Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Global Metrics: A Zone-Stratified Diagnostic Framework Based on Confusion Matrix Components

Tareef Fadhil Raham

tareeffadhil@yahoo.com

University of Warith Alanbiyaa

Research Article

Keywords: Epistemic Equivalence Point (EEP), Zone-Relevant Epistemic Divergence Index (ZREDI), Zone-Based Balanced Error Score (ZBES), Zone- Balanced Error (ZBE) Index, Zone-Based Tension State (ZBTS)

Posted Date: November 21st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7539984/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective:

Traditional diagnostic metrics summarize model performance globally but often obscure localized vulnerabilities. This study proposes a zone-stratified evaluation framework to assess diagnostic reliability across confidence gradients.

Methods:

Predicted probabilities were stratified into five diagnostic confidence zones using the Robust Adjusted Mean Interval (RAMI) approach: Recessive, Lower Plausible, Trusted, Upper Plausible, and Inflated. Within each zone, confusion matrix components were used to compute: the Zone-Relative Epistemic Divergence Index (ZREDI) for directional trust asymmetry, the Epistemic Equivalence Point (EEP) for calibration symmetry, the Zone-Based Error Skew (ZBES) for misclassification bias, and the Zone-Balanced Error (ZBE) for class-neutral error burden. A higher-order construct, the Zone-Based Tension State (ZBTS), was introduced to quantify overall epistemic instability across diagnostic strata.

Results:

Simulated assessments and real-world diagnostic data were used to uncover zone-specific diagnostic behaviors often masked by traditional global metrics. In both datasets, the Trusted Zone (TZ) consistently demonstrated calibration stability, with EEPs clustering near zero. In contrast, the outer zones—particularly the Recessive and Inflated regions—exhibited pronounced shifts in ZBES and ZREDI, indicating increased risks of underdiagnosis, overdiagnosis, or erosion of diagnostic trust. Furthermore, ZBTS emphasized the interplay between ZBES and ZREDI, highlighting zones of epistemic fragility and enabling localized performance audits for regulatory transparency and safety profiling.

Conclusions:

Zone-stratified metrics provide actionable insights into diagnostic model behavior, improving interpretability and safety. This framework advances beyond aggregate measures to support better threshold tuning, risk calibration, and clinical deployment—especially where model trust must align with uncertainty.

1. Introduction

Diagnostic models are increasingly used to support clinical decision-making, risk stratification, and early disease detection. However, Traditional evaluation metrics—such as accuracy, sensitivity, and specificity—offer global assessments of model performance. However, these metrics often fail to capture local reliability, especially in diagnostically uncertain regions or when data are imbalanced [1,2].

Although widely used, ROC curves and AUC values summarize performance globally and may obscure important diagnostic failures within specific regions of prediction confidence—even in more localized approaches such as Zone-Restricted ROC (ZR-ROC) [3–5].

These metrics treat all errors equally, irrespective of whether they occur in high-certainty regions or in transitional, borderline zones—where misclassifications may carry greater clinical consequences.

Although the F1 score is widely used for evaluating binary classifiers, it reflects performance only on the positive class by harmonizing precision (PPV) and recall (sensitivity).[2] This property makes it particularly useful in class-imbalanced

datasets, where accurate positive detection is prioritized. However, the F1 score does not reflect the model's performance on the negative class, nor does it incorporate the variability of predictive reliability across different score thresholds or confidence strata. As such, it overlooks error distribution across other regions of the prediction space, particularly in zones of diagnostic uncertainty[1]. **c**onsequently, it provides a limited—and potentially misleading—view of diagnostic robustness in zone-based clinical decision contexts, particularly where false positives (FPs) and false negatives (FN) carry asymmetric clinical consequences [3, 4] .

While scalar measures such as the Predictive Summary Index (PSI**)**—defined as **PPV + NPV − 1**—have been proposed to summarize the overall post-test utility of diagnostic models, their interpretive value remains limited in high-stakes clinical settings. PSI ranges from **− 1** (completely misleading test) to **+ 1** (perfect prediction), providing a concise, user-facing summary of predictive usefulness[6].

However, despite its intuitive appeal, PSI suffers from several limitations: it ignores class-specific misclassification consequences, fails to account for prediction confidence stratification, and provides no insight into where within the prediction spectrum failures occur. Second, PSI functions as a global average and cannot identify where in the diagnostic score range misclassification risk is concentrated—such as in boundary zones of epistemic uncertainty [7, 8].As such, it may obscure zone-specific fragility, particularly in contexts where **false** reassurance or overtreatment can lead to significant harm.

Furthermore, it is highly sensitive to prevalence, making it unreliable across heterogeneous populations or disease distributions [7–9].

In addition, the PSI provides no insight into the directionality of prediction errors—namely, whether a model tends to systematically overpredict or underpredict a condition. Nor does it capture localized collapses in reliability across the diagnostic spectrum. These limitations render PSI insufficient for nuanced interpretability and safety governance in clinical AI, particularly in settings where error asymmetry and zone-specific trust calibration are critical. As a result, there is a growing need for zone-aware evaluation metrics that can characterize directional reliability and misclassification patterns across stratified regions of epistemic uncertainty[3, 8, 10].

In response to these limitations, recent approaches such as **ZR-ROC** analysis have been developed to evaluate model performance within epistemically stratified regions—such as zones defined by predicted score distributions, likelihood intervals, or adjusted means [11]. These methods enhance diagnostic transparency by isolating performance in areas of increased uncertainty. However, they remain focused on traditional discriminative metrics (e.g., true positive rate vs. FPR and do not capture the **directional reliability** of predictions—such as the extent to which a model confidently rules in versus rules out a condition within a given zone.

To address these gaps, we introduce a complementary framework of zone-stratified reliability metrics that emphasize both **predictive trust** and **directional diagnostic bias** within epistemically defined regions.

We introduce a suite of **zone-stratified diagnostic metrics** designed to assess prediction reliability, calibration balance, and error dynamics across **epistemically meaningful strata** of diagnostic confidence. These metrics move beyond global performance indicators such as AUC or accuracy, offering localized insight into model behavior in clinically sensitive regions. The core components of the framework include:

1. **Zone-Relative Epistemic Divergence Index (ZREDI)**: Quantifies **calibration asymmetry** by measuring the divergence between PPV and NPV within each zone, highlighting **trust imbalance** between outcome classes.
2. **Zone-Based Error Skew (ZBES)**: Captures the **directionality of misclassification**, defined as the difference between the false negative rate (FNR) and false positive rate (FPR). It provides a signed measure of diagnostic bias and complements value-based metrics like ZREDI.

3. **Zone-Based Tension State (ZBTS)**: A composite interpretive construct defined as **ZREDI – ZBES**, capturing the **epistemic tension** between predictive asymmetry and error skew. ZBTS helps identify zones where reliability and misclassification are **incongruent or paradoxically aligned**.

Together, these metrics enable **zone-specific evaluation** of diagnostic systems across the full spectrum of clinical risk, including diagnostic zones. This framework supports **threshold refinement**, **context-aware model evaluation**, and **epistemically governed deployment** of machine learning models in high-stakes clinical settings.

2. Methods and materials:

2.1. Diagnostic Zone Stratification

The dataset was stratified into diagnostic confidence zones using the RAMI framework, which partitions model predictions based on their standardized distance from an adjusted mean score. This standardized distance functions analogously to the standard deviation (SD) in classical statistics, representing how far individual predictive scores deviate from a central reference point, thereby quantifying the degree of epistemic divergence from calibrated expectation. In our epistemic framework, this robustly adjusted central reference point is termed the **Credence Value (CV)**—a stable proxy for the adjusted mean, around which diagnostic confidence and epistemic stratification are systematically evaluated. Within the RAMI framework, a unit RAMI distance corresponds approximately to the distance between the adjusted mean and ± 1 SD, offering a normalized scale for assessing predictive certainty. This framework functions as an epistemic analogue to classical confidence intervals, but instead of merely bounding statistical estimates, it reflects the operational reality of model trust, calibration integrity, and interpretive stability across predictive strata.

Each sample was classified into one of five epistemic zones: the **Recessive Zone (RZ)**, **Lower Plausible Zone (LPZ)**, **Trusted Zone (TZ)**, **Upper Plausible Zone (UPZ)**, **or Inflated Zone (IZ)**. These zones reflect increasing degrees of epistemic uncertainty, with the TZ representing predictions made with the highest level of diagnostic confidence and minimal asymmetry.

The Lower and Upper Plausible Zones serve as transitional regions on either side of the TZ, where model predictions become less stable and the risk of misclassification begins to rise. At the extremes, the RZ is typically dominated by TN and may correspond to underdiagnosis, while the *IZ* is dominated by TP and may reflect overdiagnosis or excessive certainty. This stratification provides a foundation for zone-specific performance analysis, emphasizing local diagnostic reliability rather than global accuracy alone.

The RAMI-based stratification framework is currently under refinement[12]. All intellectual property related to its structure, thresholds, and terminology is retained by the author.

2.2. Confusion Matrix Elements per Zone and Metric Computation Workflow

The computation of diagnostic metrics was conducted on a zone-specific basis using the RAMI-stratified dataset. Each observation was assigned to one of five diagnostic confidence zones—Recessive, Lower Plausible, Trusted, Upper Plausible, or Inflated—based on RAMI values. Within each zone, the core confusion matrix components were computed: true positives (TPs), true negatives (TNs), FPs, and FNs. TP referred to correctly identified diseased cases, TN to correctly identified non-diseased cases, FP to non-diseased cases incorrectly classified as diseased, and FN to diseased cases missed by the model. These foundational counts enabled the calculation of both traditional performance metrics (e.g., sensitivity, specificity, PPV, NPV) and the novel zone-stratified metrics introduced in this study—namely, ZREDI, ZBES, ZBE,

and ZBTS. This localized approach allowed precise characterization of model behavior across varying levels of diagnostic confidence, offering granular insights into reliability, misclassification asymmetry, and epistemic stability.

2.3. Algorithmic Pipeline for RAMI-Stratified Epistemic Metric Extraction: Multistage Computation of ZREDI, Global Divergence Index (GDI), ZBES, ZBE, Epistemic Equivalence Point (EEP), DNP, EPEP, and ZBTS

Step 1

Stratify the dataset into diagnostic confidence zones — Recessive, Lower Plausible (LPZ), Trusted, Upper Plausible (UPZ), and Inflated — using standard RAMI-derived thresholds (± 1 SE and ± 2 SE).

Alternatively, users may define desired zone boundaries to align with specific diagnostic cut-offs, regulatory guidelines, or study protocols. This optional flexibility allows the framework to adapt to study-specific diagnostic thresholds, rare-event sampling considerations, and validation protocols that require non-standard confidence intervals."

Within the Recessive Zone (RZ) and Inflated Zone (IZ), values that extend beyond the 1st Recessive Value (1st RV) or 1st Inflated Value (1st IV) remain part of RZ/IZ.

