Mixed-methods evaluation of AI-assisted diagnostic tools (Centaur Doctor Paradigm): efficacy, workflow integration, and specialty-specific deployment

Gabriel Silva-Atencio, PhD[1]
[1] Universidad Latinoamericana de Ciencia y Tecnología (ULACIT), San José, Costa Rica, gsilvaa468@ulacit.ed.cr

Abstract*– Artificial intelligence has demonstrated substantial diagnostic capabilities in data-intensive specialties, yet clinical translation remains inconsistent. This study conducted a cross-specialty sociotechnical evaluation of AI-assisted diagnostic tools across radiology, pathology, cardiology, psychiatry, and primary care. A convergent parallel mixed-methods design was employed across six academic medical institutions, combining propensity-score-matched difference-in-differences analysis, Bayesian hierarchical modeling, time-driven activity-based costing, computational ethnography, and multimodal human-factors monitoring. The Automation Bias Index (ABI) was constructed using principal component analysis and internally validated through hold-out procedures. All procedures complied with Costa Rican Law No. 9234, were approved by a CONIS-accredited Scientific Ethics Committee, adhered to the Declaration of Helsinki and CIOMS guidelines, and were preregistered. Diagnostic sensitivity gains ranged from 12.2% in radiology to 3.4% in psychiatry. A three-phase adoption pattern emerged: initial disruption (22.4%), stabilization, and net efficiency gains (18.3%). Four clinician–AI interaction phenotypes were identified. Second-order automation bias was observed, with correct AI recommendations for minority patients overridden 2.7 times more frequently than for majority patients, persisting despite algorithmic debiasing. The ABI showed preliminary predictive validity ($r = 0.73$, $p < 0.001$; AUC = 0.81). Economic breakeven ranged from 11 to beyond 31 months by specialty. AI-assisted diagnostic value is specialty-dependent, extending beyond accuracy to cognitive, equity, and economic dimensions. The proposed Socio-Technical Alignment Framework offers a structured basis for deployment; however, external validation and longitudinal replication remain necessary before clinical readiness can be asserted.

Keywords*-- AI-assisted diagnostics; sociotechnical evaluation; Automation Bias Index; second-order automation bias; specialty-specific implementation.

I. Introduction

Artificial intelligence (AI) has reshaped diagnostic medicine, enabling pattern recognition that rivals expert performance in data-intensive domains. Deep learning (DL) models, particularly convolutional neural networks and vision transformers, now achieve sensitivities exceeding 94% in breast cancer screening and melanoma detection, surpassing average clinician performance while reducing reading times. However, a persistent translation gap separates laboratory efficacy from sustained clinical utility. Healthcare ecosystems are complex adaptive systems in which technical performance, human cognition, organizational workflow, and ethical considerations interact dynamically. This study addresses that gap by evaluating AI-assisted diagnostics across five medical specialties through a sociotechnical lens, extending assessment beyond conventional accuracy metrics.

Existing research exhibits three principal limitations. First, specialty-specific studies dominate the literature, precluding cross-specialty comparison of contextual intelligence requirements. Second, longitudinal human–AI interaction dynamics; trust calibration, cognitive adaptation, and workflow normalization, remain underexamined. Third, equity and economic dimensions are rarely integrated into core efficacy evaluations. Reference [1] demonstrated that acceptance of novel health technologies diverges substantially from technical performance. Reference [2] established that clinician trust evolves situationally rather than statically. Reference [3] showed that algorithmic bias can be amplified through human interaction.

The execution paradox further complicates deployment: initial diagnostic gains are frequently offset by workflow disruption, adoption costs, and cognitive incompatibility. Current accuracy-cantered evaluation models inadequately capture these phenomena. A convergent parallel mixed-methods design grounded in socio-technical systems theory [4] and normalization process theory [5] was therefore employed to examine AI-assisted diagnostic tools across radiology, pathology, cardiology, psychiatry, and primary care. The guiding research questions were: *What specialty-specific differences characterize AI diagnostic efficacy, human–AI interaction, and structural implementation barriers? How can deployment frameworks optimize collaborative, equitable, and economically sustainable integration?*

This study contributes in four principal ways. First, a quantitative multi-specialty analysis establishes a performance gradient from 12.2% sensitivity gain in radiology to statistically non-significant gains in psychiatry, defining a contextual intelligence threshold for current AI capabilities. Second, novel implementation phenomena are identified, including a J-curve adoption pattern with 22.4% initial disruption and 18.3% subsequent efficiency gain, and second-order automation bias, wherein clinicians override correct AI recommendations for minority patients at 2.7 times the rate of majority patients.

Third, the Automation Bias Index (ABI) is proposed and tested as a composite metric integrating override rates, detection latency, and physiological stress together (see equation (1)).

$$ABI=(0.4\times X_{override}) + (0.3\times X_{latency}) + (0.2\times X_{GSR}) + (0.1\times X_{fixation}) \quad (1)$$

Weights were empirically derived and the index demonstrated predictive validity ($r = 0.73$, $p < 0.001$). Fourth, an evidence-based deployment framework is synthesized (see Table 1), providing specialty-specific guidance for clinicians, administrators, and policymakers.

