Quantifying Explainability in Healthcare AI with the Extended Collaborative Intelligence Index (X-CII): A Synthetic Evaluation Framework

Unya Torisan (ORCID: https://orcid.org/0009-0004-7067-9765)
Independent Researcher

Abstract

Human-AI collaboration in healthcare motivates explainable AI (XAI) to promote trust, safety, and regulatory alignment for high-risk systems under the EU AI Act [1] and IMDRF GMLP guidance [2]. We propose the **Extended Collaborative Intelligence Index (X-CII)**, which integrates team *quality* (Q), *effectiveness* (E), and *safety* (S) through a risk-sensitive power mean ($\lambda = 0.25$). To link explainability directly to risk mitigation and address critiques of post-hoc XAI [3], our synthetic evaluation applies a conservative +5% multiplicative uplift in team detectability (d'), reflecting reported 5–10% task-performance gains with XAI [16,17]. Under the equal-variance binormal model this increases AUC from 0.800 to approximately 0.813. The uplift modifies only S while keeping Q and E fixed. Unless otherwise stated, relative percentages are referenced to the better individual agent (human or AI).

Using 10,000 paired Monte Carlo draws with independent skills ($\rho = 0$), the XAI-enhanced team achieved a median relative X-CII of 102.963% (IQR 101.24–104.56%) versus the better individual, outperforming it in 89.7% of cases. Versus an identical team without XAI, median X-CII rose by 0.811% (IQR 0.593–1.003%), with a 100% win rate, isolating explainability's incremental contribution. Under domain shift (AUC = 0.72 with adjusted fidelity/reliance parameters), the median remained 102.82%. Lower integration efficiency ($\eta \leq 0.8$) reduced team performance below baseline, whereas negative skill correlation ($\rho = -0.5$), indicating complementary strengths, increased gains (median 108.66%). Here ρ denotes human–AI skill correlation and η parameterizes integration efficiency ($\eta = 1$ ideal).

The X-CII framework can help quantify how explainability contributes to safe and effective human–AI teamwork and to benchmark compliance-oriented design. Safety normalization ($S = 1 - L/L_{\text{worst}}$) ensures bounded, comparable scores, though it compresses high-performance differences. This work provides no legal advice; consult the official EU AI Act and competent authorities for regulatory interpretation.

Keywords: Human-AI Collaboration, Collaborative Intelligence Metrics, Explainable AI, Healthcare Applications, X-CII Framework, Synthetic Evaluation, Regulatory Compliance
Categories: Health Informatics; Evidence-Based Medicine; Epidemiology

Submission Note: v1: Added dedicated Methods and Results sections for enhanced structure and clarity; recomputed statistics using verified code execution (e.g., baseline median updated to 102.963% to reflect precise floating-point results); incorporated recent XAI advancements from semantic search; confirmed EU AI Act details with official sources. Verified references as of September 29, 2025. No empirical claims; synthetic evaluation only. Reproducible code provided in Appendix A (requires Python 3.10+, NumPy 1.23+, SciPy 1.10+; execution time approximately 0.5 seconds on standard hardware).

Plain-Language Summary for Clinicians

In healthcare, AI tools can help with diagnosis, but their "black box" nature creates risks. If an AI suggests a treatment without a clear reason, how can you confidently explain it to a patient or justify it legally?

The Extended Collaborative Intelligence Index (X-CII) is a "report card" for a human-AI team. It grades not just accuracy (Q) and patient-centered effectiveness (E; e.g., shorter time to diagnosis and treatment, fewer avoidable delays and errors that support recovery, not speed for its own sake), but also how safe and explainable the team's decisions are (S). A high-quality AI explanation is modeled as improved teamwork, providing a conservative 5% boost (in simulation) to the team's diagnostic discrimination (ability to separate diseased from non-diseased cases). In our synthetic evaluation, this translated to a typical (median) improvement of about 3% for the human-AI team compared with the best individual (human or AI), and the team outperformed in about 90% of cases. These results are consistent with recent reviews on human-AI collaboration metrics [4] and align with transparency requirements in the EU AI Act [1] and IMDRF guidelines [2].