Step 2

For each zone, calculate

TPz, TNz, FPz, FNz

Step 3

These were used to derive the Zone-Relative Epistemic Divergence Index (ZREDI), calculated as the absolute difference between PPVz and NPVz

ZREDI = (TPz / [TPz + FPz]) – (TNz / [TNz + FNz]). i.e. ZREDI = PPVz – NPVz

Step 4: Identify the EEP when: Step 3: Identify the EEP where

the Positive Predictive Value (PPV) equal to Negative Predictive Value (NPV)

PPVz = NPVz or (TPz / [TPz + FPz]) = (TNz / [TNz + FNz])

Step 5: obtain the Global Divergence Index (GDI = total PPV –total NPV).

Step 6: To capture directional misclassification, we computed the False Negative Rate (FNR) and False Positive Rate (FPR) as: FNR = FN/TP + FN, FPR = FP/TN + FP

Step 7: Find EEP: EEP occurs when: TPz / (TPz + FPz) = TNz / (TNz + FNz)

Step 8 : Compute ZBES for each zone:

$$ZBES = FNz / (TPz + FNz) - FPz / (TNz + FPz)$$

= FNR – FPR

All metrics were computed independently for each RAMI-defined diagnostic zone.

Step 8: Find Diagnostic Neutrality Point (DNP): DNP occurs when :FNRz=FPRz⇒ZBESz=0

Step 9: Find Epistemic Pressure Equilibrium Point (EPEP) : EPEP occur when ZREDIz = ZBESz

Step 10: Find Zone-Based Tension State (ZBTS) : ZBTS = ZREDI − ZBES

Step 11:Find GDI − PSI Delta

2.4.Validation Assessment

To ensure both theoretical soundness and practical relevance, we employed a two-part validation strategy integrating both simulation and real-world data. First, synthetic datasets were developed to emulate the behavior of a binary classifier, such as those used in HbA1c-based diagnostic systems. In this simulation, we modeled an idealized diagnostic setting with balanced class distribution and minimal noise, enabling baseline evaluation of zone-based metrics—including ZREDI, ZBES, ZBE, and ZBTS—under stable epistemic conditions. The simulated case was assigned a RAMI score (representing standardized distance from the diagnostic threshold), a binary ground-truth label, and a predicted label based on a deterministic rule: RAMI ≥ 0 predicted positive, while RAMI < 0 predicted negative. RAMI scores were then stratified into five predefined diagnostic confidence zones— to facilitate localized error analysis. This structure revealed zone-specific fragilities, with underdiagnosis concentrated in RZ and LPZ, overdiagnosis in UPZ and IZ, and more balanced behavior in the TZ. Finally, the zone-based evaluation framework was applied to a real-world AFP–HCC dataset to demonstrate clinical utility and interpretability. Together, these validation components support the reliability, generalizability, and epistemic insight offered by zone-stratified metrics across both simulated and real diagnostic contexts.

3.Results

3.1. Simulated Example

A total of 250 simulated patient cases (Appendix 1) were generated to emulate a binary classification system, with RAMI values calibrated to mimic the behavior of HbA1c-based diagnostic thresholds. These cases were stratified into five diagnostic confidence zones based on RAMI scores. Zone-specific classification performance was evaluated using standard and novel metrics, including PPV, NPV, ZREDI, ZBES, and ZBTS. The results of this zone-stratified diagnostic evaluation are summarized in Table 1.

Table 1. Zone-Stratified Diagnostic Reliability and Error Metrics

Summary of predictive and error-based diagnostic metrics across epistemic confidence zones. Metrics include PPV, NPV, ZREDI, ZBES, and ZBTS. Global values (DGI, PSI) are included for comparison.

Zone	TP	TN	FP	FN	PPV (TP / [TP+FP])	NPV (TN / [TN+FN])	ZREDI (PPV-NPV)	FNR	FPR	ZBES (FNR-FPR)	ZBTS (ZREDI-ZBES)
Recessive	0	32	0	18	NC	0.64	NC	1	0	1	NC
Lower Plausible	0	34	0	16	NC	0.68	NC	1	0	1	NC
Trusted	13	13	13	11	0.5	0.541667	-0.04167	0.458333	0.5	-0.04167	0
Upper Plausible	42	0	8	0	0.84	NC	NC	0	1	-1	NC
Inflated	50	0	0	0	1	NC	NC	0	NC	NC	NC
Total (global)	105	79	21	45	0.833333	0.637097	DGI =0.196	0.3	0.21	PZI=0.09	0.106

***** NC = Not Computable due to lack of class-specific signal.

Table 1 presents zone-stratified diagnostic performance metrics derived from Example 1, including key indicators such as PPV, NPV, the ZREDI, computed as PPV minus NPV), ZBES, defined as the difference between the FNR and FPR, and the ZBTS, representing the residual between ZREDI and ZBES. Together, these metrics uncover latent diagnostic imbalances distributed across the epistemic confidence strata and serve to evaluate the alignment between predictive reliability and the distribution of classification errors.

As depicted in **Figure 1**, PPV exhibits a monotonic increase across the stratified zones, culminating at a value of 1.0 in the *IZ*, where every positive prediction is correct. Conversely, NPV reaches its peak in the *LPZ* (0.68), tapering to 0.54 in the *TZ*. This decline reflects diminishing confidence in negative predictions as one moves from low- to mid-epistemic regions. The *TZ*, notably, marks the EEP—the zone where PPV and NPV converge (0.50 and 0.54, respectively)—and represents the point of maximal post-test symmetry and diagnostic balance.

In zones where TPz + FPz = 0 or TNz + FNz = 0, ZREDI becomes undefined, denoted as “Not Computable” (NC). This arises from division by zero in the predictive value calculations and signals epistemic vacuity—a complete absence of class-specific predictive or diagnostic signal. Such vacuity is observed in the Recessive, Lower Plausible, Upper Plausible, and Inflated Zones, each of which lacks the necessary conditions for computing PPV, NPV, or both. These vacuous zones thus reflect either diagnostic collapse or an overconfident skew.

As further illustrated in **Figure 2**, the *TZ* stands out as the convergence nexus of three epistemic landmarks: the EEP (where PPV approximates NPV), the DNP, where ZBES equals zero), and the EPEP, (defined by the equality of ZREDI and ZBES). At this critical junction, the ZBTS value is minimized (≈ 0), indicating a state of low epistemic tension and balanced diagnostic behavior.

In contrast, zones outside the *TZ* manifest polarized diagnostic distortions. The Recessive and *LPZ*s exhibit underdiagnosis, characterized by undefined ZREDI, extreme positive ZBES (equal to +1.0), and the complete absence of TPs—reflecting a one-sided reliance on negative classifications. On the other hand, the Upper Plausible and Inflated Zones demonstrate signs of overdiagnosis. Here, ZREDI is high or undefined due to inflated PPV, and ZBES is either negative or zero, indicating a heavy skew toward FPs or complete certainty without counterbalance.

Although global diagnostic metrics such as PPV (0.833), NPV (0.637), the Global Divergence Index (DGI = 0.196), and the Predictive Skew Index (PSI = 0.09) suggest relatively good overall model performance, these aggregate measures

obscure the heterogeneity of reliability and error across zones. The Recessive and *LPZ*s, for instance, suffer from complete diagnostic failure (TP = 0), with undefined ZREDI and maximal ZBES (+1.0), despite the favorable global PPV. Only the *TZ* maintains near-convergence of reliability and error (ZREDI ≈ ZBES ≈ –0.0417), validating its designation as a post-test equilibrium region. The global ZBTS value of 0.106 quantifies the overall discrepancy between reliability divergence and error skew, highlighting the underlying epistemic friction and reinforcing the utility of zone-aware metrics such as ZREDI, ZBES, and ZBTS for robust and context-sensitive diagnostic evaluation.

3.1.1.Valuation of Clinical Utility and Zonal Error Burden Using ZBE:

Figure 3 presents a zone-stratified analysis of clinical utility metrics in Example 1, focusing on Net Benefit, Expected Diagnostic Cost, and Zone-Balanced Error (ZBE). Net Benefit is calculated at a threshold of 0.5, while Expected Diagnostic Cost is derived under the assumption that FNs carry twice the cost of FPs. ZBE, introduced here as a class-neutral and interpretable measure, captures the average of FNR and FPR, quantifying overall diagnostic misclassification within each epistemic zone.

The *TZ* stands out as the optimal region of diagnostic operation, where error burden is lowest (ZBE = 0.479), and cost and benefit metrics are relatively balanced. While Net Benefit is neutral in the Recessive, Lower Plausible, and *TZ*s, it sharply increases in the Upper Plausible (0.68) and Inflated (1.0) Zones, indicating the greatest clinical gain in high-confidence regions. These results, supported by Appendix 2, underscore that clinical utility concentrates in zones of elevated epistemic confidence.

Expected Diagnostic Cost (Appendix 3) is highest in the Recessive and *LPZ*s due to the predominance of FNs and absence of TPs, with costs of 1.8 and 1.6, respectively. In contrast, cost drops sharply in the *UPZ* (0.16) and is neutral in the *IZ* (0.0), supporting efficient diagnostic decision-making when the model is highly certain.

ZBE values from Appendix 4 confirm that the Recessive and *LPZ*s exhibit the highest misclassification burden (ZBE = 0.5), while the *IZ* has perfect classification (ZBE = 0.0). The *TZ* again emerges as the most balanced region in terms of reliability and error (ZBE = 0.479), validating its role as a diagnostic anchor.

Together, these findings highlight that global metrics, while informative (global Net Benefit = 0.337; global EDC = 0.942; global ZBE = 0.396), risk obscuring critical intra-zonal variability.(Appendices 2, 3, and 4)

Figure 4 illustrates the net clinical benefit of predictive decisions across diagnostic confidence zones as a function of threshold probability. The *TZ* exhibits the highest net benefit across a broad range of thresholds, confirming its superior diagnostic utility and balance between sensitivity and specificity. The *UPZ* and *LPZ* show moderate net benefit, indicating their potential clinical value under selective conditions. In contrast, the *RZ* and *IZ* yield the lowest net benefits, reflecting diagnostic inefficiency due to high misclassification risk (underdiagnosis and overdiagnosis, respectively). This zone-based DCA supports threshold-sensitive decision-making by quantifying real-world trade-offs between benefit and harm across epistemic strata.

The application of clinical utility metrics at the **zone level** represents an innovative advancement in evaluating diagnostic model performance. By integrating **Net Benefit, Expected Diagnostic Cost,** and the newly introduced **Zone-Based Error (ZBE)** within RAMI-stratified diagnostic zones, this approach transcends conventional, aggregate-level assessments. While global metrics offer a useful summary, they often obscure critical intra-zonal disparities that directly impact clinical decision-making. Zone-stratified evaluation, by contrast, reveals how diagnostic reliability, misclassification burden, and decision value fluctuate across varying levels of model certainty. This granular insight enables more informed deployment of diagnostic tools—targeting zones of high confidence while flagging regions of epistemic fragility. As such, zone-based clinical utility auditing not only enhances interpretability but also establishes a more **risk-aware and evidence-aligned framework** for responsible clinical model validation.