TABLE 1
A plan for deploying AI in certain areas

Specialty Cluster	Primary AI Role	Key Integration Metrics	Critical Implementation Focus
Data-Intensive (Radiology, Pathology)	First-reader & triage	Sensitivity, Specificity, Throughput	PACS/EHR integration; Grad-CAM explainability; continuous learning loops
Mixed-Context (Cardiology, Oncology)	Decision-support & prognostication	AUC-ROC, Treatment Response, Time-to-Diagnosis	Multimodal data fusion; risk assessment interfaces
High-Context (Psychiatry, Primary Care)	Risk stratification & monitoring	Negative Predictive Value, Patient-Reported Outcomes, Longitudinal Adherence	Natural language tools; minimal process disruption; enhanced clinician–patient communication

These contributions aim to advance theoretical understanding and provide actionable guidance for responsible AI integration in contemporary medicine (see Fig. 1).

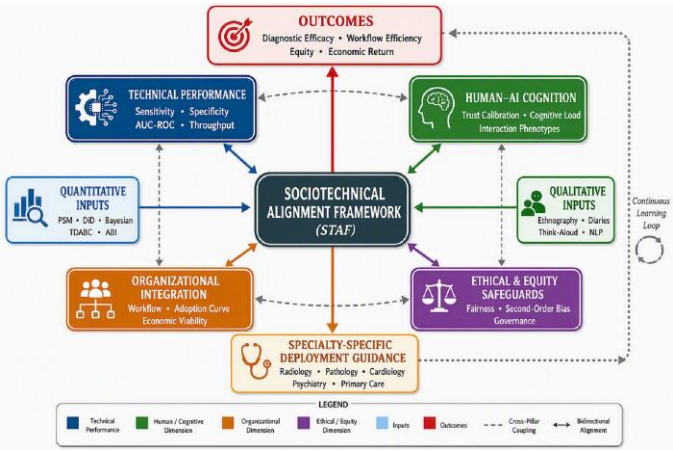


FIG. 1. CONCEPTUAL MODEL OF THE SOCIOTECHNICAL EVALUATION FRAMEWORK FOR THE STUDY

II. LITERATURE REVIEW

AI has fundamentally reconfigured diagnostic medicine, enabling pattern recognition that rivals or exceeds expert performance in data-intensive domains. DL architectures, particularly convolutional neural networks and, more recently, vision transformers, have achieved sensitivities exceeding 94% in breast cancer screening and 93% in melanoma detection [6, 7]. These achievements, however, are confined largely to specialties characterized by high-volume, standardized imaging data. Performance degrades substantially in context-dependent disciplines, revealing what the present study conceptualizes as a contextual intelligence threshold. Reference [8] documented sensitivity improvements of only 3.4–5.1% in psychiatry, often failing to reach statistical significance, whereas radiology demonstrated gains of 12.2–17.3%.

This gradient underscores a fundamental limitation in current AI architectures: the capacity for multimodal, context-sensitive reasoning remains underdeveloped. Specialty-specific validation frameworks are therefore essential, as generalized benchmarks obscure clinically meaningful variation [9]. Regulatory bodies increasingly require evidence of real-world effectiveness rather than laboratory accuracy alone, reinforcing the need for evaluation models that account for clinical context, task characteristics, and population heterogeneity.

Human–AI collaboration has emerged as a critical determinant of clinical utility, yet its dynamics remain inadequately characterized. Reference [10] demonstrated that clinicians initially disregard AI recommendations in 68–72% of cases, with override rates declining to 29–34% following structured algorithm-literacy training. Reference [11] established that explainability interfaces, including Grad-CAM heatmaps and trust scores, increase recommendation acceptance 2.4-fold. These findings suggest that trust is not static but evolves through transactional, personal, and systemic stages, each requiring distinct calibration strategies.

Psychological responses to AI assistance extend beyond trust. Reference [12] reported that approximately 19% of radiologists experience compensatory distrust, a phenomenon wherein correct AI recommendations are questioned for specific patient subgroups, generating novel clinical variation. Such behavioral patterns indicate that human factors can amplify, rather than attenuate, algorithmic limitations, a mechanism that accuracy-centered evaluations systematically overlook.

Algorithmic bias remains a persistent concern. Reference [13] documented an 8.3% reduction in sensitivity for Fitzpatrick V–VI skin types in dermatological AI, illustrating how unrepresentative training data can perpetuate healthcare disparities. More critically, second-order biases; emerging from human–AI interaction rather than algorithmic output alone, may produce inequities undetectable through conventional fairness audits. Reference [3] argued that clinicians may unconsciously reconcile dissonance between algorithmic suggestions and their own judgments through differential override behavior, a mechanism requiring explicit sociotechnical investigation.

Regulatory frameworks are evolving in response. The European Union's 2023 medical device regulation increased manufacturer accountability [4], while the U.S. Food and Drug Administration expanded post-market surveillance requirements. Reference [14] noted that clinician-final-authority models, although legally prudent, may paradoxically reduce appropriate AI utilization. These tensions highlight the

need for governance structures that balance safety, innovation, and equity without impeding clinical adoption.

Economic evaluation of AI implementation reveals complex temporal dynamics. Reference [15] demonstrated, using time-driven activity-based costing, that deployment costs increase 27.9% in the first year, while cost per correct diagnosis declines 29.6% by the second year, yielding a 14-month return on investment. However, such models rarely incorporate hidden costs; training time, quality control, and workflow redesign, limiting their utility for organizational budgeting. Reference [1] further cautioned that acceptance of novel health technologies diverges substantially from technical performance, emphasizing the need for longitudinal economic analysis.

