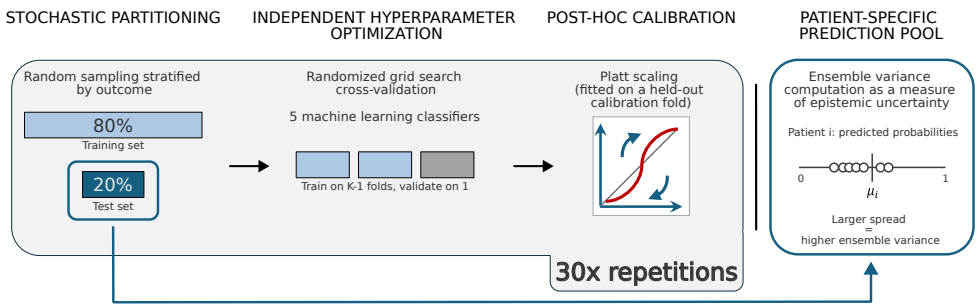


Fig. 2 Asynchronous ensemble pipeline. For each of 30 bootstrap iterations, the dataset was split into stratified 80% training and 20% test partitions while preserving the natural class distribution. Within each iteration, five heterogeneous machine learning classifiers were independently optimized using randomized grid-search cross-validation and subsequently calibrated by Platt scaling on a held-out calibration fold to obtain reliable predicted probabilities. For each patient appearing in a test set, calibrated predictions across classifiers and iterations were aggregated into a patient-specific prediction pool. The mean prediction μ_i represents the ensemble consensus, while the variance σ_i^2 quantifies prediction variability across the ensemble and is used as a proxy for epistemic uncertainty. No external resampling strategy was applied, ensuring that the resulting uncertainty estimates reflect model behavior under the natural class imbalance.

No external data resampling techniques (e.g., SMOTE, random undersampling) were applied to alter the natural class distribution of the training folds. While algorithmic cost-sensitive learning (e.g., class-weighted loss functions) was included in the hyperparameter search space to aid minority class boundary detection, post-hoc Platt scaling ensures that the final probabilities remain calibrated to the natural prior τ regardless of the training-time weighting scheme. Complete information on the computational environment and hyperparameter optimization can be found in the Appendix B.

2.5.2 Baseline and Control Methods

Six uncertainty quantification and rejection methods are compared across all experiments (Table 1). H_{Total} and Margin are established baselines. H_τ (Global) applies the proposed ambiguity score without the class-conditioned rejection mechanism. H_{Total} + CCR applies standard Shannon entropy within the class-conditioned framework, serving as a critical baseline to demonstrate the mathematical necessity of threshold-aware scoring when $\tau \neq 0.5$. H_τ + CCR is the proposed dual architecture. Random + CCR is a control condition that replaces the H_τ signal with random within-class ordering while retaining the class-conditioned rejection mechanism, enabling decomposition of the two components' independent contributions. Finally, we acknowledge Margin + CCR as a theoretical baseline, utilizing the linear decision margin ($|p_i - \tau|$) as the uncertainty signal within the CCR framework. As established in Section 2.3.1, both the margin and H_τ are strictly monotonic with respect to the distance from τ on either side of the decision boundary. Because Algorithm 1 relies exclusively on within-class ordinal ranking, applying Margin + CCR selects the exact same

patients for rejection at every threshold as H_τ + CCR, resulting in zero empirical disagreements. Consequently, the explicit Margin + CCR baseline produces an empirical rejection curve that perfectly overlaps with H_τ + CCR. To prevent visual redundancy in the figures and duplicate rows in the performance tables, this baseline is omitted from the reported results, though we fully acknowledge its operational equivalence.

Table 1 Uncertainty methods compared across all experiments. Summary of the evaluated quantification and rejection strategies. The proposed architecture (H_τ + CCR) is isolated against standard global baselines (Margin, H_{Total}), an ablation of the rejection mechanism (H_τ), a standard entropy CCR baseline (H_{Total} + CCR) to demonstrate the necessity of threshold-aware scoring, and a randomized control (Random + CCR) to isolate the mechanical benefit of class preservation.

Method	Uncertainty signal	Rejection
H_{Total} (Global)	Shannon entropy $H(p_i)$	Global
Margin (Global)	Decision margin $\lvert p_i - \tau \rvert$	Global
H_τ (Global)	Ambiguity Score $H_\tau(p_i)$	Global
H_{Total} + CCR	Shannon entropy $H(p_i)$	Class-conditioned
H_τ + CCR	Ambiguity Score $H_\tau(p_i)$	Class-conditioned
Random + CCR	Random	Class-conditioned

2.5.3 Evaluation protocol and statistical analysis

To capture the dynamic risk-coverage trade-off, all methods are evaluated across the full rejection rate range $r \in [0, 0.80]$ in steps of 0.05. Performance is tracked via three class-stratified metrics: Sensitivity (the primary safety metric for progressive diseases), Specificity, and the Area Under the Precision-Recall Curve (AUPRC) in the retained cohort.

To formally quantify these trajectories in the clinical cohorts, we compute three summary metrics. First, we calculate *Sensitivity stability*, defined as the slope of target-class Sensitivity over the 0%–50% rejection range estimated via linear regression. Because the 0%-50% range captures the most plausible deployment scenarios, a negative slope here indicates active target-class exclusion, whereas a slope near zero indicates stable protection. Second, to empirically quantify the risk-masking effect identified in standard evaluations, we compute the class-conditioned AUGRC (*ccAUGRC*) for both progressor and stable trajectories independently, integrated over the full rejection range. Finally, standard *Global AUGRC* is reported to demonstrate how class-specific failures are hidden within traditional baseline comparisons.

Due to the limited target-class sizes under sever imbalance inherent to clinical data, computing AUGRC on individual test folds yields numerically unstable artifacts. To ensure evaluation stability, we assess the final aggregated ensemble via patient-level bootstrapping. Point estimates are reported with 95% bootstrap confidence intervals (CIs) computed from $1,000$ patient-level resamples of the aggregated predictions. Pairwise statistical comparisons against the proposed H_τ + CCR framework utilize a

bootstrap sign test. To rigorously control the family-wise error rate (FWER) across the multiple baseline comparisons, we apply the Holm-Bonferroni step-down procedure independently for each disease cohort and evaluation metric. A performance difference is considered statistically significant only if its confidence interval strictly excludes zero and the comparison satisfies the adjusted Holm-Bonferroni threshold at family-wise $\alpha = 0.05$.

For the synthetic experiments, the analysis focuses on mapping performance across the grid of π and d conditions to establish the phase diagram. We report Sensitivity at a fixed 30% rejection rate as the primary phase diagram metric, alongside the Sensitivity stability slope to delineate regimes of active minority deferral. Full rejection curve panels for all 56 synthetic conditions are provided in the Supplementary Material.

3 Results

3.1 Synthetic validation

To isolate and empirically validate the structural vulnerabilities of standard uncertainty quantification (Vulnerabilities 1 and 2), we first evaluated the frameworks across 56 highly controlled synthetic conditions. This allowed us to map the precise boundaries of algorithmic failure without the confounding effects of clinical biological heterogeneity. Empirical confirmation of Vulnerability 3 (structural uncertainty divergence) is provided in Section 3.2.