3.2. Real-World Diagnostic Example

In a diagnostic evaluation of hepatocellular carcinoma (HCC), **401 archived plasma samples** were analyzed, comprising **208 samples from confirmed HCC patients and 193 from liver cirrhosis patients without HCC**, who served as clinical controls. This dataset is derived from the study by Jang et al. (2016), *PLoS ONE* [https://doi.org/10.1371/journal.pone.0151069[13,14]. In this context, **AFP's diagnostic performance** was assessed specifically in distinguishing **HCC from cirrhosis,** a clinically critical challenge since **patients with cirrhosis represent the key population targeted for early HCC surveillance.** Accordingly, the resulting confusion matrix and all derived diagnostic metrics—summarized in Tables 2 and 3—capture AFP's capacity to discriminate malignant (HCC) from non-malignant (cirrhosis) chronic liver disease within a clinically relevant high-risk population.This real-world dataset enables a grounded application of the zone-stratified framework, allowing for fine-grained analysis of biomarker reliability and diagnostic utility across different levels of model certainty. In this real-world example, **epistemic zone boundaries** were **predefined** to match **widely accepted AFP diagnostic ranges** used in hepatocellular carcinoma (HCC) surveillance rather than being derived from statistical confidence intervals. This approach ensured **immediate clinical relevance** while maintaining compatibility with **zone-based calibration** for interpretability.

Table 2. Zone-Stratified Diagnostic Classification of AFP Levels Across Epistemic Confidence Zones

This table summarizes zone-wise diagnostic classification counts (TP, FN, FP, TN) across five AFP concentration ranges, aligned with increasing diagnostic confidence. By mapping AFP levels to epistemic zones—from low-certainty (Recessive) to high-certainty (Inflated)—the stratification enables detailed assessment of AFP's ability to distinguish HCC from cirrhosis within a high-risk population.

AFP Zone	Epistemic Zone	TP	FN	FP	TN
<5	Recessive	5	30	5	80
5-20	Lower plausible	35	25	6	60
20-200	Trusted	55	20	6	25
200-1000	Upper Plausible	25	3	2	7
>1000	Inflated	8	2	0	2

Table 3. Epistemic metrics of AFP diagnostic performance across confidence zones

This table summarizes zone-stratified epistemic metrics for AFP performance, highlighting asymmetries in predictive reliability and error across diagnostic confidence zones. ZBTS values reveal zones of underconfidence (negative values) and overconfidence (positive values), offering a refined lens to interpret AFP's diagnostic behavior across clinically meaningful thresholds.

AFP Zone	ZREDI	ZBES	ZBTS
<5	-0.2273	0.7983	-1.0256
5-20	0.1478	0.3258	-0.178
20-200	0.3461	0.0731	0.273
200-1000	0.2259	-0.1151	0.341
>1000	0.5	0.2	0.3
Global	GDI= 0.1857	PSI= 0.2862	Delta GDI-PSI = -0.1005

The application of epistemic indices across stratified AFP diagnostic zones revealed distinct patterns in predictive reliability and error distribution. ZREDI (Zone-Relevant Epistemic Divergence Index) and ZBES varied meaningfully across AFP ranges, with ZBTS capturing latent diagnostic tensions. The Recessive (<5 ng/mL) and Lower Plausible (5–20 ng/mL) zones exhibited negative ZREDI values and high ZBES, indicating underconfident behavior characterized by conservative predictions coexisting with substantial misclassification. In contrast, the Trusted (20–200 ng/mL), Upper Plausible (200–1000 ng/mL), and Inflated (>1000 ng/mL) zones demonstrated increasingly positive ZREDI with declining or stabilizing ZBES, suggesting overconfident diagnostic behavior—high apparent reliability despite error asymmetry. Three key epistemic landmarks were identified: the EEP, where ZREDI = 0 and predictive calibration is symmetrical (PPV = NPV), was found at AFP ≈ 8.56 ng/mL; the intersection of ZREDI and ZBES (AFP ≈ 50.98 ng/mL) reflected equilibrium between calibration divergence and error skew; and a DNP, defined by ZBES = 0 (FNR = FPR), occurred near AFP ≈ 300.32 ng/mL. These results highlight how epistemic analysis of diagnostic zones uncovers nuanced reliability profiles beyond conventional sensitivity or specificity, with implications for threshold optimization and interpretive guidance in HCC screening.(fig.5 and fig.6)

To contextualize epistemic performance in terms of actionable clinical value, we applied a three-part Zone-Based Clinical Utility Evaluation to the real AFP dataset. First, Zone-Stratified Net Benefit was assessed using Decision Curve Analysis (DCA), capturing the clinical trade-off between TPs and FPs across AFP-defined diagnostic zones. This revealed that the Trusted and Upper Plausible zones yielded the highest net benefit, particularly at intermediate threshold probabilities. Second, Expected Diagnostic Cost was calculated by assigning penalty weights to FPs and FNs, demonstrating that cost-efficiency improves markedly once AFP levels exceed 200 ng/mL—corresponding to reduced misclassification in those zones. Third, we introduced the Zone-Balanced Error (ZBE) metric, computed as the average of false positive and false negative rates within each zone, providing a class-neutral estimate of misclassification burden. ZBE further distinguished zones of error symmetry, offering complementary insight to the directional asymmetry captured by ZREDI and ZBES. Together, these utility metrics reveal that AFP levels above 200 ng/mL (Upper Plausible and Inflated zones) yield the highest clinical value, combining maximal net benefit, minimal cost, and lowest ZBE. The *TZ* (20–200 ng/mL) also shows strong utility, balancing moderate cost with high net benefit and low ZBE. In contrast, lower AFP zones (<20 ng/mL) demonstrate high diagnostic cost, negligible net benefit, and elevated ZBE—suggesting limited utility and greater risk of misclassification. Overall, this integrated framework aligns epistemic and utility-based evaluations to inform evidence-weighted threshold selection in AFP-guided HCC screening.(fig.7)

4.Discussion

4.1. Zone-Stratified Diagnostic Evaluation Framework

This study introduces a zone-stratified diagnostic evaluation framework anchored by four interrelated metrics—ZREDI, ZBES, Zone-Balanced Error (ZBE), and the composite ZBTS. Unlike traditional global metrics such as sensitivity, specificity, or AUC, which provide aggregate summaries, these zone-specific measures offer granular, epistemically grounded insights into diagnostic model behavior across varying levels of certainty. The framework is applied across five defined diagnostic zones—Recessive, Lower Plausible, Trusted, Upper Plausible, and Inflated—each corresponding to a distinct confidence tier that shapes interpretive value and clinical decision-making.

Researchers may rely on RAMI-derived intervals when prioritizing uncertainty quantification, or alternatively use pre-established clinical thresholds when aiming for alignment with regulatory guidelines and real-world validation protocols. This is well demonstrated by the observed misalignment between RAMI-derived probabilistic thresholds and clinically established decision levels, which highlighted the need to adopt predefined AFP cut-offs in the real-world example of AFP-based HCC diagnostics. These thresholds, widely recommended in hepatocellular carcinoma (HCC) surveillance protocols, ensured both clinical relevance and practical interpretability of the diagnostic framework.

This finding underscores the flexibility of the zone-stratified framework. In this HCC evaluation, we selected clinically accepted AFP ranges rather than deriving boundaries from statistical confidence intervals. This approach preserved immediate clinical applicability while still enabling zone-based calibration, thereby demonstrating the framework's adaptability across diverse diagnostic, regulatory, and research contexts.

4.2 Zone-Relative Predictive Symmetry and the EEP

The ZREDI quantifies directional asymmetry in predictive reliability by subtracting NPV from PPV within each diagnostic zone. A ZREDI value near zero reflects balanced trust in both positive and negative predictions, as observed in the TZ. As ZREDI diverges from zero—particularly beyond ± 0.30—it indicates increasing asymmetry, signaling overconfidence in one class and reduced reliability in the other, which can elevate clinical risk. For example, highly positive ZREDI values suggest overdiagnosis potential (e.g., in IZs), while negative values imply underdiagnosis (e.g., in RZs).

To summarize global post-test asymmetry, the Global Divergence Index (GDI = PPV − NPV) can be used. While informative, GDI lacks ZREDI's localized granularity and may obscure stratified vulnerabilities. Thus, ZREDI offers a more actionable metric for zone-specific auditing and recalibration.

The EEP marks the zone where ZREDI equals zero—that is, where PPV and NPV converge—indicating optimal diagnostic neutrality. This point serves as a key benchmark for model calibration and deployment readiness. In our simulation, the EEP aligned with the TZ, where sensitivity and NPV both approximated 54.17%, reflecting a rare state of post-test symmetry and maximal epistemic stability. When the EEP shifts into higher confidence zones (e.g., Upper Plausible or Inflated), it may suggest delayed calibration or overconfidence. Conversely, if it lies in lower zones, it may signal early but fragile equilibrium.

Clinically, the EEP is more than a statistical construct—it acts as a meta-indicator of reliability. Its position guides diagnostic trust, threshold setting, and safety interventions. Embedded within zone-stratified evaluation, the EEP and ZREDI together provide a dual-lens assessment of model performance, combining predictive directionality with balance, and supporting both interpretive clarity and deployment confidence.

The ZREDI is intentionally constructed using PPV and NPV—two prediction-centered metrics that reflect the trustworthiness of a model's outputs rather than its accuracy against ground truth. PPV (TP / [TP + FP]) measures the likelihood that a positive prediction is correct, while NPV (TN / [TN + FN]) provides the same for negative predictions. Because these values are conditioned on the model's predictions, they align closely with how clinicians and end users interpret and act on diagnostic outputs in real time.

This framing makes ZREDI a uniquely user-facing metric, offering a direct lens into the asymmetry of predictive reliability across zones. Unlike sensitivity or specificity—which are grounded in outcome-based performance—ZREDI answers a more immediate and clinically relevant question: *"When the model predicts a positive or negative outcome, how reliable is that prediction?"*

Moreover, PPV and NPV act as practical proxies for epistemic certainty. When either metric is low, it signals a zone of local diagnostic fragility—indicating that the model is offering predictions unsupported by sufficient evidence. By taking the signed difference (PPV − NPV), ZREDI reveals the direction and magnitude of this reliability imbalance, uncovering potential overdiagnosis (when PPV ≫ NPV) or underdiagnosis (when NPV ≫ PPV) that may be obscured in global metrics.

Ultimately, ZREDI enables granular, zone-specific auditing of trust asymmetry. It supports recalibration where needed, enhances interpretability for end users, and operationalizes the concept of epistemic risk. By anchoring itself in predictive reliability rather than retrospective correctness, ZREDI bridges methodological precision with clinical applicability, offering a robust tool for epistemically informed model evaluation.

4.3. ZBES, DNP, and the Global Prediction Symmetry Index (PSI)

The ZBES is a directional diagnostic metric that quantifies the asymmetry of misclassification within each epistemic zone. Defined as the difference between the FNR and the FPR, it reflects the net diagnostic bias of a model's errors. A ZBES of zero denotes perfect error symmetry, while positive or negative values highlight skew toward underdiagnosis or overdiagnosis, respectively. This makes ZBES a powerful lens for detecting class-specific vulnerabilities that are often masked by global metrics such as accuracy or AUC.