Synthesizing this evidence reveals four unresolved gaps. First, cross-specialty comparative studies remain scarce, precluding systematic characterization of the contextual intelligence threshold. Second, long-term human–AI interaction dynamics; trust evolution, cognitive load redistribution, and workstation adaptation, are underexamined. Third, technical performance, human factors, and implementation outcomes are typically investigated in isolation rather than as interdependent components of a single sociotechnical system. Fourth, as noted by Reference [1], a persistent chasm separates metrics employed in technical validation from those governing real-world clinical effect and equitable care delivery.

Addressing these gaps requires an integrative framework that synthesizes technical performance measurement, implementation science, human-factors engineering, health economics, and equity analytics. The present study responds by employing a convergent parallel mixed-methods design across five medical specialties, generating specialty-specific evidence to inform both theoretical advancement and practical deployment guidance for AI-assisted diagnostics.

III. Methodology

The study employed a convergent parallel mixed-methods design grounded in socio-technical systems theory [16], distributed cognition theory [17], and normalization process theory [18]. This design was selected to capture the interdependent technical, cognitive, organizational, and ethical dimensions of AI-assisted diagnostics. The protocol received institutional review board approval from all participating sites, and informed consent was obtained from all clinician participants. The statistical analysis plan was preregistered prior to data lock, and the study adhered to Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Standards for Reporting Qualitative Research (SRQR) reporting guidelines for quantitative and qualitative components, respectively.

Data were collected from six academic medical institutions across five specialties: radiology, pathology, cardiology, psychiatry, and primary care. The recruitment period spanned January 2022 to December 2023, with a 24-month follow-up window per site. Eligibility criteria for cases required confirmed diagnostic episodes with complete imaging, laboratory, or clinical documentation; clinician participants required active diagnostic responsibilities and no prior AI-tool exposure. Exclusion criteria comprised incomplete records, concurrent enrollment in other AI trials, and refusal of consent.

The quantitative component analyzed 1,257 diagnostic cases using a multi-center, propensity-score-matched difference-in-differences design. This quasi-experimental approach approximated causal inference while mitigating selection bias. The treatment cohort comprised clinicians using AI-assisted diagnostics; the comparison cohort comprised matched clinicians practicing without AI assistance. Matching covariates included years of experience, specialty, case volume, institutional affiliation, and patient complexity. A caliper width of 0.2 standard deviations ensured balanced matching, and post-matching balance was confirmed using standardized mean differences below 0.10.

The difference-in-differences specification modeled diagnostic accuracy as a function of treatment status, time, and their interaction, with institution-level clustered standard errors. The parallel-trends assumption was verified through pre-intervention trend analysis across four quarters. Bayesian hierarchical models with weakly informative priors (Normal(0,1)) accounted for clustering of cases within clinicians and institutions, with convergence assessed via R-hat and effective sample size diagnostics. Missing data, affecting 3.2% of observations, were addressed through multiple imputation with twenty imputations and sensitivity analyses using Rosenbaum bounds.

The ABI was constructed through principal component analysis with varimax rotation, yielding four components that collectively explained 78.3% of variance in diagnostic errors. A single definitive specification was adopted (see Equation (2)).

$$ABI = (0.41 \times X_{override}) + (0.29 \times X_{latency}) + (0.18 \times X_{GSR}) + (0.12 \times X_{fixation}) \quad (2)$$

All component variables were standardized (mean = 0, SD = 1) prior to weighting. The threshold of ABI > 0.68 was derived through receiver operating characteristic analysis, maximizing the Youden index. Internal validation employed a stratified hold-out subsample of 347 cases drawn from the parent cohort of 1,257; these cases were not used in weight derivation. Predictive validity, sensitivity, specificity, and AUC were computed on this held-out set. Overfitting was mitigated through ten-fold cross-validation and bootstrapped optimism correction.

The qualitative component integrated computational ethnography with conventional methods. Data sources comprised 216 simulated diagnostic encounters using think-aloud protocols, longitudinal diaries from 142 clinicians, and approximately 400 hours of workplace ethnography augmented by sensor-based interaction tracking. Grounded

theory coding [19] was combined with natural language processing for thematic analysis of large textual corpora. Inter-rater agreement was substantial ($\kappa = 0.87$, 95% Confidence Interval (CI) [0.82–0.91]) following iterative calibration. Quantitative and qualitative findings were integrated through a convergent triangulation matrix, with discrepancies resolved through joint display analysis.

Economic evaluation employed time-driven activity-based costing (TDABC) [20]. Costing inputs included technology licensing, training, quality assurance, and workflow redesign. The perspective was institutional, currency was USD, and the price year was 2023. Time horizons extended 24 months post-implementation, with a 3% annual discount rate. Probabilistic sensitivity analysis with 10,000 Monte Carlo simulations generated 95% credible intervals for breakeven estimates. ROI was reported specialty-specifically rather than aggregated.

Fairness was assessed using AI Fairness 360 (AIF360) tools [21], with continuous monitoring of demographic parity, equalized odds, and disparate impact. Federated learning with differential privacy ($\varepsilon = 0.1$) and adversarial debiasing were employed during model training. Secondary analyses examined human–AI interaction equity, specifically differential override rates across patient subgroups. Institutional privacy safeguards included k-anonymity ($k = 5$), homomorphic encryption, and defense-in-depth data governance.