3.1.1 Phase diagram of standard UQ failure

To establish the failure landscape, we tracked minority class Sensitivity (at a fixed 30% rejection rate) across a grid of varying class priors ($\pi \in [0.05, 0.50]$) and class separability ($d \in [0.5, 3.0]$). The results for all conditions (with $N = 1000$ samples) are shown in Figure 3. Standard H_{Total} maintains high Sensitivity only when classes are easily separable ($d \geq 2.0$) or near-balanced ($\pi \geq 0.40$). However, under severe imbalance ($\pi \leq 0.20$) and overlapping distributions ($d \leq 1.0$) standard UQ undergoes a complete collapse. At $\pi = 0.10$ and $d = 1.0$ baseline minority Sensitivity drops to 0.11, indicating that standard entropy systematically excludes nearly all minority patients from the retained cohort. Under the most severe conditions ($\pi = 0.05, d = 2.0$), it yields a Sensitivity of exactly 0.00. The proposed H_τ + CCR framework successfully rescues minority coverage within this failure zone. By anchoring the entropy peak to τ (set to π) and ranking within trajectories, the framework recovers up to +0.99 Sensitivity at severe imbalance ($\pi = 0.10, d = 2.0$). Crucially, this architectural correction scales proportionally with the degree of imbalance, converging perfectly at parity: when the natural prior is balanced ($\pi = 0.50$), the difference between the proposed framework and the baseline is exactly zero. Refer to Appendix C for all conditions with $N = 300$ samples.

3.1.2 Sensitivity stability scaling law

To capture the dynamic failure of standard UQ, we evaluated the stability of Sensitivity across the continuous range of class priors (Figure 4A). Our analysis exposes a

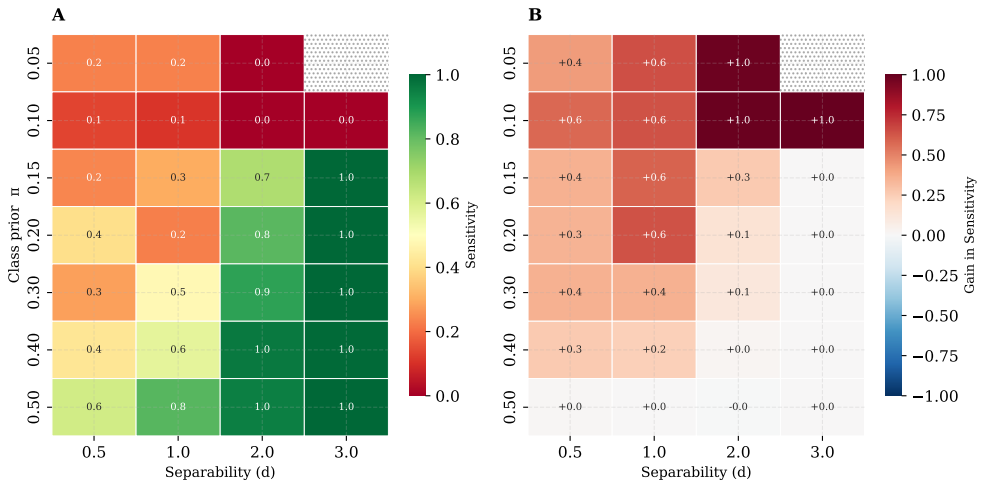


Fig. 3 Failure landscape of selective prediction under class imbalance. Heatmaps illustrate minority class Sensitivity at a fixed 30% global rejection rate across 56 synthetic conditions, parameterized by class prior ($\pi \in [0.05, 0.50]$, x-axis) and class separability ($d \in [0.5, 3.0]$, y-axis). **(A)** Standard global H_{Total} exhibits a severe structural failure boundary. Under conditions of severe imbalance and overlapping distributions ($d \leq 1.0, \pi \leq 0.20$) the baseline disproportionately defers minority patients (Sensitivity ≈ 0). **(B)** The proposed H_τ + CCR dual-architecture successfully rescues minority coverage within this diagnostic failure zone. Crucially, the architectural correction scales proportionally with the degree of imbalance, smoothly converging to identical standard baseline behavior when the natural prior is perfectly balanced ($\pi = 0.50$).

severe scaling law: systematic exclusion of minority patients by standard UQ scales inversely and non-linearly with class imbalance. While standard H_{Total} maintains a neutral stability slope (+0.13) at perfect balance ($\pi = 0.50$), it aggressively defers the minority class at extreme imbalance, driving the slope down to -2.09 at $\pi = 0.10$.

Evaluating the independent performance curves confirms that neither component alone can resolve this collapse. Applying only the H_τ signal globally successfully resolves the threshold-entropy mismatch (Vulnerability 1), sharply reducing the exclusion slope at $\pi = 0.10$ from -2.09 to -0.26. However, because the spatial penalty (Vulnerability 2) remains unaddressed, the system still leaks minority patients. Implementing the complete dual-architecture (H_τ + CCR) fully stabilizes the triage mechanism. The combined framework maintains a safe, positive stability slope (bounded between +0.13 and +0.24) across all tested priors, maintaining stable minority protection regardless of the underlying disease prevalence.

Evaluating the system at a fixed global rejection rate highlights the mechanistic trade-off inherent to selective classification (Figure 4C). Under severe imbalance, the standard baseline achieves a disproportionately high Specificity (≈ 0.94 at $\pi = 0.10$) strictly by displacing its rejections onto the minority class. By correcting this structural bias to recover minority Sensitivity, the proposed H_τ + CCR framework forces

