L A B O R A T O R I O S  A L E X A N D R I A

**D E M O  E D I T I O N  —  S E L E C T E D  S E C T I O N S**

*Full analysis available upon request*

# The Epistemic Gap in Clinical AI

*Why Medical AI Performs in Papers but Fails in Hospitals:*
*Structural Barriers Between Computational Validity and Clinical Impact*

**AIB-2026-007**
June 2026
Life Sciences × Epistemology

**ALETHEIA GRADE A**
Confidence: 0.50 | Epistemic Distance: 1.0 | Surprise: 1.0

# 1. Executive Summary

This analysis investigates a structural disconnect identified by the Alexandria epistemic engine at the intersection of Life Sciences and Epistemology: the persistent failure of artificial intelligence systems to translate high laboratory performance metrics into measurable clinical impact. The finding was surfaced through cross-domain correlation analysis of 273,000+ scientific records and subjected to adversarial deliberation through the ALETHEIA protocol.

The anchor case is a landmark paper by Christopher Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg S. Corrado, and Dominic King (Google Health / DeepMind), which identifies the key challenges for delivering clinical impact with artificial intelligence. This paper is notable not for proposing a new AI architecture or reporting a new benchmark, but for diagnosing a systemic failure: AI systems that achieve state-of-the-art performance on retrospective datasets routinely fail to produce measurable benefits when deployed in clinical settings. The authors' analysis is significant precisely because they come from within the organisations that build these systems—the diagnosis is not external critique but internal recognition of a structural problem.

The central finding of this brief is this: the gap between computational performance and clinical impact is not a deployment problem, an engineering problem, or a data quality problem. It is an **epistemic** problem. The metrics by which AI systems are evaluated (AUC, sensitivity, specificity, F1 score) measure the system's relationship to a labelled dataset. They do not measure—and cannot measure—the system's relationship to the clinical reality those labels were meant to represent. When a model achieves 0.95 AUC on a retrospective chest X-ray dataset, this number describes the model's ability to replicate the patterns in radiologist annotations. It does not describe the model's ability to detect disease, because the annotations themselves are an imperfect and context-dependent representation of disease. The metric measures alignment with a proxy, not with the target.

The ALETHEIA deliberation, involving four rounds of adversarial analysis by the SINESIE and EPISTEME models, produced a Grade A conclusion with an unusually low confidence of 0.50. This asymmetry is itself informative and warrants explanation. Grade A indicates that the structural finding is sound and generalisable: the epistemic gap between computational metrics and clinical validity is real, identifiable, and recurrent across medical AI applications. Confidence 0.50 reflects the partial state of formal evidence: while the barriers are well-described qualitatively, their complete formalisation—as rigorous epistemic invariants with quantitative boundary conditions—remains an open research programme. The reader should interpret this as: the diagnosis is correct, but the prescription is incomplete.

**Epistemic Note:** This brief employs Alexandria's standard transparency protocol. All claims sourced from the vault are marked VAULT-VERIFIED with document identifiers. Claims generated during SINESIE deliberation that cite specific numerical values, references, or experimental data not independently verified against the vault are marked as "adversarial claims raised during deliberation—not independently verified." The surprise score of 1.0 (the maximum in the entire Alexandria corpus) indicates that this cross-domain connection—Life Sciences and Epistemology—is essentially absent from existing interdisciplinary literature, despite being, as this brief argues, the most consequential blind spot in clinical AI development.

## 2. Cross-Domain Convergence Map

The convergence identified in this brief operates across two domains whose separation is maximal within the Alexandria Universal Taxonomy: Life Sciences (the empirical study of biological systems, disease, and clinical intervention) and Epistemic Foundations (the study of what constitutes valid knowledge, justification, and evidence). Epistemic distance: 1.0. Surprise: 1.0. These are not incremental values—they represent the maximum possible separation and the maximum possible novelty in the corpus. The Alexandria engine has never encountered this specific crossing in 360,000+ records. This section maps why.

### 2.1 Life Sciences: The Performance-Impact Paradox

The history of medical AI is, in significant part, a history of failed translation. The literature is replete with systems that demonstrate exceptional performance on benchmark tasks—detecting diabetic retinopathy from fundus images, identifying malignant lesions on dermoscopy, predicting sepsis onset from vital sign patterns—and that subsequently fail to produce measurable improvements in patient outcomes when deployed in clinical environments. The anchor paper by Kelly et al. provides the most authoritative analysis of why this occurs, identifying barriers that are not technical but systemic.