By measuring the value of an explanation, X-CII helps build trustworthy AI that supports clinicians and meets legal standards. This is a conceptual framework, not for direct clinical use, and should be calibrated to site-specific data in practice. Obligations under the EU AI Act are being introduced in stages; consult the official text for current dates [1].

Under domain shift, we conservatively adjust parameters by lowering explanation fidelity and increasing the misreliance penalty (to reflect reduced model confidence and a higher risk of erroneous overreliance).

1 Introduction

This work presents a theoretical framework for evaluating collaboration in medicine, grounded in the 2,500-year-old Hippocratic aphorism, which states in its entirety: "Life is short, the Art long, opportunity fleeting, experiment perilous, judgment difficult. The physician must not only be prepared to do what is right himself, but also to make the patient, the attendants, and externals cooperate."[19] We explicitly emphasize this latter clause—cooperation with the patient, the attendants, and externals—as a design requirement for modern healthcare systems. We contend that this complete principle—the acknowledgment of profound limitations followed by the mandate for cooperation—points not to a single tool, but to a necessary ecosystem of partnership. In the modern context, this ecosystem must include not only the physician and the AI system but also the patient, their family, and other attendants, all working as partners in the shared art of healing. Operationally, our simulations instantiate the clinician-AI dyad; however, the Safety component S is constructed to capture whole-team factors (e.g., patient comprehension, caregiver workflows, and organizational context) and can be calibrated accordingly.

AI integration in healthcare enhances diagnostics and treatment but faces opacity challenges, risking trust erosion and non-compliance with regulations such as the EU AI Act, which applies in stages: entry into force 20 days after publication; prohibitions 6 months after entry into force; GPAI obligations 12 months after entry into force; and most high-risk obligations 36 months after entry into force (consult the official text for the current schedule) [1]; see also IMDRF GMLP [2]. XAI mitigates this by providing interpretable insights, enabling clinicians to understand AI decisions and justify them to patients or regulators. This paper extends the Collaborative Intelligence Index (CII) to X-CII, incorporating XAI's impact on team performance via Signal Detection Theory (SDT) metrics [13–15,18]. We focus on synthetic evaluation to isolate mechanisms. We assume equal-variance Gaussian noise and a conservative +5% uplift to the collaborative d' attributable to XAI, based on systematic reviews and experimental studies

reporting 5–10% task-performance gains [16,17]. This uplift models improved detectability through better explanation fidelity and calibrated reliance, reducing errors such as over-reliance on flawed explanations [3,17].

The framework addresses key XAI challenges: post hoc vs. intrinsic interpretability [3], multimodal integration, and uncertainty quantification [6,7]. X-CII quantifies these through the Safety component S (incorporating explanation fidelity F and user reliance R), while maintaining fixed team Quality (Q) and effectiveness (E) in baselines. We demonstrate via Monte Carlo simulations on synthetic data that XAI-induced uplifts yield consistent collaborative advantages, though benefits are compressed under domain shift. This work complements recent XAI reviews [8-11] and SDT applications [13-15,18], providing a reproducible metric to support compliance for high-risk AI. Limitations include synthetic assumptions; future work should adapt and calibrate to real datasets.

2 Methods

As a synthetic study involving no human or animal subjects, this research did not require institutional review board approval. Computational procedures align with good practice guidance for medical AI development (IMDRF GMLP [2]). All simulations used a fixed random seed (42) for reproducibility, and each scenario was run with n = 10,000 replicates using the same seed to enable paired comparisons. Φ denotes the standard normal CDF and Φ^{-1} its inverse; ln denotes the natural logarithm.

2.1 Agent and Team Performance Model

We modeled individual and team performance as follows:

- Individual agent performance (human, AI) is quantified via the detectability index d' derived from the Area Under the ROC Curve (AUC) under the equal-variance binormal assumption: $d' = \sqrt{2} \cdot \Phi^{-1}(\text{AUC})$. For numerical stability, AUC is clipped to [1e-6, 1 - 1e-6] before applying Φ^{-1}.
- For each replicate, agent AUCs are drawn independently from Uniform(0.75, 0.85) and converted to d'.
- Collaborative team detectability d'_{team} follows a correlated Gaussian (Mahalanobis) formulation that captures two collaborative factors: skill correlation ρ and integration efficiency η:

$$d'_{\text{team}} = \eta \cdot \sqrt{\frac{(d'_{\text{h}})^2 + (d'_{\text{ai}})^2 - 2\rho d'_{\text{h}} d'_{\text{ai}}}{1 - \rho^2}}.$$

  For numerical stability, we clip ρ to [-0.999, 0.999] and clip the quantity inside the square root at 0 (as implemented in Appendix A).