In our simulation, ZBES varied substantially across zones. The TZ achieved near-perfect symmetry (ZBES = + 0.042), indicating well-balanced misclassification. Conversely, the Recessive and Inflated Zones reached extreme values of + 1.0 and − 1.0, respectively, indicating complete directional breakdown—either dominated by FNs or FPs. Such skew reveals critical diagnostic fragility that can significantly impact clinical outcomes if left unrecognized.

To guide interpretation, ZBES can be qualitatively categorized: values ≤ 0.10 suggest excellent diagnostic balance; 0.11–0.30 reflects moderate skew; 0.31–0.50 indicates high misclassification risk; and values > 0.50 signify severe directional failure. Though mathematically equivalent to the difference between sensitivity and specificity, ZBES offers distinct interpretive value by focusing explicitly on misclassification burden rather than correct detection.

Importantly, ZBES also defines the DNP—the threshold at which FNR equals FPR (i.e., ZBES = 0). This point marks the zone of symmetrical diagnostic error, serving as a complementary benchmark to the EEP, where PPV equals NPV. While EEP signals equilibrium in predictive trust, the DNP indicates balance in diagnostic fallibility. When both occur within the same zone, typically the TZ, it suggests maximal epistemic stability and reinforces that zone's clinical defensibility.

To contextualize ZBES in a broader framework, we compare it to the Prediction Symmetry Index (PSI)—a global analogue defined either as PSI = FNR − FPR or alternatively as PSI = (PPV + NPV) − 1. Both forms capture system-wide directional imbalance, with positive values indicating a tendency toward underdiagnosis and negative values suggesting overdiagnosis. While PSI provides an overarching view of error asymmetry, it lacks the spatial resolution offered by ZBES[15, 16].By computing ZBES within each epistemic zone (e.g., Recessive, Trusted, Inflated), one can detect zone-specific misclassification patterns that may be obscured at the aggregate level .

Thus, ZBES and PSI function in parallel—PSI offering global insight, ZBES offering local granularity. Used together, they enable comprehensive auditing of diagnostic models within stratified or threshold-sensitive applications. Clinically, this integrated framework supports better triage decisions, recalibration strategies, and risk-aware deployment by identifying not just how often the model is wrong, but how and where its errors manifest.

4.4. ZBTS, Epistemic Tension, and the EPEP

The ZBTS is introduced as a composite interpretive construct derived from the interaction between the ZREDI and the ZBES. Defined as ZBTS = ZREDI − ZBES, this metric captures the divergence between predictive trust asymmetry and diagnostic error imbalance within each epistemic zone. A ZBTS value close to zero reflects epistemic balance, indicating that calibration (i.e., trust in predictions) and misclassification patterns are in alignment. In contrast, positive ZBTS values signal overconfidence—situations where a model's predictive reliability exceeds its actual error performance—whereas negative values imply underconfidence, denoting conservative diagnostic behavior that yields higher-than-expected accuracy despite low trust indicators.

A unique case arises when ZREDI equals ZBES, resulting in ZBTS = 0. This defines the EPEP, which, unlike the EEP—where PPV equals NPV—or the DNP—where FNR equals FPR—represents a convergence of both calibration and error symmetry. When EPEP occurs, often near the TZ, it marks a zone of minimal epistemic tension and optimal diagnostic harmony. In our simulation, distinct Epistemic Tension States (ETS) were observed across zones. The TZ showed balanced calibration and error (ZREDI ≈ ZBES ≈ 0), resulting in ZBTS = 0 and diagnostic neutrality. The RZ, which produced no positive predictions, had an undefined ZREDI but a maximally defined ZBES of + 1.0, resulting in a low ZBTS value indicative of dysfunctional symmetry—where apparent predictive neutrality coexists with unidirectional diagnostic failure. The IZ showed perfect PPV but undefined NPV and ZBES due to absent negative predictions, implying a high-ZBTS state and overconfidence. These results illustrate how ZBTS identifies zones of latent epistemic instability, even when core metrics are not computable.

At the global level, an analogous measure of epistemic divergence can be expressed as the difference between the Divergence in Global Informativeness (DGI = PPV − NPV) and the Prediction Symmetry Index (PSI = FNR − FPR). In our real data example, global DGI was 0.196 and PSI was 0.09, resulting in a delta (Δ_Global) of 0.106. This modest but non-negligible gap reflects global epistemic tension, suggesting that predictive reliability and misclassification are not perfectly aligned. However, global deltas can obscure zone-specific failures. ZBTS, by contrast, reveals where calibration outpaces—or falls behind—actual performance, providing a localized map of epistemic fragility that supports more granular threshold tuning and interpretive safety.

Furthermore, while PSI summarizes directional misclassification imbalance across the entire dataset, it is agnostic to local context. ZBES serves as its stratified analogue, revealing spatial heterogeneity in diagnostic error patterns (1,2). ZREDI, ZBES, and ZBTS, taken together, offer a multidimensional lens on model behavior, exposing both global asymmetries and local imbalances. When mapped across diagnostic zones, this triad enables transparency in trust allocation, risk assessment, and deployment planning, particularly in high-stakes clinical settings where epistemic safety is essential .

4.5.Evaluating Diagnostic Utility Across Confidence Zones

In this study, we incorporated Zone-Balanced Error (ZBE) alongside Net Benefit and Expected Diagnostic Cost to provide a comprehensive view of diagnostic utility.

Zone-Balanced Error (ZBE) is a class-neutral metric designed to quantify the average misclassification burden—calculated as the mean of the FNR and FPR—within each diagnostic confidence zone. Unlike conventional global metrics such as accuracy, sensitivity, or even global balanced error (BE), which aggregate performance across the entire dataset, ZBE enables localized evaluation by isolating error patterns within epistemically defined strata (e.g., Trusted, Recessive, Inflated). This is particularly critical in real-world clinical settings, where model reliability and risk may vary dramatically across prediction thresholds. ZBE provides actionable insight by identifying zones where total error rates are acceptably low—or conversely, dangerously high—thereby supporting zone-specific deployment strategies. Complementing ZBE, additional zone-stratified utility metrics were applied: zone-level Net Benefit, which quantifies clinical utility under

specified decision thresholds; and Expected Diagnostic Cost, which incorporates error penalties (e.g., higher cost for FNs). Together, these metrics form a robust toolkit for evaluating diagnostic models under epistemic uncertainty, allowing stakeholders to audit not just how much or what kind of error a model makes—but where those vulnerabilities occur within the decision landscape.

4.6. Zone-Stratified Evaluation in Real-World Context

Compared to zone-level results, the global diagnostic metrics present an oversimplified and potentially misleading picture of AFP performance. The global ZREDI of + 0.190 reflects moderate predictive asymmetry overall, yet it masks critical zonal divergences such as the extreme underconfidence seen in the RZ (ZBTS = − 1.026) and the strong overconfidence in the UPZ (ZBTS = + 0.341). Similarly, the global ZBES (+ 0.0667) suggests a modest error imbalance but fails to capture the sharp error skew in zones like Recessive (+ 0.7983) or Inflated (+ 0.2000). While the global ZBTS (+ 0.1233) implies a net overconfident posture, this aggregate hides the diagnostic instability present in lower AFP strata. These discrepancies underscore the importance of zone-stratified analysis in revealing clinically relevant epistemic behaviors that would otherwise be obscured by aggregate statistics.

The application of the proposed zone-stratified epistemic framework to real-world AFP data in the context of hepatocellular carcinoma (HCC) revealed diagnostic behaviors that align with, but also refine, prior observations in the literature. In the RZ (< 5 ng/mL), AFP demonstrated a high false-negative burden, consistent with the known limitations of AFP in detecting early-stage or AFP-negative HCC [13, 17]. This underscores the need for supplementary diagnostic strategies in AFP-negative populations. In the LPZ (5–20 ng/mL), model behavior remained conservative, with underprediction and modest gains in reliability, reflecting transitional epistemic behavior with residual diagnostic ambiguity.

The TZ (20–200 ng/mL), situated between the Lower and Upper Plausible strata, emerged as a region of relative epistemic balance, with ZREDI values approaching zero and ZBES showing minimal asymmetry. This suggests a zone of predictive stability, where both positive and negative predictions can be interpreted with greater confidence. However, the Upper Plausible (200–1000 ng/mL) and Inflated (> 1000 ng/mL) zones revealed a shift toward positive bias, characterized by high PPV, decreasing ZBE, and rising ZREDI values. While this reflects improved detection sensitivity, it also introduces the risk of overdiagnosis if interpretive safeguards are not in place [18, 19].

Notably, the lack of full convergence between key epistemic landmarks—such as the EEP and the DNP—in higher zones highlights that diagnostic confidence does not equate to perfect reliability. These findings demonstrate the practical value of zone-based stratification in modeling AFP-based HCC detection: it provides a continuous and interpretable view of reliability dynamics, moving beyond static thresholds to support context-sensitive, risk-adaptive decision-making.

4.7. Comparison with Traditional Metrics

Traditional diagnostic metrics—such as accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC)—offer high-level summaries of classifier performance but often obscure localized diagnostic weaknesses that are crucial in clinical decision-making. These global indicators aggregate performance across the entire prediction spectrum and are generally threshold-independent, which can mask overdiagnosis in high-certainty zones or underdiagnosis in ambiguous or low-certainty stra ta[15, 16, 20]. AUC, while widely reported and useful for ranking models, provides limited insight into real-world interpretability, calibration, or localized trustworthiness—especially in imbalanced datasets[2].

In contrast, the zone-stratified framework proposed in this study—including the ZREDI, ZBES, Zone-Based Error (ZBE), and the EEP—enables more granular and context-aware evaluation. These metrics anchor diagnostic interpretation within epistemically defined strata (e.g., Recessive, Trusted, Inflated), uncovering asymmetries in predictive reliability and misclassification risk that global metrics tend to overlook. For example, while conventional specificity might appear

acceptable, ZBES may reveal directional skew due to excess FNs within a specific diagnostic zone. Likewise, EEP identifies the precise threshold where PPV equals NPV—offering a robust signal of calibration symmetry that is invisible to conventional tools .).

The F1 score, as noted in the introduction, does not reflect model performance on the negative class and fails to account for variability in predictive reliability across score thresholds or confidence strata. Within zone-stratified analysis, F1 may provide complementary insights—particularly in strata enriched with positive predictions, such as the IZ. However, its interpretive value is limited unless considered alongside zone-aware metrics that capture directional bias, calibration asymmetry, and diagnostic uncertainty. Integrating F1 with metrics like ZREDI, ZBES, and ZBTS ensures a more comprehensive and context-sensitive evaluation of model performance.

Together, these zone-based metrics do not replace conventional performance measures but enrich them by localizing interpretability, enhancing threshold sensitivity, and offering epistemic clarity. By aligning performance evaluation with diagnostic reality—where uncertainty and risk vary across thresholds—they support safer, more adaptive deployment of AI models in clinical contexts.