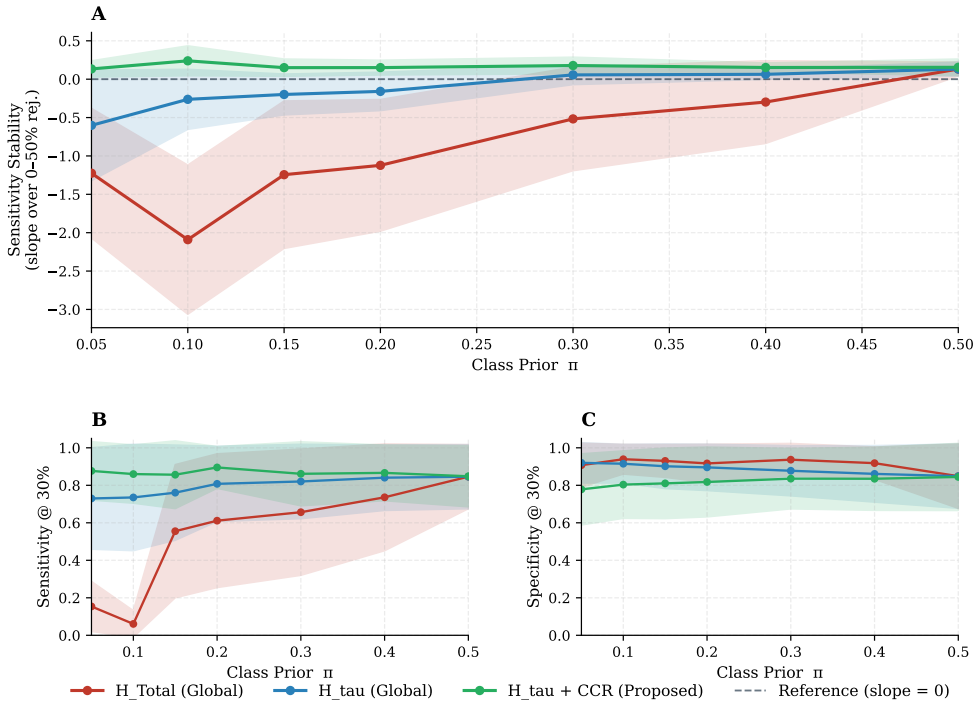


Fig. 4 Scaling law of minority Sensitivity under dynamic rejection across class priors. Panels show the mean (solid lines) and standard deviation (shaded regions) aggregated across varying levels of intrinsic class separability ($d \in [0.5, 3.0]$) at a sample size of $n = 1000$. **(A)** Sensitivity stability, defined as the linear slope of minority Sensitivity across the 0%–50% rejection range. Standard global H_{Total} (red) exhibits a severe non-linear scaling law, plunging into severe minority exclusion as class imbalance increases. The proposed H_τ + CCR dual-architecture (green) fully neutralizes this artifact, maintaining a safe trajectory near zero across all priors. **(B)** Sensitivity evaluated at a fixed 30% global rejection rate. Standard UQ systematically collapses under severe imbalance ($\pi \leq 0.20$), whereas the proposed framework guarantees robust minority retention. **(C)** Specificity (majority class retention) evaluated at the exact same 30% rejection rate. Under a fixed global rejection constraint, standard UQ maximizes Specificity by disproportionately concentrating deferrals within the minority class. By recovering minority Sensitivity (Panel B), the proposed framework mathematically redistributes these deferrals, resulting in a balanced and class-proportional rejection profile.

a mathematical reallocation of the deferrals. The resulting trade-off, which is a modest reduction in Specificity to secure a substantial recovery in minority Sensitivity, demonstrates a balanced distribution of uncertainty across both disease trajectories.

3.2 Real-world validation

To validate these findings in clinical practice, we evaluated the framework across three real-world cohorts (MS, PD, AD) whose class priors ($\pi \in \{0.13, 0.39, 0.54\}$) strategically span the structural failure zones identified in the synthetic experiments.

3.2.1 Empirical confirmation of uncertainty divergence

To validate the theoretical presence of Vulnerability 3 (structural uncertainty divergence), we analyzed the relationship between the predicted probabilities and the ensemble disagreement σ^2 (used as a proxy of epistemic uncertainty) across the clinical trajectories.

Figure 5 illustrates the highly asymmetric uncertainty profiles characteristic of imbalanced prognostic tasks. To formally test this divergence while accounting for the limited minority sample size, predictions were grouped into deciles based on their relative distance from the class-conditioned boundary (cumulative class proportions). For the densely sampled majority class (stable patients, $y = 0$), the model demonstrates solid certainty: as predictions move further from the decision boundary, the ensemble disagreement smoothly converges to zero. In contrast, the sparse minority class lacks this representational density. Consequently, even when the model correctly assigns high-confidence probabilities to progressor patients, these predictions are accompanied by a structurally higher ensemble disagreement. A paired Wilcoxon signed-rank test across these distance deciles formally confirms this asymmetry. In the severely imbalanced MS cohort, the minority class exhibits significantly higher variance than the majority class at equivalent relative distances from the boundary ($p < 0.01$, with $n \approx 23$ minority and $n \approx 48$ majority patients per decile bin). This structural divergence persists in the moderately imbalanced PD cohort ($p < 0.01$) but naturally dissolves in the near-balanced AD cohort ($p =$ n.s.). This confirms that standard global ranking systems inherently penalize minority patients not due to inherent biological ambiguity (aleatoric noise), but primarily due to the epistemic sparsity of their cohort.

To isolate the effect of class frequency and confirm this divergence is not an artifact of varying clinical feature spaces, we replicated this exact analysis under tightly controlled synthetic conditions (Supplementary Figures C5, C6). By fixing the underlying class separability at $d = 1.0$ and exclusively varying the minority prevalence, we reproduced the identical structural pattern. At a high imbalance of $\pi = 0.15$, the uncertainty asymmetry gap between classes is highly significant ($p < 0.01$). As the prevalence is artificially increased to $\pi = 0.30$, the gap narrows but remains significant ($p < 0.01$), before fully dissipating at parity ($\pi = 0.50$, $p =$ n.s.).

3.2.2 Rejection curve analysis

To trace the dynamic degradation of target-class coverage, we evaluated class-conditioned performance trajectories across the full continuum of rejection rates ($r \in [0, 0.80]$) for the three clinical cohorts (Figure 6). These empirical curves visually confirm the structural vulnerabilities of standard uncertainty metrics under severe class imbalance. Under severe imbalance (MS cohort, $\pi = 0.13$), H_{Total} triggers a catastrophic failure in minority protection. As the rejection rate increases, H_{Total} systematically excludes the progressor class, driving minority Sensitivity to zero by a 35% rejection rate. While applying the global H_τ correction without class-conditioning delays this collapse by fixing the threshold-entropy mismatch, it ultimately falls victim to epistemic compression and spatial penalties, reaching zero Sensitivity by 60% rejection. The Margin baseline instead empirically instantiates the spatial penalty: because

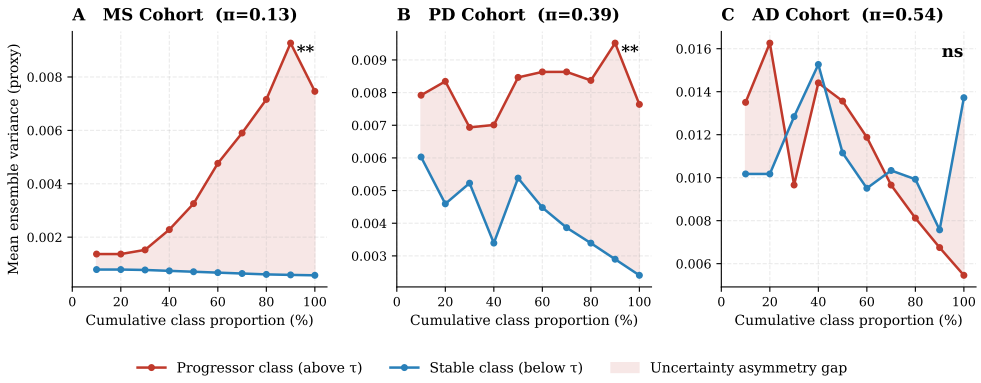


Fig. 5 Empirical confirmation of asymmetric uncertainty divergence. Mean ensemble disagreement (used as a proxy for epistemic uncertainty) is evaluated as a function of the prediction's relative distance from the class-conditioned decision boundary, binned by class-specific distance deciles. Statistical significance is evaluated via a paired Wilcoxon signed-rank test across the distance deciles (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns: not significant). Because data is partitioned into deciles, the sample size per bin remains rigorously constant within each cohort. **(A)** Under severe class imbalance (MS cohort), a profound uncertainty asymmetry gap emerges (**). As minority progressor predictions (red) move away from the decision threshold, their variance proportionally increases, whereas majority stable predictions (blue) converge near zero ($n \approx 23$ minority / 48 majority patients per decile bin). **(B)** In the moderately imbalanced PD cohort, minority variance remains structurally elevated across boundary distances (**; $n \approx 5$ minority / 9 majority per bin). **(C)** In the near-balanced AD cohort, this asymmetry naturally dissolves (ns; $n \approx 8$ minority / 6 majority per bin). This cross-cohort progression empirically confirms that uncertainty divergence is a mechanical artifact of class imbalance rather than intrinsic clinical noise.