Note that η ('integration efficiency') is an algorithmic/compositional parameter and is distinct from E (patient-centered effectiveness). Here, ρ denotes the within-class correlation of human and AI scores under a shared covariance Σ (equal across classes); thus the team discriminability uses the Mahalanobis distance $\Delta\mu^T \Sigma^{-1} \Delta\mu$.

2.2 XAI Impact and Evaluation Metrics

- In X-CII, E denotes patient-centered care effectiveness (timeliness and coordination that support recovery and avoid harm), not speed or throughput. To isolate explainability effects, we model XAI as a conservative +5% multiplicative uplift ($\times 1.05$) on detectability d' in designated routing modes (described below), based on prior studies reporting 5–10%

task-performance gains from XAI integration [16,17]. Unless otherwise stated, this uplift is applied at evaluation time only; it does not alter Q or E (it perturbs S only via d').

- Uplift routing and propagation. We consider three routing modes for how the 5% uplift is applied:
  - Team-only: the uplift is applied only after composition of the human and AI signals, i.e., $d'_{\text{team}} := 1.05 \times \text{team\_d'}(d'_{\text{h}}, d'_{\text{ai}})$, while the individual agents remain unchanged (see Appendix A for the composition model).
  - Single-only: the uplift is applied to the individual agents before composition, i.e., $d'_{\text{h}} := 1.05 \times d'_{\text{h}}$ and $d'_{\text{ai}} := 1.05 \times d'_{\text{ai}}$; because d'_{team} is a deterministic function of $(d'_{\text{h}}, d'_{\text{ai}})$, the collaborative detectability increases indirectly even when no explicit team uplift is applied. Note that under single-only, the 'better individual' denominator is also uplifted, which naturally reduces the relative margin versus the team-only setting.
  - Both: both operations are applied (individual detectabilities are uplifted first, then the composed team detectability is additionally uplifted).

  This design reflects realistic propagation of improved individual signal quality into the collaborative signal. Throughout, Q and E are fixed so that any change in X-CII arises via S through the induced change in expected loss $L(d')$.

- Performance evaluation uses the Extended Collaborative Intelligence Index (X-CII), aggregating Quality (Q), Effectiveness (E), and Safety (S). To isolate S, we fix Q = 0.75 and E = 0.75 in all scenarios (each bounded in [0, 1]).

- Safety S is grounded in Signal Detection Theory (SDT) with decisions evaluated at the Bayes-optimal threshold under equal variances:
  - Expected loss: $L = c_{\text{FN}} \cdot \pi \cdot (1 - \text{TPR}) + c_{\text{FP}} \cdot (1 - \pi) \cdot \text{FPR}$, where π is prevalence P(Y=1), and $c_{\text{FN}}, c_{\text{FP}} \geq 0$ are misclassification costs.
  - Optimal threshold: $\tau^* = 0.5d' + \ln(c_{\text{FP}} \cdot (1 - \pi)/(c_{\text{FN}} \cdot \pi))/d'$, with class-conditional scores N(0,1) vs N(d',1). For numerical stability, divisions by d' and logarithm arguments are guarded to avoid division by zero and nonpositive log inputs (see Appendix A).
  - TPR = 1 - $\Phi(\tau^* - d')$, FPR = 1 - $\Phi(\tau^*)$, where Φ is the standard normal CDF.
  - Normalized base Safety: $S_{\text{base}} = 1 - L/L_{\text{worst}}$, where $L_{\text{worst}} = \max((1-\pi) \cdot c_{\text{FP}}, \pi \cdot c_{\text{FN}})$.
  - Final Safety with fidelity/misreliance penalty scaling and clipping: $S = \text{clip}(S_{\text{base}} \cdot (\alpha + (1 - \alpha) \cdot F) \cdot (1 - R), 0, 1)$, where R represents the misreliance penalty (0: no erroneous over- or under-reliance, 1: full penalty for complete misreliance), with $\alpha = 0.5$. Baseline F (explanation fidelity) and R (misreliance penalty) are set to F = 1.0, R = 0.0 for human/AI alone (ideal, no collaboration-induced risks) and F = 0.95, R = 0.05 for collaboration (minor fidelity loss and reliance risk from team integration); in the domain-shift scenario (Section 2.3), these are adjusted downward for F and upward for R to reflect reduced model confidence and heightened misreliance risks.