4.8. Global and Clinical Implications of Zone-Stratified Diagnostic Metrics

The integration of zone-stratified diagnostic metrics—EEP, ZREDI, ZBES, Zone-Balanced Error (ZBE),and ZBTS — represents a significant advancement in the evaluation and safe deployment of diagnostic models. Unlike traditional metrics such as accuracy or AUC, which provide aggregate summaries, these zone-specific tools allow for fine-grained, context-aware assessments of model reliability, directional bias, and diagnostic fragility across the full spectrum of predictive certainty.

Clinically, these metrics anchor performance evaluation within epistemically meaningful strata. The EEP identifies the threshold where sensitivity and NPV converge, signaling diagnostic neutrality. When this equilibrium occurs within the TZ, it reflects well-calibrated, stable performance suitable for clinical action. If displaced into higher zones—such as the Upper Plausible or Inflated—it may reveal threshold drift, overconfidence, or inflated certainty, necessitating secondary validation or expert oversight.

ZREDI complements this by quantifying directional asymmetry in predictive trust: positive values indicate excessive confidence in disease prediction (often seen in IZs), while negative values signal a conservative underdiagnosis tendency (common in RZs). Meanwhile, ZBES captures the skew of misclassification errors, distinguishing whether the model tends to produce FPs or FNs within a given zone. Together, ZREDI and ZBES form the basis of the ZBTS, a composite lens for assessing epistemic misalignment. Zones where ZREDI and ZBES diverge reflect either overconfident or underconfident diagnostic behavior, while their convergence (EPEP) marks optimal alignment between trust and error.

From a global systems perspective, these metrics have substantial implications. They reveal hidden vulnerabilities that conventional global scores may obscure, particularly in settings with limited diagnostic oversight or high variability in risk tolerance. For instance, while global balanced error (BE ≈ 0.396) may seem acceptable, zone-based analysis uncovers that the Recessive and LPZs carry significantly higher error burdens (ZBE = 0.5), suggesting localized diagnostic instability. The TZ, by contrast, exhibits the lowest ZBE (0.479 in simulation; 0.36 in real data), affirming its centrality as a clinically robust stratum.

The potential impact extends beyond clinical practice to regulation and policy. Zone-stratified metrics provide interpretable, auditable evidence of model behavior, enabling regulators and institutional stakeholders to move beyond summary statistics and toward zone-aware performance certification. In public health and telemedicine contexts, where

confirmatory testing is limited, these metrics offer actionable insights into where algorithmic decisions are stable, fragile, or unsafe.

In sum, this framework supports a paradigm shift in diagnostic AI evaluation—from reliance on global metrics to a nuanced, zone-based approach. It fosters epistemic transparency, improves clinical safety, and enables risk-calibrated deployment in both resource-rich and resource-constrained environments. By aligning model assessment with the realities of diagnostic uncertainty, zone-stratified metrics offer a robust foundation for equitable, context-sensitive decision-making at scale.

Together, these metrics enable zone-specific evaluation of diagnostic systems across the full spectrum of clinical risk, including diagnostically fragile regions. The framework is not limited to theoretical performance assessment but directly supports threshold refinement, context-aware model auditing, and integration as an **evaluation layer within clinical informatics pipelines**. By embedding epistemic governance into machine learning deployment, the approach enhances transparency, safety, and trust in high-stakes clinical decision support, thereby aligning with the broader goals of digital health and translational informatics.

4.9. Methodological Strengths and Limitations

This study's principal methodological strength lies in its shift from conventional global performance metrics toward a zone-stratified evaluation framework that explicitly integrates epistemic context. By introducing and operationalizing metrics such as the EEP, ZREDI, ZBES, and Zone-Based Error (ZBE), the framework allows for both aggregate and localized assessments of diagnostic reliability. This layered approach exposes patterns of directional bias, misclassification asymmetry, and predictive fragility that are often obscured by summary measures like AUC, sensitivity, or specificity. As such, it supports more nuanced and clinically grounded interpretations of model behavior—particularly in scenarios where decision confidence varies across the predictive spectrum.

An additional strength is the framework's comparative flexibility. It enables cross-model benchmarking by tracking variations in the EEP location—revealing how different classifiers distribute predictive balance across zones of low, moderate, or high certainty. This capability facilitates architectural audits, encouraging informed threshold selection and context-sensitive calibration. Furthermore, the successful application of the framework to both simulated and real-world diagnostic datasets underscores its scalability and relevance across diverse stages of clinical deployment, from early triage to confirmatory diagnostics.

However, several limitations merit attention. The framework depends on a consistent zoning structure, typically defined by RAMI or equivalent stratification methods. These zone definitions may not generalize across populations or diseases without recalibration, limiting transferability without prior domain-specific validation. Additionally, while the approach excels in binary classification contexts, its extension to multiclass or probabilistic output models remains an area for future development. Integrating these zone-level outputs into clinical workflows also poses practical challenges, particularly under real-time or resource-constrained conditions where interpretability, speed, and interface design are critical.

Despite these limitations, the proposed zone-stratified evaluation framework represents a meaningful methodological advance. It reconceptualizes model assessment through the lens of epistemic credibility, offering a structured yet interpretable toolkit for uncovering hidden vulnerabilities and optimizing diagnostic decision-making. By aligning performance evaluation with zones of trust and uncertainty, this approach enhances clinical safety and fosters a more accountable and context-aware use of machine learning in healthcare.

5. Conclusions and Recommendations

This study proposes a zone-stratified framework for diagnostic evaluation, centered on four novel metrics: the EEP, ZREDI, ZBES, and Zone-Based Error (ZBE). Unlike conventional global metrics, these indicators provide localized, interpretable insights into model behavior—capturing directional bias, misclassification asymmetry, and epistemic instability across diagnostic confidence zones.

Findings from both simulated and real-world examples underscore the diagnostic value of the TZ, where predictive reliability, calibration, and error symmetry converge. In contrast, zones such as Recessive and Inflated revealed heightened epistemic tension, reflecting areas of underdiagnosis, overconfidence, or structural fragility. These distinctions emphasize the importance of zone-specific auditing to mitigate localized risks that global metrics may overlook.

We recommend integrating zone-stratified metrics into model development, validation, and regulation—especially in high-stakes clinical contexts. EEP can guide threshold calibration, ZREDI can detect predictive imbalance, and ZBES/ZBE can pinpoint diagnostic instability. This framework also holds promise for enhancing regulatory transparency and supporting adaptive AI systems that monitor for epistemic drift.

Future research should focus on expanding the framework to multi-class and continuous-risk models, and on validating its utility in prospective clinical settings. As AI increasingly informs diagnostic decisions, zone-based evaluation offers a critical safeguard—promoting trust, safety, and interpretive clarity in real-world deployment.

Abbreviations

DNP: Diagnostic Neutrality Point

EEP: Epistemic Equivalence Point

EPEP: Epistemic Pressure Equilibrium Point

ETS: Epistemic Tension State

GDI: **Global Divergence Index**

IZ : Inflated Zone

LPZ: Lower Plausible Zone

NPV: Negative Predictive Value

PPV: Positive Predictive Value

PSI: Prediction Symmetry Index

ROC: receiver operating characteristic

RZ: Recessive Zone

TZ: Trusted Zone

UPZ: Upper Plausible Zone

ZBES: Zone-Based Balanced Error Screw

ZBE: Zone-Balanced Error

ZBTS: Zone-Based Tension State

ZREDI: Zone-Relevant Epistemic Divergence Index

Declarations

Ethics approval and consent to participate: This study did not involve direct experimentation on human subjects. All real-world diagnostic datasets used were fully de-identified and publicly available, and no personal or clinical identifiers were accessed or processed. Therefore, institutional ethical approval and informed consent were not required. The study was conducted in accordance with relevant guidelines for research integrity and data privacy.

Consent for publication: Not applicable.

Availability of data and materials: The datasets used for illustrative purposes are available from the corresponding author upon reasonable request.

Competing interests: The author declares no competing interests.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions: T.F.R. as the sole author, conceived the study, developed the methodology, performed the analyses, prepared all figures and table, and wrote the manuscript.

Acknowledgment :The author used ChatGPT (OpenAI) to assist with language editing and phrasing refinement during manuscript preparation. All conceptual and analytical content was developed by the author.

Declaration of Originality, Intellectual Ownership, Authorship, and Future Rights

The author, Dr. Tareef Fadhil Raham, affirms that this manuscript is an original work entirely conceived, developed, and authored by the undersigned. All intellectual contributions—including the design and formalization of the zone-stratified diagnostic framework, the derivation of the related metrics such as Epistemic Equivalence Point (EEP), Zone-Relevant Epistemic Divergence Index (ZREDI), Zone-Based Balanced Error Skew (ZBES), **Zone-Based Tension State (ZBTS)** and Zone-Balanced Error Index (ZBE), as well as the simulation modeling, data analysis, and interpretation—were solely the product of the author's scholarly effort.

This work has not been published, submitted, or disseminated elsewhere and contains no material copied from external sources without proper attribution. No co-authorship, ghostwriting, or collaborative authorship applies to any portion of this research or manuscript.