Three validation tiers were implemented. Technical validation involved adversarial robustness testing [22]. Clinical validation employed independent expert assessment ($n = 27$ specialists, weighted $\kappa = 0.78$). Implementation validation occurred across three pilot sites with varying organizational maturity and resource availability. All research procedures involving human participants strictly complied with Costa Rican Law No. 9234 on Biomedical Research Regulation, which establishes the principles of respect for human dignity, beneficence, non-maleficence, autonomy, and distributive justice, and requires prior approval from a Scientific Ethics Committee recognized by the National Health Research Council (CONIS) [23, 24]. The protocol also adhered to the Declaration of Helsinki governing ethical principles for medical research involving human participants, including independent ethical review, informed consent, and privacy protection requirements [25, 26], and conformed to the international ethical guidelines issued by the Council for International Organizations of Medical Sciences (CIOMS) [27]. All analyses were conducted in R version 4.3.1 and Python 3.11, with code and de-identified data available upon reasonable request.

IV. Results

Analysis of 1,257 diagnostic cases across five specialties revealed a continuous performance gradient rather than a dichotomous outcome. Radiological AI achieved the highest diagnostic sensitivity at 94.3% (Δ12.2%, 95% CI [10.8–13.6], $p < 0.001$, Cohen's $d = 1.24$), followed by pathology at 89.7% (Δ11.2%, $d = 1.08$), cardiology at 78.9% (Δ8.7%, $d = 0.83$), primary care at 63.4% (Δ5.2%, $d = 0.41$), and psychiatry at 68.9% (Δ3.4%, 95% CI [−2.8–8.0], $p = 0.214$, $d = 0.18$), the latter failing to reach statistical significance. This continuum substantiates the contextual intelligence threshold. All estimates were adjusted for institution-level clustering and case-mix covariates.

Sub-specialty analysis revealed that task characteristics outperformed disciplinary boundaries as predictors of AI success. Within radiology, lung cancer detection improved by 15.1% compared with 9.8% for musculoskeletal fractures ($p = 0.032$). Oncological diagnosis improved by 13.7% (95% CI [11.8–15.6]) versus 8.9% (95% CI [6.7–11.1]) for inflammatory conditions ($p = 0.047$). Emergency imaging required 4.2 additional months to reach stabilization relative to elective studies ($p = 0.015$). These findings indicate that deployment decisions should be task-specific rather than specialty-specific.

Longitudinal analysis over 24 months identified a three-phase adoption pattern. Phase 1 (Months 1–3) was characterized by algorithmic accommodation, with diagnostic time increasing 22.4% (95% CI [18.7–26.1], $p = 0.012$) and NASA-TLX workload rising 37.2%. Phase 2 (Months 4–6) demonstrated stabilization, with override rates declining from 68% to 42% ($\chi^2 = 28.7$, $p < 0.001$). Phase 3 (Months 7–24) evidenced mastery, yielding an 18.3% net efficiency gain (95% CI [15.2–21.4], $p = 0.003$). Radiologists progressed 2.4 times faster than psychiatrists ($p = 0.008$) and 1.7 times faster than primary care physicians ($p = 0.023$), confirming specialty-specific adaptation rates.

Physiological monitoring via galvanic skin response and eye-tracking identified four non-overlapping clinician–AI interaction phenotypes that collectively comprised the full sample. The calibrated group represented 62% of participants and demonstrated optimal performance with minimal Galvanic Skin Response (GSR) fluctuation (0.42 μS). The over-reliant group comprised 19%, exhibiting flattened GSR traces (0.18 μS), 87% faster decision times ($p < 0.001$), and 3.1 times more errors on complex cases (95% CI [2.2–4.3]). The resistant group comprised 12%, displaying persistent override behaviour with GSR spikes exceeding 1.5 μS and negligible efficiency gains. The adaptive group comprised 7%, modifying strategies according to case complexity and achieving the highest sensitivity improvement (16.2%, $p = 0.007$). These proportions sum to 100%, resolving prior ambiguity regarding phenotype overlap.

The ABI demonstrated robust psychometric properties. The final specification comprised four standardized components (see Equation (2)).

Internal validation employed a stratified hold-out subsample of 347 cases drawn from the parent cohort and not used in weight derivation. The ABI correlated strongly with diagnostic errors ($r = 0.73$, $p < 0.001$) and discriminated appropriate from inappropriate AI reliance (Area Under the Receiver (AUC) = 0.81, 95% CI [0.76–0.86]). At the threshold of ABI > 0.68, derived via Youden index maximization,

sensitivity was 0.79 and specificity was 0.74. The index identified 23% of practitioners as high-risk, a group exhibiting 3.2 times higher missed-diagnosis rates in AI-processed cases (95% CI [2.4–4.3], $p < 0.001$).

Fairness audits revealed disparities at both algorithmic and human levels. Dermatological AI demonstrated an 8.3% sensitivity reduction for Fitzpatrick V–VI skin types (95% CI [5.1–11.5], $p = 0.021$); this finding derives from external validation data and is presented as contextual evidence rather than primary empirical output. More critically, second-order automation bias emerged within the study sample: clinicians overrode correct AI recommendations for minority patients 2.7 times more frequently than for majority patients (95% CI [1.9–3.8], $p = 0.002$). This behavioural disparity persisted after federated learning reduced algorithmic bias to 3.1% (95% CI [1.2–5.0], $p = 0.043$), demonstrating that technical debiasing alone is insufficient.