- X-CII is computed via a power mean:

$$\text{X-CII} = \left[\frac{Q^\lambda + E^\lambda + S^\lambda}{3} \right]^{1/\lambda},$$

  with $\lambda = 0.25$ (sensitivities at $\lambda \to 0$ geometric, $\lambda = 1$ arithmetic).

- The primary outcome is the relative X-CII: $100 \cdot \text{X-CII}_{\text{collab}} / \max(\text{X-CII}_{\text{human}}, \text{X-CII}_{\text{ai}})$.

Note: Under the single-only routing, the team metric increases indirectly via composition, so "single-only" and "both" can yield numerically close results. The implementation follows this routing exactly (Appendix A).

2.3 Simulation Scenarios and Parameters

- Parameter priors (literature-informed):
  - Prevalence $\pi \sim \text{Beta}(6, 14)$ for moderate class imbalance.
  - Asymmetric costs: $c_{\text{FN}} \sim \text{Uniform}(2, 5)$, $c_{\text{FP}} \sim \text{Uniform}(0.5, 2)$ [13,18].
- Baseline scenario (all unlisted parameters at baseline values):
  - Skill correlation $\rho = 0$ (independent agents).
  - Integration efficiency $\eta = 1.0$.
  - XAI uplift routing: team-only ($d'_{\text{team}} := 1.05 \times \text{team\_d'}(d'_{\text{h}}, d'_{\text{ai}})$; d'_{h} and d'_{ai} unchanged).
- Sensitivity analyses (one-at-a-time, holding others at baseline):
  - XAI application (uplift routing):
    - Team-only: $d'_{\text{team}} := 1.05 \times \text{team\_d'}(d'_{\text{h}}, d'_{\text{ai}})$; d'_{h} and d'_{ai} unchanged.
    - Single-only: $d'_{\text{h}} := 1.05 \times d'_{\text{h}}$ and $d'_{\text{ai}} := 1.05 \times d'_{\text{ai}}$; then $d'_{\text{team}} = \text{team\_d'}(d'_{\text{h}}, d'_{\text{ai}})$ increases indirectly via composition.
    - Both: apply single-only first, then $d'_{\text{team}} := 1.05 \times \text{team\_d'}(d'_{\text{h}}, d'_{\text{ai}})$.
  - Skill correlation ρ: -0.5, 0, +0.5.
  - Integration efficiency η: 0.6, 0.8, 1.0.
  - Power mean exponent λ: 0 (geometric), 0.25 (baseline), 0.5, 1.0 (arithmetic).
- Domain shift scenario:
  - Fixed lower AUC = 0.72 for both agents (converted to d').
  - Adjusted fidelity/reliance: human/AI F = 0.98, R = 0.02; collaboration F = 0.92, R = 0.08.

Unless otherwise stated, all random draws (e.g., agent AUCs, prevalence, costs) are mutually independent. Numerical clipping and guards follow Appendix A.

2.4 Statistical Analysis and Reproducibility

- We report medians and interquartile ranges (IQRs) via NumPy percentiles. Win rate is the proportion of replicates with relative X-CII > 100%.
- For the paired uplift analysis, we compute $\Delta_{\text{team}} = 100 \cdot \text{X-CII}_{\text{team}}(\text{XAI}) / \text{X-CII}_{\text{team}}(\text{no-XAI})$ using identical draws (same seed), and report its median/IQR and win rate (% with $\Delta_{\text{team}} > 100\%$).
- Sensitivity analyses vary one parameter at a time; no hypothesis testing is performed and results are illustrative.
- All simulations are reproducible with Appendix A code; no empirical data were collected. Final S is clipped to [0, 1], and numerical guards (e.g., for logarithms and divisions) are applied as in Appendix A.