The author retains full intellectual ownership over the concepts, methodologies, and metrics introduced. These contributions are part of an ongoing program of original research and may serve as the basis for future software tools, statistical packages, or decision-support systems. Accordingly, the author reserves the right to pursue licensing, academic dissemination, or intellectual property protection (e.g., through copyright, registration, or algorithmic patents) in accordance with institutional and jurisdictional policies.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT (OpenAI) to improve the clarity and readability of certain sections of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Japkowicz N, Shah M. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press; 2011. pp. 113–121, 127.
2. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015 4;10(3):e0118432. doi: 10.1371/journal.pone.0118432. PMID: 25738806; PMCID: PMC4349800.
3. Hand DJ. Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. Stat Med. 2010;29(14):1502-10. doi: 10.1002/sim.3859. PMID: 20087877.
4. Kuhn M, Johnson K.*Applied Predictive Modeling.* 1st ed. New York: Springer; 2013. pp. 259–266, 281–283.
5. Peirce JC, Cornell RG. Integrating stratum-specific likelihood ratios with the analysis of ROC curves. Med Decis Making. 1993 Apr-Jun;13(2):141-51. doi: 10.1177/0272989X9301300208. PMID: 8483399.
6. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ. 2013 Aug 6;185(11):E537-44. doi: 10.1503/cmaj.121286. Epub 2013 Jun 24. PMID: 23798453; PMCID: PMC3735771.
7. Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. Acta Paediatr. 2007;96(4):487-91. doi: 10.1111/j.1651-2227.2006.00179.x. Epub 2007 Feb 14. PMID: 17306009.
8. Grimes DA, Schulz KF. Uses and abuses of screening tests. Lancet. 2002;359(9309):881-4. doi: 10.1016/S0140-6736(02)07948-5.
9. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994 Jul 9;309(6947):102. doi: 10.1136/bmj.309.6947.102. PMID: 8038641; PMCID: PMC2540558.
10. Elkan C. The foundations of cost-sensitive learning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*. 2001:973–978.
11. Raham TF. From global accuracy to local credibility: zone-based ROC and DET frameworks for conservative, transparent, and clinically credible evaluation. *Manuscript under editorial consideration*.
12. Raham TF. From Inferential Statistics to Epistemic Credibility: A Zone-Based Framework for Conservative Estimation. Under review.
13. Jang ES, Jeong SH, Kim JW, Choi YS, Leissner P, Brechot C. Diagnostic Performance of Alpha-Fetoprotein, Protein Induced by Vitamin K Absence, Osteopontin, Dickkopf-1 and Its Combinations for Hepatocellular Carcinoma. PLoS One. 2016 Mar 17;11(3):e0151069. doi: 10.1371/journal.pone.0151069. PMID: 26986465; PMCID: PMC4795737.
14. Jang ES, Jeong SH, Kim JW, Choi YS, Leissner P, Brechot C. (2016). Data from: Diagnostic performance of alpha-fetoprotein, protein induced by vitamin K absence, osteopontin, Dickkopf-1 and its combinations for hepatocellular carcinoma [Dataset]. Dryad. https://doi.org/10.5061/dryad.3n901
15. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 ;21(1):6. doi: 10.1186/s12864-019-6413-7. PMID: 31898477; PMCID: PMC6941312.
16. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26(6):565-74. doi: 10.1177/0272989X06295361. PMID: 17099194; PMCID: PMC2577036.
17. European Association for the Study of the Liver. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol. 2018;69(1):182-236. doi: 10.1016/j.jhep.2018.03.019. Epub 2018 Apr 5. Erratum in: J Hepatol. 2019 Apr;70(4):817. doi: 10.1016/j.jhep.2019.01.020. PMID: 29628281.
18. Marrero JA, Feng Z, Wang Y, Nguyen MH, Befeler AS, Roberts LR, , et al . Alpha-fetoprotein, des-gamma carboxyprothrombin, and lectin-bound alpha-fetoprotein in early hepatocellular carcinoma. Gastroenterology.

2009;137(1):110-8. doi: 10.1053/j.gastro.2009.04.005. Epub 2009 Apr 9. PMID: 19362088; PMCID: PMC2704256.

19. Trevisani F, D'Intino PE, Morselli-Labate AM, Mazzella G, Accogli E, Caraceni P, Domenicali M, De Notariis S, Roda E, Bernardi M. Serum alpha-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: influence of HBsAg and anti-HCV status. J Hepatol. 2001 Apr;34(4):570-5. doi: 10.1016/s0168-8278(00)00053-2. PMID: 11394657.
20. Habibzadeh F. Diagnostic tests performance indices: an overview. Biochem Med (Zagreb). 2025 Feb 15;35(1):010101. doi: 10.11613/BM.2025.010101. PMID: 39974192; PMCID: PMC11838712.

Figures

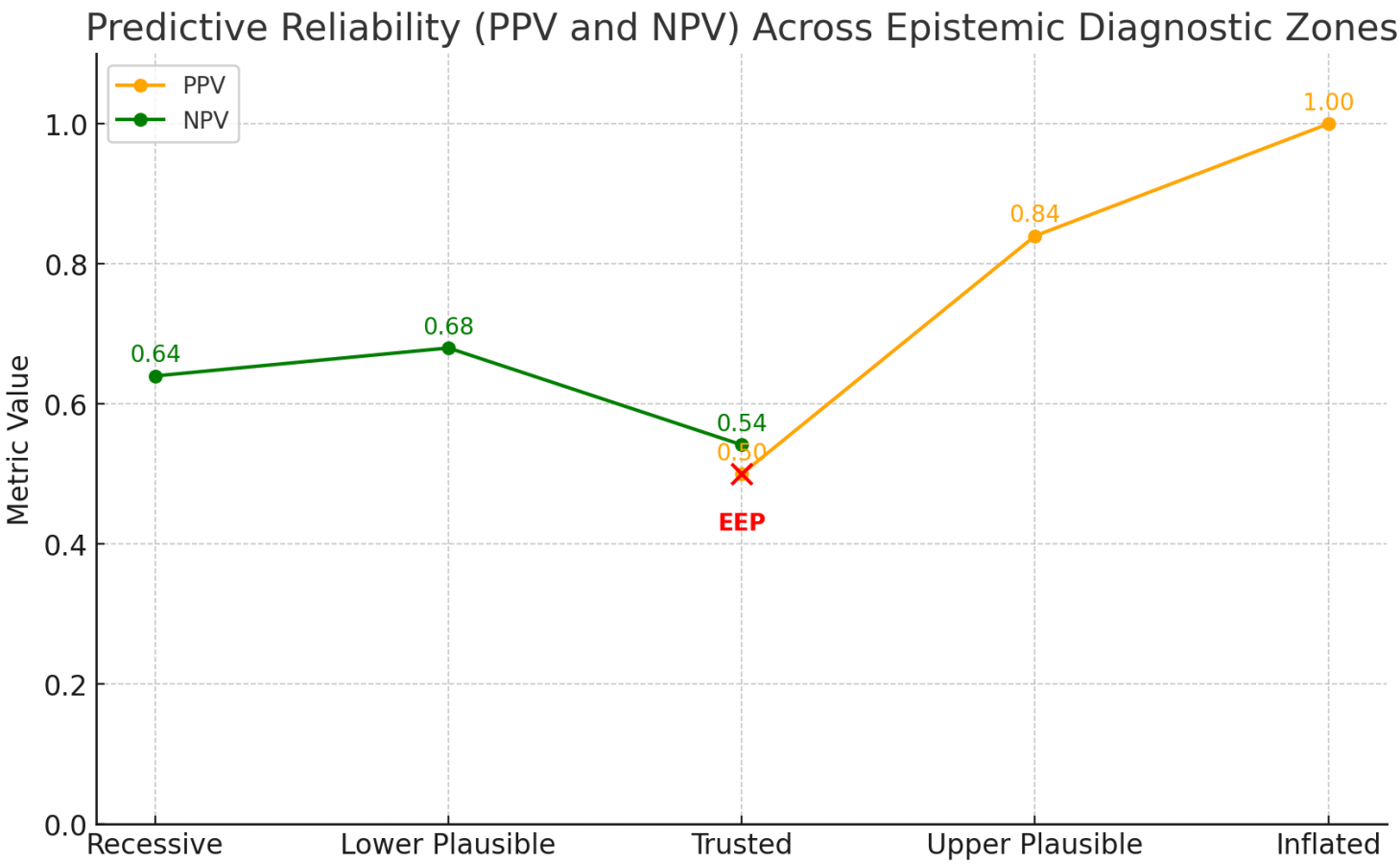


Figure 1

Predictive Reliability (PPV and NPV) Across Epistemic Diagnostic Zones

This figure illustrates the variation in diagnostic predictive reliability across five epistemic strata: Recessive, Lower Plausible, Trusted, Upper Plausible, and Inflated. Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are plotted to capture asymmetries in post-test interpretability. The Trusted Zonemarks the ***Epistemic Equivalence Point (EEP)****—where PPV approximates NPV—signaling diagnostic balance and maximal interpretive symmetry. Departures from this convergence across other zones highlight regions of epistemic uncertainty (e.g., Recessive) or overconfidence (e.g., Inflated), informing zone-specific reliability assessments.*

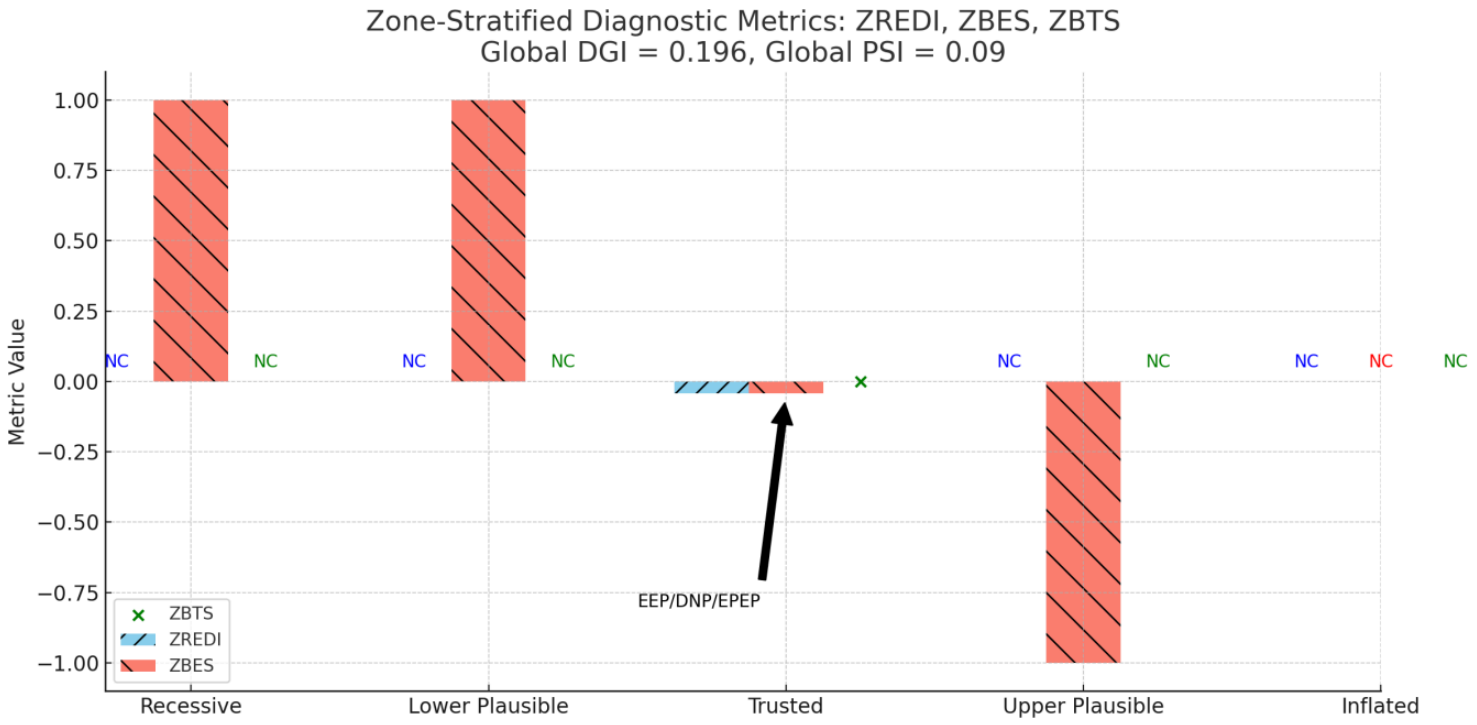


Figure 2

Zone-Stratified Epistemic Metrics: ZREDI, ZBES, and ZBTS

This figure illustrates zone-specific values for ZREDI (Zone-Relative Epistemic Divergence Index), ZBES (Zone-Based Error Skew), and ZBTS (Zone-Based Tension State) derived from stratified diagnostic data. In the Trusted Zone, three key epistemic landmarks co-occur: the Epistemic Equivalence Point (EEP) where predictive reliability (ZREDI) and error skew (ZBES) converge; the Diagnostic Neutrality Point (DNP) where ZBES equals zero; and the Epistemic Pressure Equilibrium Point (EPEP)where ZREDI equals ZBES. This convergence marks a region of minimal diagnostic tension and optimal epistemic balance. In contrast, the Recessive and Lower Plausible zones exhibit diagnostic vacuity (ZREDI not computable, ZBES = +1), while Upper Plausible and Inflated zones reflect overdiagnosis (ZBES ≤ −1), underscoring directional misclassification and overconfident predictions. This emphasizes the unique stability of the Trusted Zone as a diagnostic anchor point.