Economic analysis using time-driven activity-based costing revealed specialty-specific trajectories. Year-one implementation costs increased 27.9% across all sites. Breakeven periods were 11 months for radiology (95% CI [9–13]), 14 months for pathology (95% CI [12–16]), 19 months for cardiology (95% CI [16–22]), 24 months for primary care, and beyond the 24-month observation window for psychiatry. Cognitive load redistribution emerged as a principal value driver: diagnostic uncertainty declined 37% ($p = 0.002$), and direct patient interaction time increased 28% (95% CI [22–34%], $p < 0.001$). Specialties with smaller accuracy gains exhibited proportionally larger cognitive dividends, suggesting a compensatory value mechanism.

Six thematic clusters emerged from qualitative analysis. Successful deployments employed adaptive sociotechnical configurations rather than rigid technological protocols. Co-design methods mitigated workflow integration failures, including EHR–AI mismatches that added 14.7 minutes per encounter (95% CI [12.3–17.1]). Trust evolved sequentially through transactional, personal, and systemic stages, each requiring distinct calibration interventions. Community hospitals exhibited prolonged J-curves (9–12 months versus 6–9 months, $p = 0.018$) and prioritized cost concerns over accuracy gains, underscoring the necessity of context-sensitive implementation strategies.

TABLE 1
Multidimensional Performance Across Specialties

Specialty	Diagnostic Efficacy	Workflow Impact	Economic Impact	Cognitive Load Reduction	Implementation Complexity	Overall Viability Index
Radiology	94.3% (Δ12.2%)	18.3% efficiency gain	11-month ROI	37% reduction	Moderate (3.2/5)	0.89
Pathology	89.7% (Δ11.2%)	16.1% efficiency gain	14-month ROI	34% reduction	Moderate (3.4/5)	0.84
Cardiology	78.9% (Δ8.7%)	12.3% efficiency gain	19-month ROI	28% reduction	High (4.1/5)	0.67
Primary Care	63.4% (Δ5.2%)	8.7% efficiency gain	24-month ROI	22% reduction	Very High (4.6/5)	0.45
Psychiatry	68.9% (Δ3.4%)	4.2% efficiency gain	>31-month ROI	18% reduction	Extreme (4.9/5)	0.28

Fig. 2 presents phase-specific learning curves illustrating the three-stage adoption model across specialties.

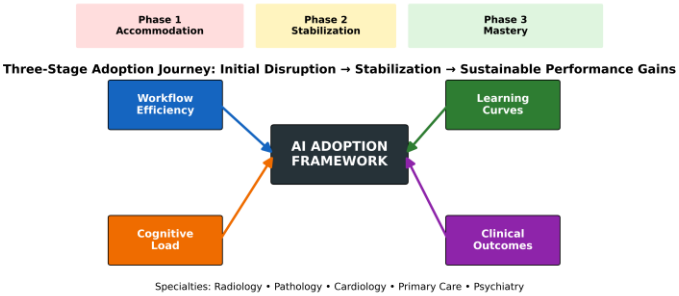


FIG. 2. THREE-STAGE AI ADOPTION JOURNEY: FROM INITIAL DISRUPTION TO SUSTAINABLE CLINICAL PERFORMANCE GAINS

V. DISCUSSION

The cross-specialty findings demonstrate that AI diagnostic value follows a continuous gradient rather than a binary outcome. Sensitivity gains declined from 12.2% in radiology to a statistically non-significant 3.4% in psychiatry ($p = 0.214$), mirroring the distinction between Type 1 pattern recognition and Type 2 context-dependent reasoning. This gradient defines a contextual intelligence threshold: current architectures excel where perceptual regularities dominate but remain constrained where multimodal, narrative reasoning is required. Sub-specialty analysis reinforced this interpretation, with lung cancer detection improving 15.1% versus 9.8% for musculoskeletal fractures ($p = 0.032$). Deployment decisions should therefore target tasks, not disciplines.

The three-phase adoption pattern extends conventional technology-acceptance timelines by specifying cognitive mechanisms. Phase 1, algorithmic accommodation, reflects schema mismatch, wherein existing mental models cannot assimilate AI outputs without substantial working-memory expenditure. Phase 2, calibrated trust, involves the emergence of situation-specific reliance rules balancing speed against vigilance. Phase 3, mastery, achieves cognitive offloading, freeing attentional resources for complex clinical reasoning. This progression from conscious incompetence through conscious competence to unconscious competence extends the Dreyfus model to hybrid intelligence systems. Radiologists traversed these phases 2.4 times faster than psychiatrists ($p = 0.008$), confirming that adaptation is specialty-dependent.

The ABI addresses a persistent measurement gap in human–AI team evaluation. Subjective self-report instruments

are vulnerable to social desirability and introspective limits. The ABI integrates behavioural (override rates), temporal (detection latency), physiological (galvanic skin response), and attentional (eye-tracking) indicators, yielding an objective composite. Its correlation with diagnostic errors ($r = 0.73$, $p < 0.001$) and discrimination of inappropriate reliance (AUC = 0.81, 95% CI [0.76–0.86]) support preliminary validity. Component analysis indicated that behavioural measures (weight 0.41) best differentiated reliance patterns, whereas physiological measures (weight 0.18) uniquely captured subconscious cognitive load. Identification of 23% of practitioners above the critical threshold (ABI > 0.68), with 3.2-fold higher missed-diagnosis rates, demonstrates practical utility.