the MS decision threshold ($\tau = 0.13$) is tightly compressed against the majority class space, the Margin metric systematically rejects highly confident stable (majority) patients. While this artificially inflates minority Sensitivity, it triggers a collapse in majority Specificity, which plunges from 0.70 down to nearly 0.20 at high rejection rates. This Sensitivity-Specificity trade-off, where an algorithm can only protect the minority by systematically reducing majority class coverage, renders global Margin structurally and practically unviable for these conditions. Combining CCR with a threshold-aware metric, such as our proposed H_τ (or the equivalent Margin), is the only framework that breaks this dynamic. By generating a centered uncertainty signal and enforcing class-conditioned triage queues, the framework smoothly and simultaneously improves both target-class Sensitivity and stable-class Specificity.

Empirical verification of the internal queue dynamics confirmed this continuous, symmetric error-filtering mechanism across the full continuum of rejection rates. For example, at a standard 30% global rejection rate in the MS cohort, the framework deferred false negatives from the $\mathcal{D}_0$ queue at more than twice the rate of true negatives (58.1% vs. 27.9%). Simultaneously, it actively reduced the number of false positives in the $\mathcal{D}_1$ queue at a substantially higher rate than confident true positives (33.5% vs. 19.7%). These targeted filtering trends hold across all evaluated cohorts (Supplementary Figure D7). This mechanism drives overall predictive utility (AUPRC) upward as uncertain predictions are safely deferred. These structural trends persist

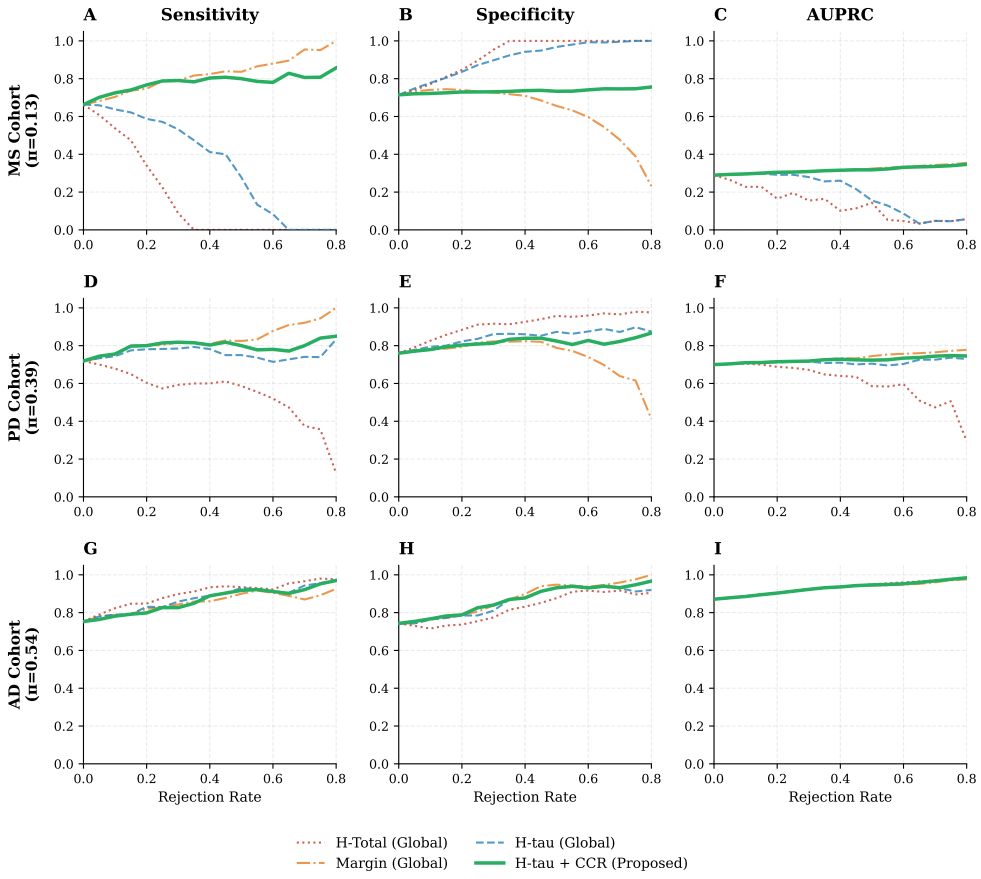


Fig. 6 Rejection trajectories across real-world clinical cohorts. Performance is evaluated across the full continuum of rejection rates ($r \in [0, 0.80]$) for severe imbalance (MS, top row), moderate imbalance (PD, middle row), and near-balance (AD, bottom row). Under severe imbalance, standard global H_{Total} (red dashed line) induces a complete Sensitivity collapse, systematically excluding minority progressor patients from the retained cohort. Conversely, the global Margin baseline (orange) empirically instantiates the spatial penalty: while it artificially protects minority Sensitivity, it systematically depletes majority class Specificity. The proposed H_τ + CCR dual-architecture (green) breaks this Sensitivity-Specificity trade-off, strictly bounding risk for both clinical trajectories and safely improving overall clinical utility (AUPRC). In the near-balanced AD cohort, these structural vulnerabilities inherently dissolve, and all frameworks safely converge.

under moderate imbalance (PD cohort, $\pi = 0.39$), where standard UQ continually erodes minority Sensitivity while the proposed framework maintains stable, bounded protection for both classes. Finally, on the near-balanced AD cohort ($\pi = 0.54$), the structural vulnerabilities dissolve, and all uncertainty frameworks smoothly converge to nearly identical, safe trajectories.

3.2.3 Class-conditioned safety evaluation

Evaluation in real-world clinical cohorts demonstrates that standard global uncertainty quantification fails under class imbalance, a vulnerability masked by standard global metrics (Table 2). To rigorously map the statistical stability of the generalized risk across patient trajectories, we performed post-prediction non-parametric bootstrapping ($B = 1000$ iterations) and calculated the class-conditioned Area Under the Generalized Risk-Coverage Curve (ccAUGRC). Under severe imbalance (MS) and moderate imbalance (PD), standard H_{Total} registers a biased statistical risk for the progressor class. While the Margin baseline appears to improve progressor risk, it simply displaces the error, resulting in statistically worse stable class ccAUGRC. Applying CCR with a threshold-aware metric (instantiated here by H_τ + CCR) is the only framework that successfully bounds risk for both clinical trajectories simultaneously. The Random + CCR control isolates the contribution of the class-conditioned rejection mechanism from the uncertainty signal. At MS, Random + CCR achieves a progressor ccAUGRC of 0.1605 [0.1084, 0.2123], significantly better than H_{Total} ($p < 0.001$) and global H_τ ($p = 0.004$), but significantly worse than H_τ + CCR ($p < 0.001$), a performance gap that holds even after Holm-Bonferroni correction. Furthermore, the H_{Total} + CCR ablation demonstrates that enforcing class-preservation on a misaligned global uncertainty signal is insufficient: it yields a progressor ccAUGRC of 0.1467, remaining significantly worse than the fully proposed framework ($p = 0.008$ after FWER adjustment). This decomposition quantifies the independent contribution of each component: the class-preservation mechanism alone accounts for the majority of the improvement over standard methods, while the threshold-aware H_τ signal provides additional gain by correctly ranking the most threshold-ambiguous patients within each trajectory. To ensure our partial-area formulation does not bias this comparative evaluation, all metrics were recalculated over a strict common support (Supplementary Table D3), fully confirming this relative performance ordering.