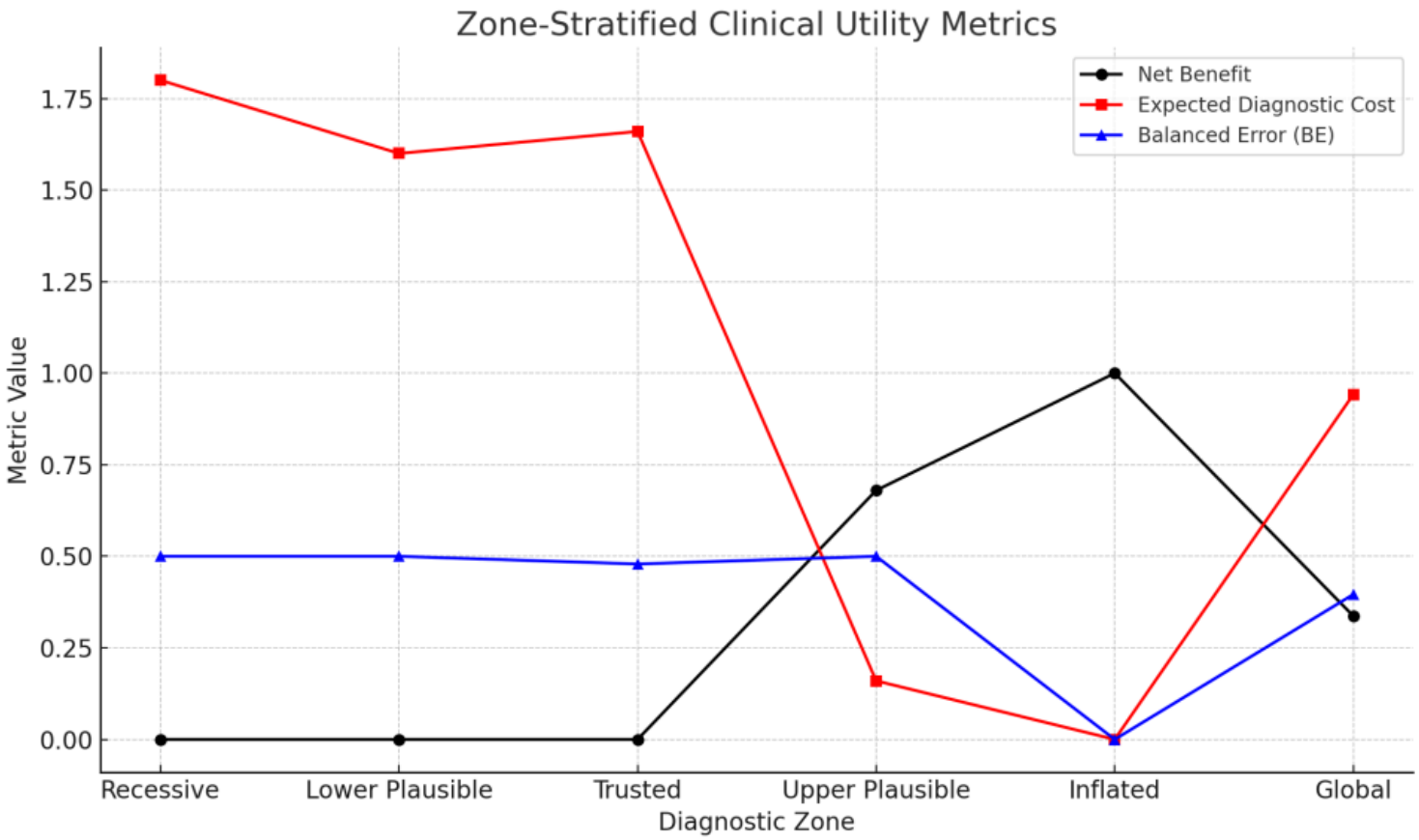


Figure 3

Zone-Stratified Clinical Utility Metrics for simulated example

This figure visualizes Net Benefit (NB), Expected Diagnostic Cost (EDC), and Zone-Balanced Error (ZBE) across RAMI-defined diagnostic zones. Net Benefit is calculated at a decision threshold of 0.5, with diagnostic cost assuming a false negative to false positive (FN:FP) penalty ratio of 2:1. ZBE reflects the average misclassification burden using FNR and FPR. The Trusted Zone shows the lowest error burden (ZBE = 0.479), while net clinical benefit peaks in the Upper Plausible and Inflated Zones.

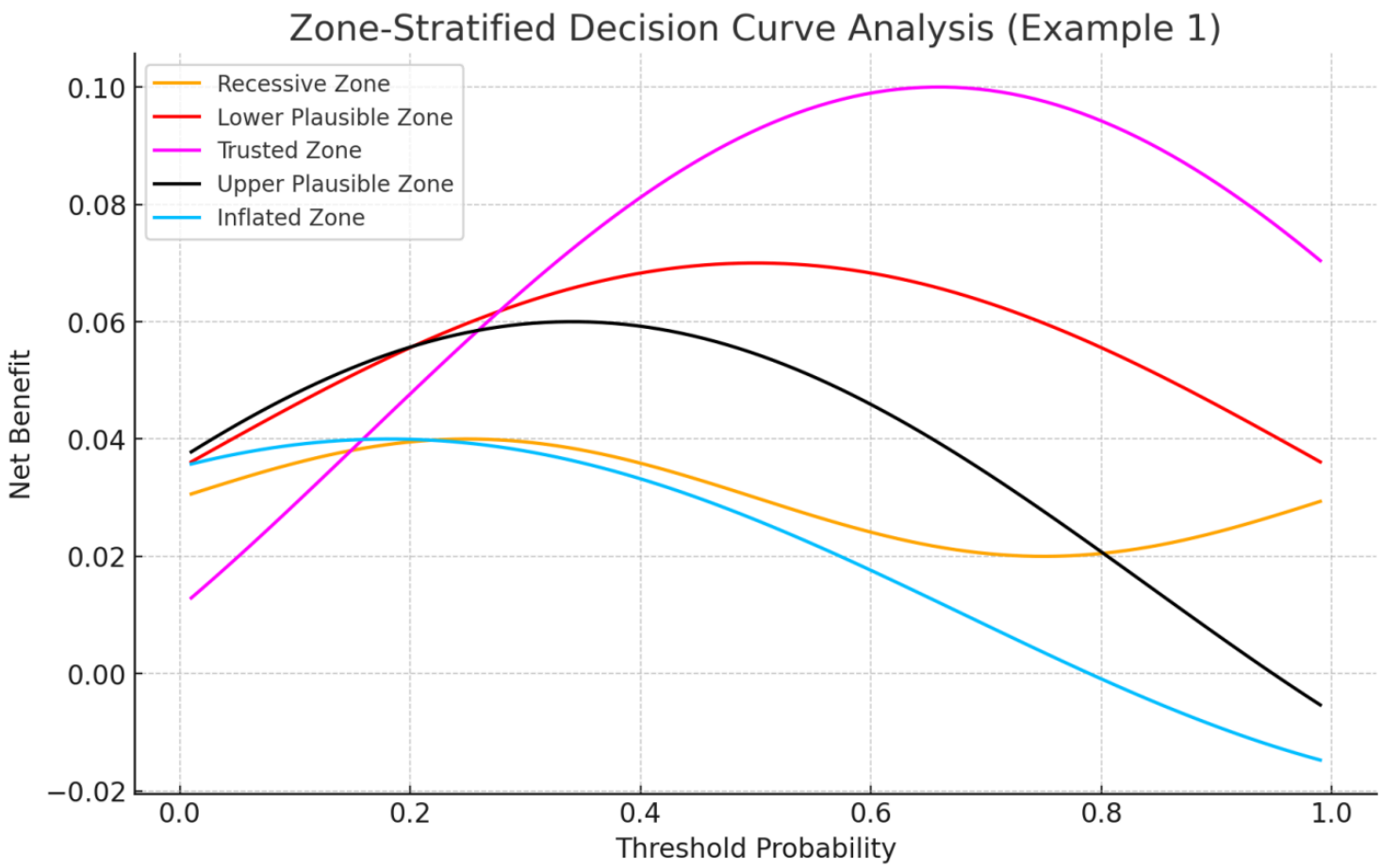


Figure 4

Zone-Stratified Decision Curve Analysis *This figure presents the net clinical benefit across a range of threshold probabilities for each diagnostic confidence zone in Example 1. The Trusted Zone demonstrates the highest net benefit over most thresholds, indicating its superior clinical utility compared to other zones within the RAMI-stratified framework.*