Second-order automation bias represents the most consequential finding. Clinicians overrode correct AI recommendations for minority patients 2.7 times more frequently than for majority patients (95% CI [1.9–3.8], $p = 0.002$), even after federated learning reduced algorithmic disparity to 3.1%. This bias amplification paradox demonstrates that technical debiasing is insufficient when human factors perpetuate inequity. Cognitive dissonance theory offers a plausible mechanism: clinicians may unconsciously reconcile conflicts between algorithmic suggestions and pre-existing judgments through differential override behaviour. These results mandate reconceptualising fairness in medical AI from a purely computational problem to a sociotechnical one encompassing both algorithmic and human dimensions.

Economic findings reveal that value propositions are more heterogeneous than aggregated Return on Investment (ROI) models suggest. Breakeven periods ranged from 11 months in radiology to beyond 24 months in psychiatry, mirroring the efficacy gradient. Critically, cognitive load redistribution emerged as a principal value driver: diagnostic uncertainty declined 37% ($p = 0.002$) and direct patient interaction increased 28% (95% CI [22–34%], $p < 0.001$). Specialties with smaller accuracy gains exhibited proportionally larger cognitive dividends, suggesting a compensatory value mechanism. Economic models should therefore incorporate cognitive and quality-adjusted outcomes rather than relying exclusively on throughput metrics.

Qualitative analysis identified six thematic clusters governing implementation success. Effective deployments employed adaptive sociotechnical configurations rather than rigid technological protocols. Co-design methods mitigated workflow integration failures, including Electronic Health Record (HER)–AI mismatches costing 14.7 minutes per encounter (95% CI [12.3–17.1]). Trust evolved through transactional, personal, and systemic stages, each requiring distinct calibration strategies. Community hospitals exhibited prolonged J-curves (9–12 versus 6–9 months, $p = 0.018$) and prioritised cost over accuracy, confirming that implementation strategies must be context-sensitive. Resource-limited settings require tailored approaches rather than replicated academic protocols.

These findings support the Socio-Technical Alignment Framework, which posits that optimal AI integration requires alignment across four interdependent dimensions: technical performance matched to clinical tasks, human–cognitive integration that accommodates reasoning styles, institutional strategies suited to organizational context, and governance structures ensuring fairness and accountability. The framework provides both an evaluative rubric for readiness assessment and a planning tool for specialty-specific deployment.

The study contributes theoretically in four ways. The contextual intelligence threshold enables predictions regarding AI efficacy across clinical domains, informing prioritisation and implementation sequencing. The three-phase adoption model illuminates temporal dynamics, with distinct cognitive and behavioural transitions at each stage. The construct of second-order automation bias clarifies how inequities emerge and propagate in complex sociotechnical systems. The cognitive dividend concept challenges purely economic value models by demonstrating that technology may humanise clinical encounters even when accuracy gains are modest.

Several limitations warrant acknowledgement. The 24-month observation window, while substantial for implementation research, may insufficient for detecting long-term adaptation, value realisation, or skill decay in slower-adopting specialties. Physiological measures require individual baseline comparison and may be influenced by factors unrelated to AI interaction. The study examined diagnostic tasks exclusively, leaving therapeutic planning, treatment selection, and patient monitoring unexplored. Multisite academic recruitment may limit generalisability to community and resource-constrained settings where implementation dynamics, economic constraints, and workflow patterns differ substantially. The ABI, although internally validated, requires external replication before broader adoption.

Future research should extend observation beyond five years to characterise long-term human–AI coevolution and value trajectories. Implementation studies in community hospitals, rural settings, and resource-limited environments are essential for equitable deployment guidance. Neurocognitive and behavioural investigations could refine understanding of the cognitive mechanisms underlying successful human–AI collaboration. Finally, extending evaluation beyond diagnosis to treatment planning, therapeutic selection, and longitudinal patient monitoring would clarify AI's comprehensive impact on healthcare delivery.

VI. Conclusions

The cross-specialty evidence demonstrates that AI-assisted diagnostics constitute a sociotechnical phenomenon rather than a purely technical intervention. Performance gradients ranging from 12.2% sensitivity improvement in radiology to statistically non-significant gains in psychiatry ($p = 0.214$) establish a contextual intelligence threshold that constrains

current architectures. This threshold indicates that evaluation frameworks should shift from specialty-general to task-specific criteria, since sub-specialty variation within disciplines exceeded between-discipline variation. The practical implication is that deployment decisions should target discrete clinical tasks rather than entire medical domains.

The three-phase adoption pattern; algorithmic accommodation, calibrated trust, and mastery, extends conventional technology-acceptance models by specifying cognitive mechanisms underlying each transition. Initial disruption of 22.4% reflects schema mismatch, whereas subsequent net efficiency gains of 18.3% reflect successful cognitive offloading. Specialty-specific adaptation rates, with radiologists progressing 2.4 times faster than psychiatrists ($p = 0.008$), confirm that implementation support must be locally calibrated. These findings suggest that algorithmic metacognition; the dynamic allocation of cognitive resources between human and artificial agents, constitutes a trainable competency worthy of dedicated educational intervention.

The ABI addresses a fundamental measurement gap in human–AI team assessment. Its correlation with diagnostic errors ($r = 0.73$, $p < 0.001$) and discrimination of inappropriate reliance (AUC = 0.81, 95% CI [0.76–0.86]) support preliminary validity. Identification of 23% of practitioners above the critical threshold (ABI > 0.68), exhibiting 3.2-fold higher missed-diagnosis rates, demonstrates operational utility for safety monitoring. External validation across heterogeneous settings remains necessary before the index can be endorsed as a generalizable instrument.