The near-balanced AD cohort ($\pi = 0.54$) validates the framework's behavior outside of the failure zone. In near-balance, H_{Total} achieves a lower progressor ccAUGRC than H_τ + CCR (0.0567 vs 0.0767, $p < 0.001$ adjusted), while simultaneously showing significantly higher stable ccAUGRC (0.1008 vs 0.0748, $p < 0.001$ adjusted). This reflects the exact inter-class substitution artifact identified in the Methods: at $\tau = \pi = 0.54$, where progressors are the majority class, H_{Total} preferentially retains progressors at the cost of stable patients rather than excluding them as it does at lower τ. As explicitly quantified by the class disparity column in Table 2, standard global entropy yields a highly skewed risk profile (disparity = 0.0441), even while its global AUGRC remains superficially competitive. The proposed H_τ + CCR framework prevents this unchecked substitution and reduces this disparity to near-zero (0.0019), preserving class-proportional risk profiles across both classes. In this near-balance scenario, the proposed architecture acts as a robust sanity check, safely matching the performance of standard methods without introducing detrimental artifacts, while the Random + CCR baseline naturally degrades to uninformative deferral.

Table 2 Generalized risk-coverage evaluation across clinical cohorts. Performance is measured via the class-conditioned Area Under the Generalized Risk-Coverage Curve (ccAUGRC) and Global AUGRC. The class disparity column (|Progressor ccAUGRC − Stable ccAUGRC|) explicitly quantifies between-class equity. Values are point estimates with 95% Confidence Intervals derived from 1000 post-prediction bootstrap iterations. Significance testing compares each baseline to the proposed reference framework (H_τ + CCR), with p-values adjusted for multiple comparisons using the Holm-Bonferroni method. Applying CCR with a threshold-aware metric (instantiated here by H_τ + CCR) is the only evaluated method to safely bound risk for both clinical trajectories under severe imbalance while minimizing class disparity. Lower values indicate better safety/lower risk.

Dataset	Uncertainty Framework	Progressor ccAUGRC (95% CI)	Stable ccAUGRC (95% CI)	Class Disparity	Global AUGRC (95% CI)
MS ($\tau = \pi = 0.13$)	H_{Total} (Global)	0.2756 [0.2080, 0.3357]$^{\uparrow***}$	0.0413 [0.0317, 0.0518]$^{\downarrow***}$	0.2343	0.0679 [0.0556, 0.0813]
	Margin (Global)	0.0931 [0.0562, 0.1378]$^{\downarrow**}$	0.1529 [0.1313, 0.1750]$^{\uparrow***}$	0.0598	0.1431 [0.1235, 0.1625]
	H_τ (Global)	0.2178 [0.1535, 0.2762]$^{\uparrow***}$	0.0561 [0.0449, 0.0680]$^{\downarrow***}$	0.1617	0.0734 [0.0605, 0.0867]
	Random + CCR	0.1605 [0.1084, 0.2123]$^{\uparrow***}$	0.1374 [0.1201, 0.1538]$^{\uparrow**}$	0.0231	0.1404 [0.1231, 0.1577]
	H_{Total} + CCR	0.1467 [0.0941, 0.2013]$^{\uparrow**}$	0.1395 [0.1230, 0.1562]$^{\uparrow*}$	0.0072	0.1404 [0.1235, 0.1577]
	H_τ + **CCR**	**0.1115 [0.0682, 0.1564]**	**0.1301 [0.1135, 0.1466]**	**0.0186**	**0.1275 [0.1119, 0.1442]**
PD ($\tau = \pi = 0.39$)	H_{Total} (Global)	0.1891 [0.1288, 0.2469]$^{\uparrow***}$	0.0431 [0.0218, 0.0660]$^{\downarrow***}$	0.1460	0.0918 [0.0646, 0.1191]
	Margin (Global)	0.0840 [0.0454, 0.1275]$^{\downarrow**}$	0.1066 [0.0681, 0.1468]$^{\uparrow ns}$	0.0226	0.0973 [0.0688, 0.1255]
	H_τ (Global)	0.1160 [0.0676, 0.1651]$^{\uparrow*}$	0.0776 [0.0469, 0.1118]$^{\downarrow*}$	0.0384	0.0926 [0.0649, 0.1188]
	Random + CCR	0.1351 [0.0890, 0.1851]$^{\uparrow*}$	0.1147 [0.0798, 0.1505]$^{\uparrow*}$	0.0204	0.1228 [0.0940, 0.1541]
	H_{Total} + CCR	0.1052 [0.0600, 0.1519]$^{\uparrow ns}$	0.0950 [0.0612, 0.1306]$^{\uparrow ns}$	0.0102	0.0987 [0.0720, 0.1267]
	H_τ + **CCR**	**0.1031 [0.0577, 0.1494]**	**0.0912 [0.0564, 0.1259]**	**0.0119**	**0.0958 [0.0681, 0.1227]**
AD ($\tau = \pi = 0.54$)	H_{Total} (Global)	0.0567 [0.0352, 0.0834]$^{\downarrow***}$	0.1008 [0.0676, 0.1382]$^{\uparrow***}$	0.0441	0.0756 [0.0570, 0.0966]
	Margin (Global)	0.0800 [0.0525, 0.1115]$^{\uparrow ns}$	0.0733 [0.0459, 0.1057]$^{\downarrow ns}$	0.0067	0.0763 [0.0572, 0.0966]
	H_τ (Global)	0.0708 [0.0463, 0.1001]$^{\downarrow ns}$	0.0818 [0.0520, 0.1162]$^{\uparrow ns}$	0.0110	0.0755 [0.0567, 0.0964]
	Random + CCR	0.1180 [0.0852, 0.1540]$^{\uparrow***}$	0.1228 [0.0859, 0.1626]$^{\uparrow***}$	0.0048	0.1203 [0.0940, 0.1479]
	H_{Total} + CCR	0.0773 [0.0521, 0.1068]$^{\uparrow ns}$	0.0755 [0.0473, 0.1071]$^{\uparrow ns}$	0.0018	0.0764 [0.0573, 0.0973]
	H_τ + **CCR**	**0.0767 [0.0515, 0.1067]**	**0.0748 [0.0465, 0.1061]**	**0.0019**	**0.0758 [0.0570, 0.0966]**

Significance vs. Proposed (H_τ + CCR): *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns: not significant.