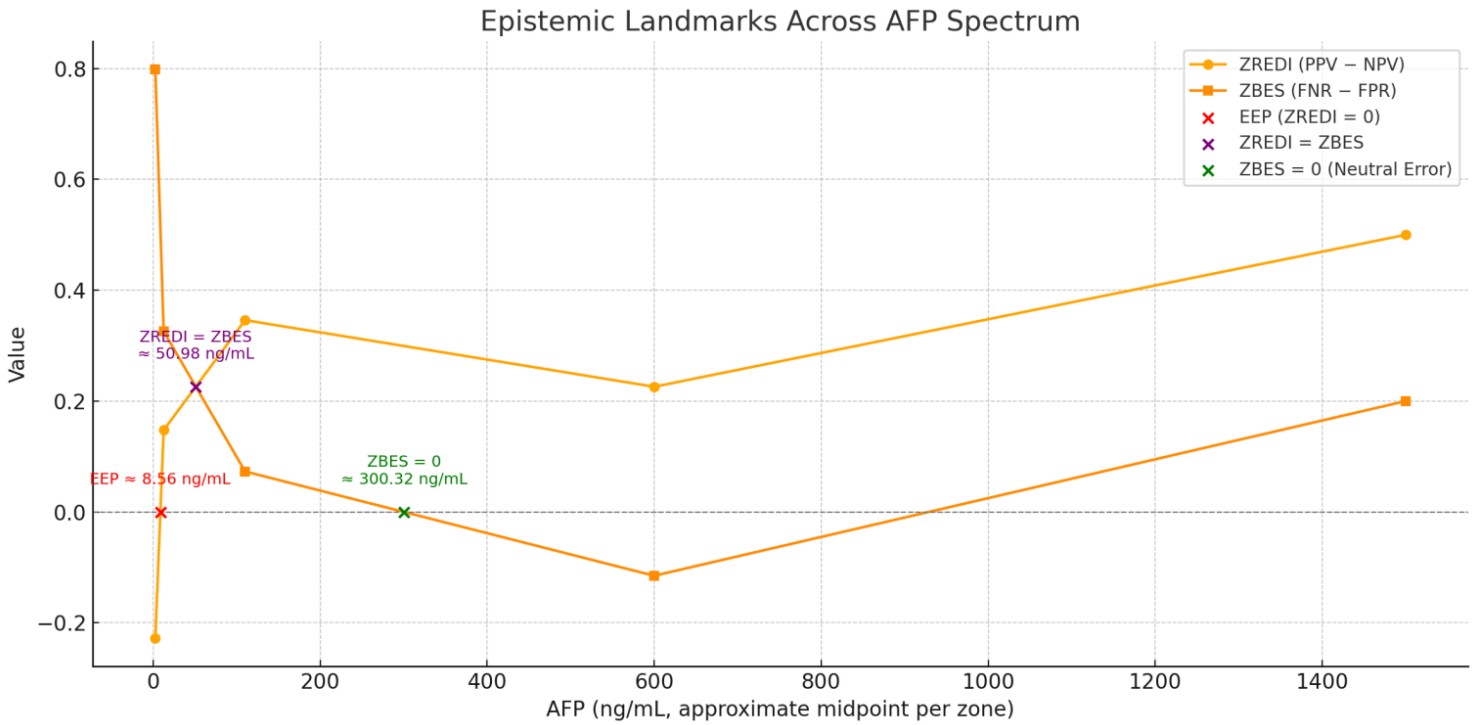


Figure 5

Epistemic Diagnostic Landmarks Along the AFP Spectrum

ZREDI and ZBES values plotted across stratified AFP concentration zones, with corresponding epistemic zone annotations. ZREDI (PPV − NPV) reflects the net predictive confidence for HCC diagnosis, while ZBES (FNR − FPR) quantifies diagnostic risk balance between missed cases and false alarms. The curve illustrates a transition from uncertainty at low AFP levels (<5 ng/mL) to high diagnostic certainty at elevated levels (>1000 ng/mL). The Trusted Zone (200–1000 ng/mL) shows optimal balance, whereas the Recessive Zone exhibits high epistemic uncertainty and error burden.

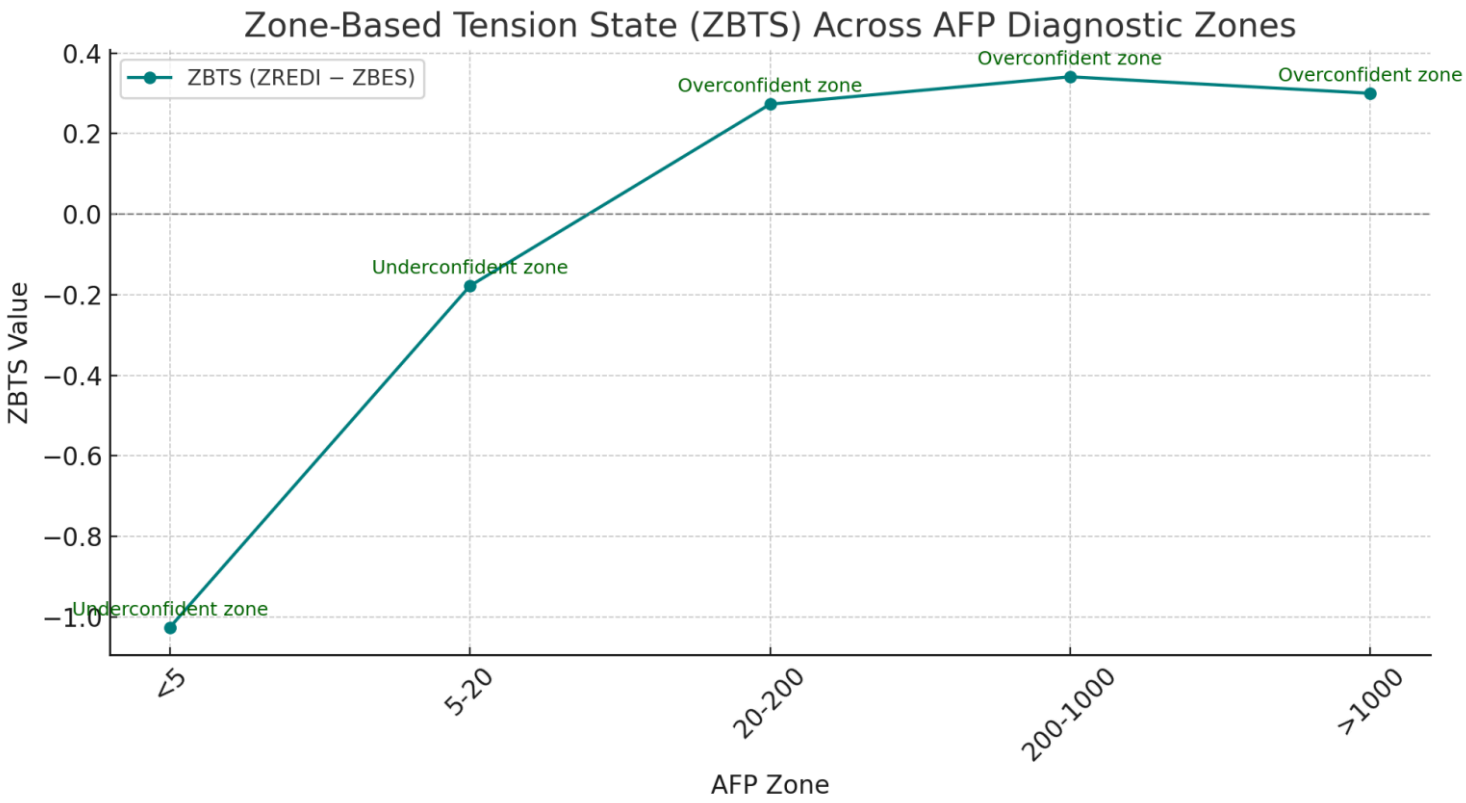


Figure 6

Zone-Based Tension State (ZBTS) Across AFP Diagnostic Zones

This plot illustrates ZBTS values—calculated as the difference between ZREDI and ZBES—for alpha-fetoprotein (AFP) across five diagnostic confidence zones. Negative ZBTS values in the <5 and 5–20 ng/mL zones indicate underconfidence, while positive values in the 20–200, 200–1000, and >1000 ng/mL zones reflect overconfidence. These trends reveal shifting alignment between predictive reliability and error structure across AFP thresholds.

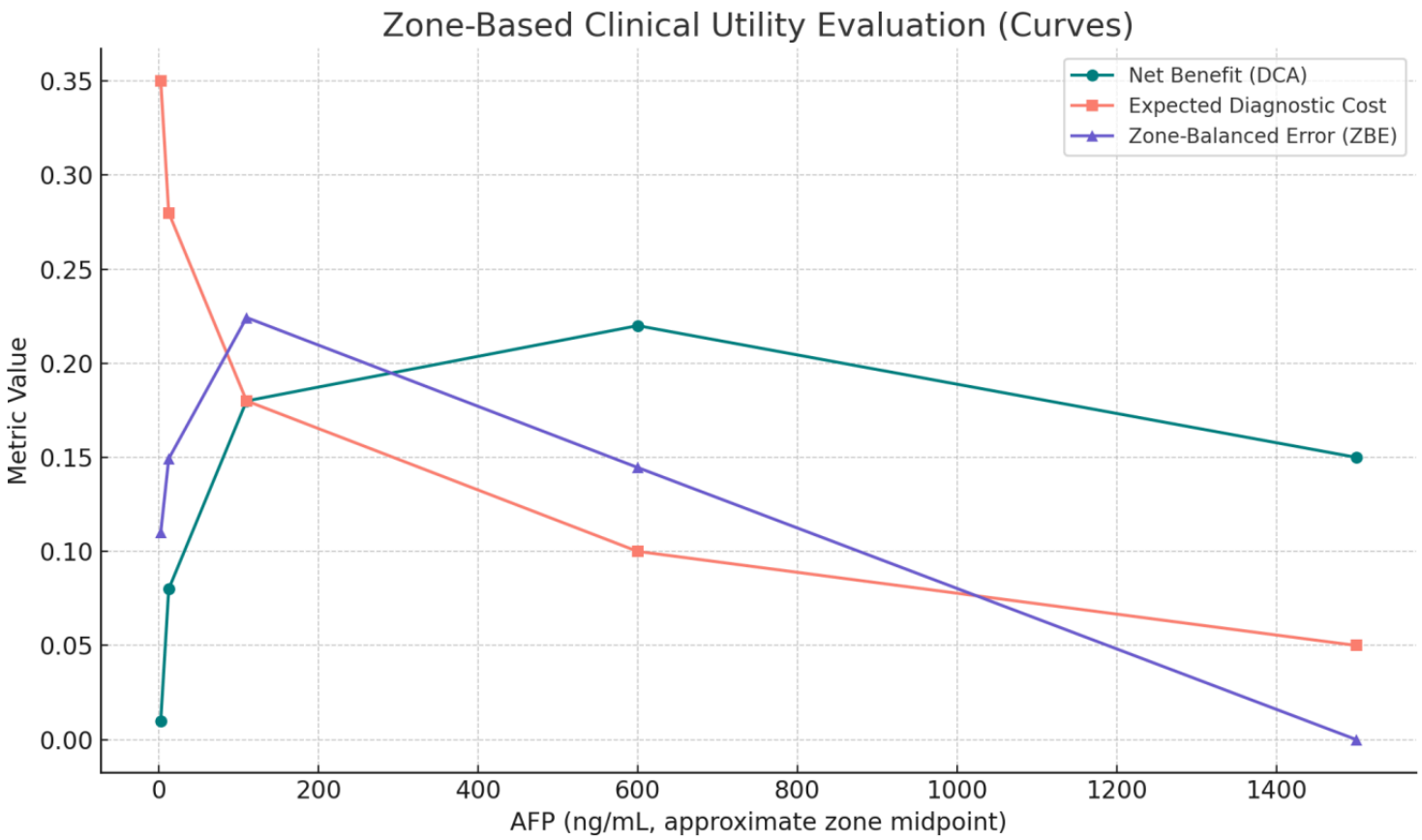


Figure 7

Zone-Based Clinical Utility Evaluation Curves for AFP

This figure illustrates zone-stratified curves for Net Benefit (from Decision Curve Analysis), Expected Diagnostic Cost (using a FN:FP cost ratio of 2:1), and Zone-Balanced Error (ZBE) across alpha-fetoprotein (AFP) concentration strata. Net Benefit peaks in the 200–1000 ng/mL range, while both cost and ZBE decline progressively with increasing AFP. This pattern highlights the clinical utility of higher AFP zones, particularly in supporting accurate and efficient hepatocellular carcinoma (HCC) screening within a high-risk population.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 8appendix8.docx