3 Results

Synthetic simulations demonstrate consistent collaborative advantages. Baseline relative X-CII: median 102.963% (IQR: 101.236–104.560%), win rate 89.7%. Table 1 summarizes sensitivities.

Table 1: Sensitivity analyses of relative X-CII. Win Rate = % of simulations where collaborative > max(human, AI). All values from 10,000 replicates; illustrative only.

Scenario	Median Relative X-CII (%)	IQR (%)	Win Rate (%)
Baseline (team uplift)	102.963	101.236–104.560	89.7
Uplift single-only	102.130	100.682–103.492	85.4
Uplift both	102.959	101.302–104.510	90.5
$\lambda = 0$ (geometric)	103.051	101.251–104.774	89.7
$\lambda = 0.5$	102.874	101.221–104.359	89.7
$\lambda = 1$ (arithmetic)	102.694	101.193–103.960	89.7
$\eta = 0.6$	95.055	94.093–95.951	0.0
$\eta = 0.8$	99.156	98.439–99.980	24.6
$\rho = -0.5$	108.659	105.466–111.481	99.2
$\rho = 0.5$	99.637	98.892–100.466	37.9
Shift AUC=0.72	102.818	100.339–105.303	78.5

Under domain shift, benefits compress but the median relative X-CII remains above 100%. Results are directionally consistent with prior reports [16,17], though synthetic. Under single-only routing, team performance improves indirectly via composition of uplifted individual signals; single-only and both can therefore be numerically close. In paired comparisons of the XAI-enhanced team versus an identically parameterized team without XAI (identical draws, Q and E held fixed), XAI increased X-CII in every replicate (win rate 100%; median +0.811%, IQR +0.593–+1.003%), isolating explainability's incremental contribution via Safety. This holds by monotonicity under positive cost assumptions ($c_{\mathrm{FP}}, c_{\mathrm{FN}} > 0$) and the equal-variance model: increasing d' strictly reduces L and increases S at the Bayes-optimal threshold.

4 Discussion

X-CII formalizes XAI's value in healthcare human-AI teams, showing modest but consistent uplifts via synthetic evaluation. Integration with multimodal foundation models (MFMs) [7] and uncertainty metrics [6] can enhance conceptual robustness. The high sensitivity of X-CII to η (integration efficiency) and ρ (skill correlation) quantitatively highlights the importance of not only technical performance but also team composition and training in AI implementation. This study is a simulation to verify the theoretical validity and sensitivity of the X-CII framework. The presented values (e.g., collaborative superiority in 90% of cases) indicate potential under the assumed parameters and do not directly predict real clinical outcomes. Future research requires calibration and validation using actual data.

As philosophical perspectives underscore [5], the deployment of explainable AI in healthcare must grapple with deeper ethical dimensions, such as the normative implications of human-AI trust and accountability, to avoid unintended biases in collaborative decision-making. Furthermore, practical implications from emerging AI watch lists [12], highlighting issues like implementation barriers and equity in high-risk systems—emphasize the need for X-CII calibration in diverse clinical settings to ensure equitable and compliant adoption. The introduction of X-CII should be accompanied by careful ethical and organizational considerations, including redefining clinician roles, responsibilities, and addressing risks like deskilling and automation bias. Limitations include synthetic assumptions (e.g., equal-variance, fixed Q/E); empirical

validation is needed. Calibration procedure is outlined in Appendix B. Future work: real-data adaptation, interface factors, and regulatory pilots.

Acknowledgments

The author acknowledges the significant assistance of several generative AI systems—including Grok (xAI), Gemini (Google), ChatGPT (OpenAI), and Claude (Anthropic)—in the research process. Tool names are listed for transparency and do not imply endorsement. These tools were utilized under the author's direct supervision and critical oversight for tasks such as literature synthesis, methodological brainstorming, and code verification. All code was authored and executed by the human author, and no confidential or patient-identifiable data were provided to these systems. The vendors had no role in study design, analysis, or the decision to submit this manuscript. The conceptual framework, all final analyses, interpretations, and the scholarly integrity of this work remain the sole responsibility of the human author. This collaborative process mirrors the principles of trustworthy human-AI synergy that the X-CII framework itself seeks to promote.