Arrows indicate if the baseline is structurally worse (↑, higher risk) or better (↓, lower risk).
Red markers ($*$) indicate that statistical significance is maintained after Holm-Bonferroni correction ($\alpha = 0.05$).

4 Discussion

Our findings demonstrate that under severe class imbalance, standard uncertainty quantification (UQ) fundamentally ceases to measure clinically aligned ambiguity, instead functioning as a proxy for class membership. Evaluation in synthetic and real-world neurodegenerative cohorts confirms that global metrics, such as Shannon entropy and classification margin, structurally penalize minority patients. This penalization is driven by three compounding mathematical failure modes: threshold-entropy mismatch, asymmetric spatial penalties, and structural uncertainty divergence.

Clinically, this measurement failure has severe implications. In prognostic settings, the minority class often represents the most severe and clinically urgent trajectories, such as rapid progressors in multiple sclerosis and Parkinson's Disease.

Because minority patients are typically represented by sparser feature distributions, they exhibit structurally higher ensemble disagreement. By conflating this dataset sparsity with decision-level ambiguity, standard global UQ systems systematically defer high-risk patients, actively excluding them from the automated pipeline and creating an untenable algorithmic trade-off.

The empirical phase diagram defines clear boundaries under which threshold-aware, class-conditioned frameworks (such as H_τ + CCR) are strictly necessary, namely when class prior and class separability jointly fall below $\pi \approx 0.20$ and $d \approx 1.0$. In this region, standard metrics lead to Sensitivity collapse under severe imbalance and to progressive minority erosion at moderate imbalance. Both components of the dual-architecture are therefore required to recover system performance. The proposed

threshold-aware ambiguity score (H_τ) corrects the signal layer by realigning maximum uncertainty with the natural class prior, while Class-Conditioned Rejection (CCR) corrects the triage layer by ranking patients within their predicted trajectories. In doing so, CCR neutralizes the structurally elevated absolute uncertainty floor associated with the minority class.

Crucially, the ablation study using the Random + CCR control isolates the contribution of the class-conditioned rejection mechanism. The results show that class-preservation alone substantially mitigates minority exclusion, while the H_τ signal provides an additional benefit by correctly ordering ambiguous cases within those protected queues.

Conversely, in near-balanced clinical tasks, such as the Alzheimer's Disease cohort ($\pi = 0.54$), these structural vulnerabilities naturally attenuate. Under these balanced conditions, the proposed framework converges to standard baseline behavior, indicating that the architectural correction scales proportionally with the degree of imbalance and introduces no measurable penalty when the classes are approximately balanced. Notably, even in this near-balanced setting, H_{Total} exhibits a residual substitution artifact in the opposite direction, preferentially protecting the now-larger progressor class at the expense of stable patients. This suggests that inter-class substitution is a general property of global aggregation, rather than a phenomenon specific to severe imbalance.

These findings extend recent theoretical work questioning the appropriateness of conditional entropy and mutual information as uncertainty measures under distribution shift [5, 9]. While prior work has largely focused on probabilistic recalibration or multimodal data acquisition to improve predictive confidence, these results show that standard selective failure is distinctly geometric. This vulnerability persists even after rigorous Platt scaling because it arises fundamentally from the structural misalignment between entropy, the shifted decision boundary, and the asymmetric probability space.

A core consequence of these vulnerabilities is that standard evaluation paradigms, such as the Area Under the Generalized Risk-Coverage Curve (AUGRC), may fail to detect automated minority exclusion. Because global multi-threshold metrics aggregate errors across the full population, a selective prediction system can mathematically improve its overall safety score by shifting residual risk onto the harder-to-predict minority class. To address this inter-class substitution artifact, we introduced class-conditioned AUGRC (ccAUGRC). By quantifying the generalized risk-coverage trade-off separately for each clinical trajectory, ccAUGRC prevents risk-masking and makes subgroup-specific failures directly observable. This addresses a critical gap identified by recent literature [14, 16], which successfully improved multi-threshold metrics but had yet to decompose risk by class subpopulation. Integrating such class-stratified metrics into reproducible machine learning pipelines is crucial to ensure that automated classification mechanisms maintain strictly bounded risk across all subgroups of patients, rather than concealing algorithmic bias behind improved overall safety.

Operationalizing the CCR framework within clinical workflows requires moving from absolute uncertainty thresholding to queue-based triage. Rather than attempting to mathematically optimize a global cost function, which requires assigning precise relative weights to false negatives, false positives, and clinician workload, CCR operationalizes deferral strictly as a fixed resource constraint. In practice, the hospital system first defines the decision threshold (τ) according to the observed disease prevalence (π) in its target population. Incoming patient predictions are then assigned to their respective diagnostic trajectories and ranked by their H_τ scores. Administrators can subsequently define distinct, clinically appropriate rejection rates (r) for each queue based on available expert capacity. This ensures that manual review resources are allocated proportionally across the cohort, ensuring that the most ambiguous cases from every diagnostic trajectory receive expert review, rather than allowing one class to monopolize the triage bandwidth or erroneously automating boundary cases. Crucially, however, CCR bases triage decisions exclusively on predicted trajectories rather than true clinical states. Consequently, a true progressor confidently misclassified by the base classifier as stable (a false negative with $p \ll \tau$) is assigned to the predicted-stable queue and does not benefit from the protective balancing mechanisms applied to the progressor queue. CCR is not designed to correct confident classifier errors; rather, its specific, narrower purpose is to prevent the selective prediction mechanism from systematically starving the minority class of manual review capacity due to the structural geometry of the uncertainty score.

Positioning our framework relative to the conformal prediction literature clarifies the practical trade-offs of our approach. Class-conditional conformal prediction (CCCP) and distribution-free risk control provide rigorous, finite-sample guarantees on error rates by dynamically adapting the size of prediction sets. While theoretically optimal for bounding risk, this adaptive approach often yields variable rejection volumes that are difficult to align with fixed hospital resources (e.g., a specialist can only review a strict quota of 10 cases per day). Furthermore, under severe class imbalance, CCCP often satisfies its mathematical guarantees by generating uninformative prediction sets for the minority class (e.g., empty sets or multi-label predictions), which complicates binary clinical triage. Conversely, CCR sacrifices formal distribution-free coverage bounds in favor of strict operational constraints. By combining a fixed quota system with threshold-aware uncertainty ranking, CCR allows clinics to set a predictable budget for manual chart reviews (e.g., deferring exactly 30% of cases) while structurally ensuring that this fixed bandwidth is distributed proportionally across predicted clinical trajectories, thereby preventing minority exclusion.

Although the present evaluation sets τ equal to the observed class prior π, in line with standard clinical practice, the proposed framework does not depend on this equivalence. By construction, H_τ reaches its maximum at $p = \tau$ for any $\tau \in (0, 1)$, regardless of how τ is selected. In clinical settings, where false negatives carry substantially greater cost than false positives (such as multiple sclerosis screening, where missing a rapid progressor may be more harmful than unnecessarily flagging a stable patient), clinicians may deliberately set τ below π to prioritize Sensitivity. The piecewise rescaling and class-conditioned rejection mechanism remain valid under this shift: H_τ continues to anchor maximum uncertainty to the selected operating point,

while CCR preserves proportional representation around the resulting decision structure. The framework therefore generalizes to cost-sensitive threshold selection without requiring modification. However, when τ is shifted substantially away from π, the queue sizes preserved by CCR still reflect the empirical dataset distribution rather than the shifted clinical prior. Practitioners should account for this distinction when interpreting the resulting triage proportions.