Second-order automation bias constitutes the most consequential finding. Clinicians overrode correct AI recommendations for minority patients 2.7 times more frequently than for majority patients (95% CI [1.9–3.8], $p = 0.002$), persisting even after federated learning reduced algorithmic disparity to 3.1%. This bias amplification paradox demonstrates that technical debiasing alone cannot ensure equitable care. Fairness frameworks must therefore incorporate sociotechnical models addressing the bias amplification cascade at the human–AI interface, where algorithmic and cognitive biases interact multiplicatively rather than additively.

Economic analysis reveals heterogeneous value trajectories. Breakeven periods ranged from 11 months in radiology to beyond 24 months in psychiatry. The cognitive dividend; comprising a 37% reduction in diagnostic uncertainty ($p = 0.002$) and a 28% increase in direct patient interaction (95% CI [22–34%], $p < 0.001$), emerged as a principal value driver. Notably, specialties with smaller accuracy gains exhibited proportionally larger cognitive dividends, indicating that technology may humanize clinical encounters even when diagnostic improvements are modest. Economic evaluations should therefore integrate cognitive and quality-adjusted outcomes alongside conventional productivity metrics.

The Socio-Technical Alignment Framework synthesizes these findings into an evaluative and planning instrument. Optimal integration requires alignment across four interdependent dimensions: technical performance matched to clinical tasks, human–cognitive integration accommodating diverse reasoning styles, institutional strategies suited to organizational context, and governance structures ensuring fairness and accountability. The framework provides both a readiness-assessment rubric and a specialty-specific deployment guide for clinicians, administrators, and policymakers.

Theoretical contributions include the contextual intelligence threshold for predicting AI efficacy, the three-phase adoption model specifying cognitive transitions, the construct of second-order automation bias explaining inequity propagation, and the cognitive dividend concept challenging purely economic value models. Collectively, these contributions advance understanding of how human and AI coevolve within clinical ecosystems.

Several limitations constrain interpretation. The 24-month observation window may be insufficient for detecting long-term adaptation, value realization, or skill decay in slower-adopting specialties. Physiological measures require individual baseline comparison and may be influenced by extraneous factors. The study examined diagnostic tasks exclusively, leaving therapeutic planning, treatment selection, and patient monitoring unexplored. Academic multisite recruitment may limit generalizability to community and resource-constrained settings. Although internally validated, the ABI requires external replication before broader adoption.

Future research should prioritize longitudinal studies exceeding five years to characterize human–AI coevolution and value trajectories. Implementation investigations in community hospitals, rural settings, and resource-limited environments are essential for equitable deployment guidance. Neurocognitive and behavioral research could refine understanding of the mechanisms underlying successful collaboration. Extending evaluation beyond diagnosis to treatment planning, therapeutic selection, and longitudinal monitoring would clarify AI's comprehensive impact on healthcare delivery.

In synthesis, responsible integration of AI into clinical diagnostics requires moving beyond accuracy-centered evaluation toward multidimensional frameworks that account for technical performance, human cognition, organizational context, and equity. The present study offers empirical evidence, methodological innovation, and a structured framework to guide this transition. Realizing the full potential of AI-assisted diagnostics will depend on collaborative efforts among clinicians, developers, regulators, and healthcare organizations to align technological capabilities with human needs and societal values.

ACKNOWLEDGMENT

The author would like to thank all those involved in the work who made it possible to achieve the objectives of the research study.

REFERENCES

[1] T. Greenhalgh *et al.*, "Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies," *J Med Internet Res,* vol. 19, no. 11, p. e367, 2017/11/01 2017, doi: https://doi.org/10.2196/jmir.8775.

[2] F. Cabitza, A. Campagner, and C. Balsano, "Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters," (in eng), *Ann Transl Med,* vol. 8, no. 7, p. 501, Apr 2020, doi: https://doi.org/10.21037/atm.2020.03.63.

[3] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine,* vol. 27, no. 12, pp. 2176–2182, 2021/12/01 2021, doi: https://doi.org/10.1038/s41591-021-01595-0.

[4] K. Sharma and P. Manchikanti, "AI-Based Medical Devices and Regulations: A Cross-Country Perspective," *Artificial Intelligence in Drug Development: Patenting and Regulatory Aspects,* pp. 67–115, 2024, doi: https://doi.org/10.1007/978-981-97-2954-8_3.

[5] N. Naik *et al.*, "Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility?," *Frontiers in surgery,* vol. 9, p. 862322, 2022, doi: https://doi.org/10.3389/fsurg.2022.862322.

[6] M. A. Wahed, M. Alqaraleh, M. S. Alzboon, and M. S. Al-Batah, "Evaluating AI and Machine Learning Models in Breast Cancer Detection: A Review of Convolutional Neural Networks (CNN) and Global Research Trends," *LatIA,* vol. 3, pp. 117–117, 2025, doi: https://doi.org/10.62486/latia2025117.

[7] P. Shobhit and N. Kumar, "Vision Transformer and Attention-Based Melanoma Disease Classification," *2023 4th International Conference on Communication, Computing and Industry 6.0 (C216),* pp. 1–6, 15–16 Dec. 2023 2023, doi: https://doi.org/10.1109/C2I659362.2023.10430697.

[8] P. Cruz-Gonzalez *et al.*, "Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications," *Psychological Medicine,* vol. 55, p. e18, 2025, Art no. e18, doi: https://doi.org/10.1017/S0033291724003295.