Funding

The author received no external funding for this work.

Competing Interests

The author declares no competing interests.

Ethics Statement

This study is based entirely on synthetic data simulations and does not involve human participants, human data, or animals. Therefore, institutional review board (IRB) approval was not required.

Correspondence

Correspondence and requests for materials should be addressed to Unya Torisan (unya-torisan@gmail.com).

A Appendix A: Reproducible Code

```
try:
    from scipy.stats import norm
except Exception as e:
    raise ImportError("SciPy (>=1.10) is required to run Appendix A.
        ↪ Please install scipy. Details: %s" % e)

import numpy as np

def auc_to_dprime(auc):
    return np.sqrt(2) * norm.ppf(np.clip(auc, 1e-6, 1-1e-6))

def team_dprime(d_h, d_ai, rho=0.0, eta=1.0):
    rho = np.clip(rho, -0.999, 0.999)
    num = d_h**2 + d_ai**2 - 2 * rho * d_h * d_ai
```

```
    den = np.maximum(1 - rho**2, 1e-12)
    return eta * np.sqrt(np.maximum(num / den, 0.0))

def expected_loss(d_prime, pi, c_fp, c_fn):
    mu0 = 0.0
    mu1 = d_prime
    delta = np.maximum(mu1 - mu0, 1e-6)
    log_k = np.log(np.maximum(c_fp, 1e-12) * (1 - pi)) -
       ↪ np.log(np.maximum(c_fn, 1e-12) * np.maximum(pi, 1e-12))
    tau_star = 0.5 * (mu0 + mu1) + log_k / delta
    tpr = 1 - norm.cdf(tau_star - mu1)
    fpr = 1 - norm.cdf(tau_star - mu0)
    return c_fn * pi * (1 - tpr) + c_fp * (1 - pi) * fpr

def safety(L, pi, c_fp, c_fn, F, R, alpha=0.5):
    L_allow = (1 - pi) * c_fp
    L_block = pi * c_fn
    L_worst = np.maximum(L_allow, L_block)
    base = 1 - L / np.maximum(L_worst, 1e-6)
    F = np.clip(F, 0.0, 1.0)
    R = np.clip(R, 0.0, 1.0)
    return np.clip(base * (alpha + (1 - alpha) * F) * (1 - R), 0, 1)

def core_xcii(q, e, s, lam=0.25):
    q = np.clip(q, 1e-12, 1)
    e = np.clip(e, 1e-12, 1)
    s = np.clip(s, 1e-12, 1)
    if abs(lam) < 1e-12:
        return np.exp((np.log(q) + np.log(e) + np.log(s)) / 3)
    else:
        return ((q**lam + e**lam + s**lam) / 3)**(1 / lam)

def run_scenario(rho=0.0, eta=1.0, lam=0.25, uplift=1.05,
                 uplift_team=True, uplift_single=False,
                 auc_fixed=None,
                 F_h=1.0, R_h=0.0, F_ai=1.0, R_ai=0.0,
                 F_collab=0.95, R_collab=0.05,
                 seed=42, n=10000, include_stats=True):
    rng = np.random.default_rng(seed)
    auc_h = rng.uniform(0.75, 0.85, n) if auc_fixed is None else
       ↪ np.full(n, auc_fixed)
    auc_ai = rng.uniform(0.75, 0.85, n) if auc_fixed is None else
       ↪ np.full(n, auc_fixed)
    d_h = auc_to_dprime(auc_h)
    d_ai = auc_to_dprime(auc_ai)
    pi = rng.beta(6, 14, n)
    c_fn = rng.uniform(2, 5, n)
    c_fp = rng.uniform(0.5, 2, n)
    if uplift_single:
        d_h_eff = d_h * uplift
        d_ai_eff = d_ai * uplift
    else:
        d_h_eff, d_ai_eff = d_h, d_ai
    d_team = team_dprime(d_h_eff, d_ai_eff, rho=rho, eta=eta)
    if uplift_team:
        d_team = d_team * uplift
    L_h = expected_loss(d_h_eff, pi, c_fp, c_fn)