Several limitations should be acknowledged. First, the framework is currently formulated for binary prognostic tasks; extension to multiclass settings will require rethinking both the entropy rescaling procedure, including which τ value anchors H_τ in the presence of multiple boundaries, and the CCR partitioning scheme. Furthermore, while the intra-class ordinal ranking of CCR makes it highly robust to monotonic miscalibration, severe violations of the logistic assumption during post-hoc calibration (e.g., Platt scaling) could distort absolute probabilities enough to shift boundary patients across the decision threshold τ, thereby placing them into the incorrect triage queue. Additionally, the decision threshold τ is treated as known and fixed, and is set equal to the observed class prior. In deployment, however, τ may vary as disease prevalence changes or as patient populations differ across clinical sites. Regarding data availability, the Italian MS and PD cohorts cannot be publicly shared because of patient privacy regulations; however, we release the synthetic data and use the publicly available ADNI cohort in our analysis. Computationally, the ensemble pipeline is more demanding than single-model approaches, and the behavior of H_τ + CCR in single-model or deep learning settings remains to be investigated. Finally, because the bootstrap procedure was applied post-prediction to the aggregated ensemble outputs, the resulting confidence intervals capture test-set sample variability but do not account for the variance of the model training process itself.

5 Conclusion

The fundamental premise of selective prediction is that algorithms should defer to human experts when their predictions are uncertain. This study shows, however, that under severe clinical class imbalance, this safety mechanism can be compromised by the way uncertainty is operationalized. In these settings, standard uncertainty quantification may fail to capture clinically aligned ambiguity and instead reflect dataset sparsity.

The primary takeaway of this work is that reliance on standard, global uncertainty metrics can create a misleading illusion of safety. Because probabilistic models are structurally less confident for underrepresented classes, minority patients may receive systematically higher uncertainty scores that reflect data sparsity rather than triage-relevant ambiguity. Standard triage systems can therefore defer these patients at disproportionately high rates, while global evaluation metrics fail to detect the resulting bias, instead reporting improved overall performance simply by routing the most clinically urgent cases to the manual review queue.

By showing that combining a threshold-aware metric (H_τ) with CCR mitigates this failure mode, our results demonstrate that this algorithmic bias is not an unavoidable consequence of imbalanced data, but a correctable limitation in how uncertainty

is measured, ranked, and operationalized. By ensuring that rejection is applied proportionally across predicted trajectories (up to finite-sample rounding errors), and by realigning the uncertainty signal to aggressively filter false negatives near the decision boundary, clinical pipelines can safely defer ambiguous cases without systematically excluding the minority class from the automated pipeline.

These findings also highlight the need to distinguish between AI systems designed for biomarker discovery and those intended for clinical deployment. In exploratory discovery, high predictive variance might provide useful evidence of biological heterogeneity. In automated clinical triage, however, uncorrected variance can become a mechanism that deprives minority patients of the benefits of automation. Future clinical machine learning pipelines should therefore move beyond global risk-coverage evaluation and incorporate class-stratified metrics such as ccAUGRC. Algorithmic safety cannot be achieved by simply shifting the entire minority class to human experts; it requires risk to be directly measured and controlled within each patient trajectory, rather than inferred from the population as a whole.

6 Declarations

6.1 Funding

This work was partially funded under the National Plan for Complementary Investments to the NRRP, project "D34H—Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care" (project code: PNC0000001), Spoke 2: "Multilayer platform to support the generation of the Patients' Digital Twin", CUP: B53C22006170001, funded by the Italian Ministry of University and Research.

Conflict of interest

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The data that support the findings of this study are available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

The remaining datasets analysed during the current study are not publicly available due to patient privacy and institutional clinical regulations, but are available from the corresponding author on reasonable request. These private datasets, including their

baseline characteristics and validation protocols, have been previously described in detail in [1] and [39].

Code availability

The code generated and analysed during the current study is available in a GitHub repository, https://github.com/endrwalter/uncertainty_ml.

Acknowledgements

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

Author contribution

WE conceptualized the study, developed the methodology and software, and wrote the original draft of the manuscript. FR, RM, SB, and MM contributed to data curation, methodological development, and the interpretation of results, with FR additionally handling data visualization. FR and VO critically reviewed and finalised the manuscript. GJ and VO provided senior supervision and project administration. All authors read and approved the final manuscript.

References

[1] Poretto, V., Endrizzi, W., Betti, M., Bovo, S., Bellinvia, A., Ragni, F., Lapucci, C., Moroni, M., Marangoni, S., Portaccio, E., Longo, C., Gios, L., Chierici, M., Jurman, G., Giometto, B., Inglese, M., Osmani, V., Marenco, M., Uccelli, A., Amato, M.P.: Machine learning analysis applied to prediction of early progression independent of relapse activity in multiple sclerosis patients. European Journal of Neurology **32**(12), 70417 (2025) https://doi.org/10.1111/ene.70417 . e70417 EJoN-25-0800

[2] Cabral, C., Morgado, P.M., Costa, D.C., Silveira, M., Disease Neuroimaging Initiative, A., *et al.*: Predicting conversion from mci to ad with fdg-pet brain images at different prodromal stages. Computers in biology and medicine **58**, 101–109 (2015)

[3] Malaguti, M.C., Longo, C., Moroni, M., Ragni, F., Bovo, S., Chierici, M., Gios, L., Avanzino, L., Marchese, R., Di Biasio, F., Pardini, M., Cerne, D., Mandich, P., Marenco, M., Uccelli, A., Giometto, B., Jurman, G., Osmani, V., NeuroArtP3 Network: Machine learning predicts risk of falls in parkinson's disease patients in a multicenter observational study. European Journal of Neurology **32**(5), 70118 (2025) https://doi.org/10.1111/ene.70118 . e70118 EJoN-24-2520.R1

[4] Lopez, L.J.L., Elsharief, S., Jorf, D.A., Darwish, F., Ma, C., Shamout, F.E.: Uncertainty quantification for machine learning in healthcare: A survey. In: Xu, X.O., Choi, E., Singhal, P., Gerych, W., Tang, S., Agrawal, M., Subbaswamy, A., Sizikova, E., Dunn, J., Daneshjou, R., Sarker, T., McDermott, M., Chen, I. (eds.) Proceedings of the Sixth Conference on Health, Inference, and Learning. Proceedings of Machine Learning Research, vol. 287, pp. 862–907. PMLR, ??? (2025). https://proceedings.mlr.press/v287/lopez25a.html

[5] Jiménez, S., Jurgens, M., Waegeman, W.: Position: Epistemic uncertainty estimation methods are fundamentally incomplete. In: International Conference on Machine Learning (2026). https://api.semanticscholar.org/CorpusID:278996371