The first barrier is **distributional shift**. A model trained on images from one hospital's scanner, annotated by that hospital's radiologists, encodes not only disease patterns but also the specific characteristics of that scanner's imaging chain and that institution's diagnostic culture. When deployed at a different hospital with different equipment and different annotation norms, the model's performance degrades—not because the diseases are different, but because the *representation* of disease is different. The model has learned a proxy (institutional imaging patterns) rather than the target (disease biology).

The second barrier is **label validity**. Every supervised learning system requires labelled training data. In medical AI, labels typically derive from clinical diagnoses, which are themselves probabilistic judgments made by clinicians under conditions of uncertainty, time pressure, and incomplete information. A chest X-ray labelled "pneumonia" does not contain an objective ground truth—it contains a clinician's judgment, which may have been influenced by the patient's history, the time of day, the clinician's experience, and institutional diagnostic conventions. The label is not a fact about the image; it is a fact about the clinical encounter that produced the label.

The third barrier is **integration into clinical workflows**. A model that correctly identifies a condition is clinically useless if its output does not arrive at the right point in the clinical decision-making process, in a format the clinician can act upon, with sufficient contextual information to justify changing the planned course of action. Clinical decision-making is not a classification task—it is a process of reasoning under uncertainty that integrates imaging data, laboratory results, patient history, clinical experience, and institutional protocols. An AI system that provides a probability score without engaging this broader epistemic framework does not supplement clinical reasoning; it interrupts it.

The fourth barrier, less discussed but perhaps most fundamental, is **the feedback gap**. In most clinical AI deployments, the model does not receive outcome feedback—it never learns whether

its recommendations led to better or worse patient outcomes. Without this feedback loop, the model cannot self-correct, and its operators cannot distinguish between a well-calibrated system and one that is confidently wrong. The absence of outcome feedback is not an engineering oversight; it reflects the deep difficulty of attributing patient outcomes to specific diagnostic decisions in complex clinical pathways.

### 2.2 Epistemic Foundations: The Validity of Computational Evidence

*This section analyses the epistemic frameworks (proxy problem, calibration gap, boundary problem, self-evaluation paradox) that structurally correspond to the clinical AI barriers identified in §2.1. Full analysis available in the complete version.*

*Request access: intelligence@laboratoriosalexandria.com*

### 2.3 Structural Correspondence

*This section presents the formal structural isomorphism between clinical AI barriers and fundamental epistemic constraints, including immediate practical consequences for the solution space. Full analysis available in the complete version.*

*Request access: intelligence@laboratoriosalexandria.com*

# 3. Epistemic Confidence Assessment

## 3.1 ALETHEIA Deliberation Metrics

| Metric | Value | Threshold | Assessment |
|---|---|---|---|
| Conclusion Grade | A | ≥ B | Exceeds threshold. Structural finding is sound and generalisable. |
| Confidence Level | 0.50 | ≥ 0.70 | Below threshold. Formal evidence is partial. See Section 3.3. |
| Epistemic Distance | 1.0 | ≥ 0.5 | Maximum distance. Domains never cross in existing literature. |
| Surprise Index | 1.0 | ≥ 0.4 | Maximum surprise. Highest rarity score in entire corpus. |
| Recurrence | 2 | ≥ 2 | Meets threshold. Pattern detected across 2 independent correlation passes. |
| Deliberation Rounds | 4 | ≥ 3 | Full adversarial cycle: SINESIE × 4 + EPISTEME arbiter. |
| Maturity | NEAR MATURE | N/A | Thesis formulated and graded. Awaiting external validation pathway. |

## 3.2 Interpreting Grade A / Confidence 0.50

This brief presents an unusual metric profile that requires explicit interpretation. Grade A with confidence 0.50 is not a contradiction—it reflects two distinct dimensions of epistemic assessment.

Grade A assesses the *structural soundness* of the finding: is the claimed pattern real, and does it generalise beyond the specific anchor case? The ALETHEIA deliberation concluded that yes—the epistemic gap between computational metrics and clinical validity is a recurring, domain-independent structure that manifests wherever proxy-based evaluation is mistaken for target-based validation. This pattern is observable across medical imaging, drug discovery, genomic prediction, and clinical decision support. It is not an artefact of the specific paper analysed.