    L_ai = expected_loss(d_ai_eff, pi, c_fp, c_fn)
```

```
    L_team = expected_loss(d_team, pi, c_fp, c_fn)
    alpha = 0.5
    q = e = np.full(n, 0.75)
    s_h = safety(L_h, pi, c_fp, c_fn, F_h, R_h, alpha)
    s_ai = safety(L_ai, pi, c_fp, c_fn, F_ai, R_ai, alpha)
    s_c = safety(L_team, pi, c_fp, c_fn, F_collab, R_collab, alpha)
    core_h = core_xcii(q, e, s_h, lam)
    core_ai = core_xcii(q, e, s_ai, lam)
    core_c = core_xcii(q, e, s_c, lam)
    rel = 100 * core_c / np.maximum(core_h, core_ai)
    median = np.median(rel)
    iqr = np.percentile(rel, [25, 75])
    win_rate = (rel > 100).mean() * 100
    if include_stats:
        mean = np.mean(rel)
        std = np.std(rel)
        return median, iqr, win_rate, mean, std
    return median, iqr, win_rate

def team_xcii_array(seed=42, n=10000, rho=0.0, eta=1.0, lam=0.25,
                    uplift=1.05, uplift_team=True, uplift_single=False,
                    auc_fixed=None, F_collab=0.95, R_collab=0.05):
    rng = np.random.default_rng(seed)
    auc_h = rng.uniform(0.75, 0.85, n) if auc_fixed is None else
       ↪ np.full(n, auc_fixed)
    auc_ai = rng.uniform(0.75, 0.85, n) if auc_fixed is None else
       ↪ np.full(n, auc_fixed)
    d_h = auc_to_dprime(auc_h); d_ai = auc_to_dprime(auc_ai)
    pi = rng.beta(6,14,n); c_fn = rng.uniform(2,5,n); c_fp =
       ↪ rng.uniform(0.5,2,n)
    if uplift_single:
        d_h, d_ai = d_h * uplift, d_ai * uplift
    d_team = team_dprime(d_h, d_ai, rho=rho, eta=eta)
    if uplift_team:
        d_team *= uplift
    L_team = expected_loss(d_team, pi, c_fp, c_fn)
    s_team = safety(L_team, pi, c_fp, c_fn, F_collab, R_collab,
       ↪ alpha=0.5)
    q = e = np.full(n, 0.75)
    return core_xcii(q, e, s_team, lam)

# Paired uplift analysis (team-only)
core_with = team_xcii_array(uplift_team=True, uplift_single=False,
   ↪ seed=42)
core_without = team_xcii_array(uplift_team=False, uplift_single=False,
   ↪ seed=42)
rel_pair = 100 * core_with / np.maximum(core_without, 1e-12)
print("Paired uplift (team-only):",
      np.median(rel_pair),
      np.percentile(rel_pair,[25,75]),
      (rel_pair > 100).mean() * 100)

# Baseline and sensitivities
print("Baseline:", run_scenario())
print("$\lambda$=0 (geom):", run_scenario(lam=0))
print("$\lambda$=0.5:", run_scenario(lam=0.5))
print("$\lambda$=1.0 (arithmetic):", run_scenario(lam=1.0))
print("$\eta$=0.6:", run_scenario(eta=0.6))
```

```
print("$\eta$=0.8:", run_scenario(eta=0.8))
print("$\rho$=-0.5:", run_scenario(rho=-0.5))
print("$\rho$=0.5:", run_scenario(rho=0.5))
print("uplift single-only:", run_scenario(uplift_team=False,
    ↪ uplift_single=True))
print("uplift both:", run_scenario(uplift_team=True,
    ↪ uplift_single=True))
# Domain shift (AUC=0.72) + F/R adjustments
print("Shift AUC=0.72:", run_scenario(auc_fixed=0.72,
      F_h=0.98, R_h=0.02, F_ai=0.98, R_ai=0.02, F_collab=0.92,
         ↪ R_collab=0.08))
```

B Appendix B: Calibration Checklist and Alternatives

This appendix outlines proposed methodology for calibrating X-CII in real environments.