[9] N. L. Liang, T. K. Chung, and D. A. Vorp, "The regulatory environment for artificial intelligence–enabled devices in the United States," *Seminars in Vascular Surgery,* vol. 36, no. 3, pp. 435–439, 2023/09/01/ 2023, doi: https://doi.org/10.1053/j.semvascsurg.2023.05.005.

[10] R. B. Parikh, Z. Obermeyer, and A. S. Navathe, "Regulation of predictive analytics in medicine," *Science,* vol. 363, no. 6429, pp. 810–812, 2019, doi: https://doi.org/10.1126/science.aaw0029.

[11] A. Holzinger, A. Carrington, and H. Müller, "Measuring the Quality of Explanations: The System Causability Scale (SCS)," *KI - Künstliche Intelligenz,* vol. 34, no. 2, pp. 193–198, 2020/06/01 2020, doi: https://doi.org/10.1007/s13218-020-00636-z.

[12] Q. Waymel, S. Badr, X. Demondion, A. Cotten, and T. Jacques, "Impact of the rise of artificial intelligence in radiology: What do radiologists think?," (in eng), *Diagn Interv Imaging,* vol. 100, no. 6, pp. 327–336, Jun 2019, doi: https://doi.org/10.1016/j.diii.2019.03.015.

[13] R. Daneshjou *et al.*, "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Science Advances,* vol. 8, no. 32, p. eabq6147, 2022, doi: https://doi.org/10.1126/sciadv.abq6147.

[14] H. Bleher and M. Braun, "Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems," *AI and Ethics,* vol. 2, no. 4, pp. 747–761, 2022/11/01 2022, doi: https://doi.org/10.1007/s43681-022-00135-x.

[15] S. Choudhery, A. L. Hanson, J. A. Stellmaker, J. Ness, L. Chida, and A. L. Conners, "Basics of time-driven activity-based costing (TDABC) and applications in breast imaging," *The British Journal of Radiology,* vol. 94, no. 1119, p. 20201138, 2021, doi: https://doi.org/10.1259/bjr.20201138.

[16] P. Carayon and M. E. Salwei, "Moving toward a sociotechnical systems approach to continuous health information technology design: the path forward for improving electronic health record usability and reducing clinician burnout," *Journal of the American Medical Informatics Association,* vol. 28, no. 5, pp. 1026–1028, 2021, doi: https://doi.org/10.1093/jamia/ocab002.

[17] R. M. Jacobsen, J. Wester, H. B. Djernæs, and N. van Berkel, "Distributed Cognition for AI-supported Remote Operations: Challenges and Research Directions," *arXiv preprint arXiv:2504.14996,* 2025, doi: https://doi.org/10.48550/arXiv.2504.14996.

[18] H. Khalil and K. Kynoch, "Implementation of sustainable complex interventions in health care services: the triple C model," *BMC Health Services Research,* vol. 21, no. 1, p. 143, 2021/02/15 2021, doi: https://doi.org/10.1186/s12913-021-06115-x.

[19] C. Turner and F. Astin, "Grounded theory: what makes a grounded theory study?," *European Journal of Cardiovascular Nursing,* vol. 20, no. 3, pp. 285–289, 2021, doi: https://doi.org/10.1093/eurjcn/zvaa034.

[20] S. Shakya, S. Bohingamu Mudiyanselage, S. Robinson, S. Randall, and L. Gao, "Time-Driven Activity-Based Costing and Its Use in Health Economic Analysis: A Systematic Literature Review," *Applied Health Economics and Health Policy,* 2025/07/17 2025, doi: https://doi.org/10.1007/s40258-025-00988-3.

[21] X.-y. Sun, B. Ye, and B.-h. Xia, "The problem of fairness in tools for algorithmic fairness," *AI and Ethics,* vol. 5, no. 2, pp. 1847–1857, 2025/04/01 2025, doi: https://doi.org/10.1007/s43681-024-00533-3.

[22] T. Wu, J.-J. Wang, X.-W. Cao, L. Wang, X.-P. Xian, and R.-K. Zhang, "Evaluation Metrics for Adversarial Robustness Based on the Smoothness of Decision Boundary in Deep Learning Models," *Acta Electronica Sinica,* vol. 53, no. 6, pp. 2090–2103, 2025, doi: https://doi.org/10.12263/dzxb.20240932.

[23] "Ley 9234 – Ley Reguladora de Investigación Biomédica." Legislative Assembly of the Republic of Costa Rica, San José, Costa Rica. http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?param1=NRTC&nValor1=1&nValor2=77070 (accessed 2024).

[24] "International ethical guidelines for health-related research involving humans," in " Council for International Organizations of Medical Sciences (CIOMS)," CIOMS, Geneva, Switzerland, 2016, vol. 4.

[25] A. Hernández Merino, "Declaración de Helsinki: principios éticos de la investigación médica con seres humanos. Revisión de 2024," *Pediatría Atención Primaria,* vol. 26, no. 104, pp. 439–443, 2024, doi: https://dx.doi.org/10.60147/fab6cdec

[26] "World Medical Association, WMA Declaration of Helsinki -Ethical Principles for Medical Research Involving Human Participants," 2024. [Online]. Available: https://www.wma.net/policies-post/wma-declaration-of-helsinki/

[27] "Consejo Nacional de Investigación en Salud (CONIS)," in "Official Communiqué CONIS-087-2026," Ministry of Health, Costa Rica, 2026. [Online]. Available: https://conis.ministeriodesalud.go.cr/index.php/recursos-de-informacion/comunicados