[6] Bickford Smith, F., Kossen, J., Trollope, E., Van Der Wilk, M., Foster, A., Rainforth, T.: Rethinking aleatoric and epistemic uncertainty. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (eds.) Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 4345–4359. PMLR, ??? (2025). https://proceedings.mlr.press/v267/bickford-smith25a.html

[7] Jong, I.P., Sburlea, A.I., Sabatelli, M., Valdenegro-Toro, M.: Measuring Orthogonality as the Blind-Spot of Uncertainty Disentanglement (2026). https://arxiv.org/abs/2408.12175

[8] Chen, J., Su, B.: Transfer Knowledge from Head to Tail: Uncertainty Calibration

under Long-tailed Distribution (2023). https://arxiv.org/abs/2304.06537

[9] Wimmer, L., Sale, Y., ..., P.H., Bischl, B., Hüllermeier, E.: Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In: Evans, R.J., Shpitser, I. (eds.) Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence. Proceedings of Machine Learning Research, vol. 216, pp. 2282–2292. PMLR, ??? (2023). https://proceedings.mlr.press/v216/wimmer23a.html

[10] Collell, G., Prelec, D., Patil, K.R.: A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. Neurocomputing **275**, 330–340 (2018) https://doi.org/10.1016/j.neucom.2017.08.035

[11] Jiang, H., Kim, B., Guan, M., Gupta, M.: To trust or not to trust a classifier. In: Advances in Neural Information Processing Systems, vol. 31 (2018)

[12] Cortes, C., DeSalvo, G., Mohri, M.: Learning with rejection. In: Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings, pp. 67–82. Springer, Berlin, Heidelberg (2016). https://doi.org/10.1007/978-3-319-46379-7_5

[13] Rezaei, M., Näppi, J.J., Bischl, B., Yoshida, H.: Bayesian uncertainty estimation for detection of long-tailed and unseen conditions in medical images. Journal of Medical Imaging **10**(5), 054501 (2023) https://doi.org/10.1117/1.jmi.10.5.054501

[14] Pugnana, A., Ruggieri, S.: Auc-based selective classification. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, pp. 8061–8082 (2023). PMLR

[15] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[16] Traub, J., Bungert, T.J., Lüth, C.T., Baumgartner, M., Maier-Hein, K.H., Maier-Hein, L., Jäger, P.F.: Overcoming common flaws in the evaluation of selective classification systems. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems, vol. 37, pp. 2323–2347. Curran Associates, Inc., ??? (2024). https://doi.org/10.52202/079017-0076 . https://proceedings.neurips.cc/paper_files/paper/2024/file/047c84ec50bd8ea29349b996fc64af4b-Paper-Conference.pdf

[17] Chow, C.: On optimum recognition error and reject tradeoff. IEEE Trans. Inf. Theor. **16**(1), 41–46 (2006) https://doi.org/10.1109/TIT.1970.1054406

[18] El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. Journal of Machine Learning Research **11**(53), 1605–1641 (2010)

[19] Gangrade, A., Kag, A., Saligrama, V.: Selective classification via one-sided prediction. CoRR **abs/2010.07853** (2020) 2010.07853

[20] Del Corso, G., Colantonio, S., Caudai, C.: Shedding light on uncertainties in machine learning: formal derivation and optimal model selection. Journal of the Franklin Institute **362**(3), 107548 (2025) https://doi.org/10.1016/j.jfranklin.2025.107548

[21] Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: specialized uncertainties for specialized tasks. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. NIPS '24. Curran Associates Inc., Red Hook, NY, USA (2024)

[22] Bindini, L., Perini, L., Nistri, S., Davis, J., Frasconi, P.: Dealing with Uncertainty in Contextual Anomaly Detection (2026). https://arxiv.org/abs/2507.04490

[23] Jones, E., Sagawa, S., Koh, P.W., Kumar, A., Liang, P.: Selective Classification Can Magnify Disparities Across Groups (2021). https://arxiv.org/abs/2010.14134

[24] López, L.J.L., Shamout, F.E., Rudner, T.G.J.: An Empirical Analysis of Calibration and Selective Prediction in Multimodal Clinical Condition Classification (2026). https://arxiv.org/abs/2603.02719

[25] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

[26] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007 (2017). https://doi.org/10.1109/ICCV.2017.324

[27] Narasimhan, H., Menon, A.K., Jitkrittum, W., Gupta, N., Kumar, S.: Learning to reject meets long-tail learning. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=ta26LtNq2r

[28] Hoarau, A., Quost, B., Destercke, S., Waegeman, W.: Reducing aleatoric and epistemic uncertainty through multi-modal data acquisition. Machine Learning **115**(5), 121 (2026)

[29] Sağlam, F., Özgen, Uygun, A., Dinçer, O.S., Albayrak, C.: Selective classification under imbalance in multiclass settings: A novel metric for bias-aware risk–coverage evaluation. Journal of Biomedical Informatics **181**, 105084 (2026) https://doi.org/10.1016/j.jbi.2026.105084

[30] Angelopoulos, A.N., Bates, S.: A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification (2022). https://arxiv.org/abs/

2107.07511

[31] Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association **114**(525), 223–234 (2018) https://doi.org/10.1080/01621459.2017.1395341

[32] Ding, T., Angelopoulos, A.N., Bates, S., Jordan, M.I., Tibshirani, R.J.: Class-Conditional Conformal Prediction with Many Classes (2023). https://arxiv.org/abs/2306.09335

[33] Vickers, A.J., Elkin, E.B.: Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making **26**(6), 565–574 (2006)

[34] Elkan, C.: The foundations of cost-sensitive learning. In: International Joint Conference on Artificial Intelligence (IJCAI), vol. 1, pp. 973–978 (2001)

[35] Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X.: A survey of uncertainty in deep neural networks. Artif. Intell. Rev. **56**(Suppl 1), 1513–1589 (2023) https://doi.org/10.1007/s10462-023-10562-9

[36] Smith, L., Gal, Y.: Understanding measures of uncertainty for adversarial example detection. arXiv (2018) https://doi.org/10.48550/arxiv.1803.08533

[37] et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods (1999)

[38] Müller, J., Cagol, A., Lorscheider, J., Tsagkas, C., Benkert, P., Yaldizli, Ö., Kuhle, J., Derfuss, T., Sormani, M.P., Thompson, A., Granziera, C., Kappos, L.: Harmonizing definitions for progression independent of relapse activity in multiple sclerosis: A systematic review. JAMA Neurology **80**(11), 1232–1245 (2023) https://doi.org/10.1001/jamaneurol.2023.3331

[39] Endrizzi, W., Campese, N., Ragni, F., Moroni, M., Bovo, S., Longo, C., Gios, L., Uccelli, A., Giometto, B., Jurman, G., Osmani, V., Malaguti, M.C., Falini, A., Castellano, A., Rossi, A., Tortora, D., Parodi, C., Verrico, A., Sabatini, F., Portaccio, E., Betti, M., Pasquini, G., Gerli, F., Niccolai, C., Cama, I., Campi, C., Cirone, A., Garbarino, S., Piana, M., Ottaviani, D., Di Giacopo, R., Bacchin, R.: Explainable machine learning reveals the challenges of predicting new-onset motor complications in parkinson's disease. npj Parkinson's Disease (2026) https://doi.org/10.1038/s41531-026-01550-1