- ρ estimation: Compute Pearson/Spearman on z-transformed paired scores; stratify by case-mix; robustify with bootstrapping and binormal ROC fitting for covariance; pool within-class covariances (estimate ρ_0, ρ_1 and variances per class 0/1, weight by sample size under equal-variance); check equal-variance via slope $s \approx 1$. Use split cross-validation to avoid leakage in experiments collecting both human confidence scores and AI outputs on the same cases.
- F/R calibration: Counterfactual surveys, calibration curves. For α: SOP compliance rate (e.g., documentation adherence threshold 0.8, double-check rates >0.9); F: explanation fidelity from blind evaluations (threshold 0.9); R: over-/under-reliance from behavioral metrics (threshold 0.1), normalized.
- Alternative Safety normalizations (not used; for illustration): Log-compressed: $1 - \log(1 + L/L_{\text{ref}})/\log(1 + L_{\text{worst}}/L_{\text{ref}})$; Power-transformed: $1 - (L/L_{\text{worst}})^{0.5}$ (expands high-performance differences).

References

[1] European Parliament and Council. (2024). Regulation (EU) 2024/1689... Official Journal of the European Union, L 206, 12.7.2024, p. 1-252. CELEX: 32024R1689.

[2] IMDRF. (2025). Good machine learning practice for medical device development: Guiding principles. IMDRF/AIML WG/N88 FINAL:2025. Published January 29, 2025.

[3] Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5), 206-215. DOI: 10.1038/s42256-019-0048-x.

[4] Fragiadakis, G., et al. (2024). Evaluating Human-AI Collaboration: A Review and Methodological Framework. arXiv:2407.19098 [cs.HC].

[5] Sun, Q., et al. (2024). Explainable Artificial Intelligence for Medical Applications: A Review. arXiv:2412.01829 [cs.AI].

[6] Farquhar, S., et al. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), 625-630. DOI: 10.1038/s41586-024-07421-0.

[7] Lu, M. Y., et al. (2024). A multimodal generative AI copilot for human pathology. *Nature*, 634(8033), 466-473. DOI: 10.1038/s41586-024-07618-3.

[8] Giorgetti, C., et al. (2025). Healthcare AI, explainability, and the human-machine relationship: a (not so) novel practical challenge. *Frontiers in Medicine*, 12:1545409. DOI: 10.3389/fmed.2025.1545409.

[9] Muhammad, D., et al. (2024). Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. *Computational and Structural Biotechnology Journal*, 24:542-560. DOI: 10.1016/j.csbj.2024.08.005.

[10] El-Geneedy, M., et al. (2025). A comprehensive explainable AI approach for enhancing transparency and interpretability in stroke prediction. *Scientific Reports*, 15:26048. DOI: 10.1038/s41598-025-11263-9.

[11] Vani, M. S., et al. (2025). Personalized health monitoring using explainable AI: bridging trust in predictive healthcare. *Scientific Reports*, 15:31892. DOI: 10.1038/s41598-025-15867-z.

[12] CADTH. (2024). CADTH's 2024 Watch List. NCBI Bookshelf ID: NBK602921.

[13] Kovesdi, C., et al. (2025). Application of Signal Detection Theory in Evaluating Trust of Information Produced by Large Language Models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. DOI: 10.1177/10711813251368829.

[14] Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley.

[15] Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide (2nd ed.)*. Lawrence Erlbaum Associates. DOI: 10.4324/9781410611147.

[16] Ameen, S. Y., et al. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. *Computers in Biology and Medicine*, 166:107543. DOI: 10.1016/j.compbiomed.2023.107543.

[17] Martell, M. J., et al. (2024). Mitigative Strategies for Recovering From Large Language Model Trust Violations. DOI: 10.1177/15553434241303577.

[18] Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes*, 60(1), 1-13. DOI: 10.1006/obhd.1994.1072.

[19] Hippocrates. Aphorisms I.1. English translation by W. H. S. Jones. Loeb Classical Library, Harvard University Press.