Confidence 0.50 assesses the *evidential completeness* of the finding: how fully has the pattern been formalised, quantified, and empirically tested? Here the answer is: partially. The qualitative barriers are well-documented by Kelly et al. and extensively discussed in the medical AI literature. But their formalisation as rigorous epistemic invariants—with quantitative boundary conditions, measurable parameters, and falsification criteria—remains incomplete. The literature describes *what* goes wrong but has not yet produced a formal framework for *predicting when* it will go wrong or *quantifying how much* validity is lost.

**The honest interpretation:** we know the disease but not the dose-response curve. The diagnosis is Grade A; the pharmacokinetics of the remedy are at confidence 0.50.

## 3.3 Sources of Uncertainty

*This section details four sources of uncertainty: formalisation gap, selection bias in the anchor paper, deliberation bias, and temporal uncertainty. Full analysis available in the complete version.*

*Request access: intelligence@laboratoriosalexandria.com*

## 3.4 Falsifiability Conditions

*This section presents four falsifiability conditions (FC-1 through FC-4) that would invalidate the central claims of this brief. Full analysis available in the complete version.*

*Request access: intelligence@laboratoriosalexandria.com*

## 4. Structural Correspondence Table

The following table maps the structural parallels identified between clinical AI barriers and epistemic frameworks. The “Verification” column distinguishes between claims grounded in the vault corpus and claims generated during SINESIE adversarial deliberation.

| Life Sciences (Domain A) | Epistemology (Domain B) | Structural Pattern | Verification |
|---|---|---|---|
| Model trained on Hospital A’s data degrades at Hospital B (distributional shift) | Formal system’s outputs lose validity outside its boundary conditions | Context-dependence of validity. Epistemic claims are bound to the conditions under which they were generated. | Vault-verified (doc_480f2cd51c0547c7) + established ML theory |
| Clinical labels encode clinician judgment, not objective ground truth | Proxy evidence: measuring correlation with phenomenon, not phenomenon itself | The proxy problem. Evaluation against proxies produces proxy validity, not target validity. | Vault-verified (anchor paper analysis of label provenance) |
| High-AUC model fails to improve patient outcomes when deployed | Calibration gap: stated confidence does not match actual reliability | Performance metrics measure internal consistency, not external correspondence. | Vault-verified (performance-impact gap documented in anchor paper) |
| Model validated against same labellers who produced training data | Self-referential evaluation: system cannot detect its own blind spots | The self-evaluation paradox. Circular validation systematically overestimates validity. | *SINESIE synthesis (structural connection to AIB-2026-009)* |
| No outcome feedback: model never learns if recommendations helped patients | Open-loop knowledge: assertions without correction mechanism | Knowledge without accountability. Claims that cannot be corrected cannot be validated. | Vault-verified (feedback gap identified in anchor paper) |
| RLHF-trained models collapse clinical judgment into scalar reward | Category error: multidimensional value reduced to single ordering | The dimensional collapse. Complex epistemic structures are destroyed by scalar aggregation. | *SINESIE synthesis (structural connection to AIB-2026-011)* |

## 5. Adversarial Findings

The ALETHEIA deliberation protocol subjected the convergence thesis to four rounds of adversarial scrutiny. The challenges raised are not weaknesses to be minimised but integral components of the epistemic assessment.

### 5.1 Challenge: Is This Simply a Known Problem?

SINESIE Round 1 raised the most immediate objection: the gap between AI performance and clinical impact is widely discussed in the medical AI literature.

> *Full adversarial analysis and resolution for this challenge available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

### 5.2 Challenge: Confidence Calibration of the Confidence Claim

EPISTEME raised a meta-level concern: this brief argues that poorly calibrated confidence estimates are a core problem in clinical AI.

> *Full adversarial analysis and resolution for this challenge available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

### 5.3 Challenge: Anchor Paper Authors' Institutional Position

SINESIE Round 2 noted that the anchor paper's authors are or were affiliated with Google Health and DeepMind—organisations that are among the world's largest developers of medical AI systems.

> *Full adversarial analysis and resolution for this challenge available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

### 5.4 Challenge: Fabricated Specificity in Deliberation

During the SINESIE deliberation, participants cited specific numerical values, journal references with precise DOIs, and statistical claims attributed to named studies.

> *Full adversarial analysis and resolution for this challenge available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

### 5.5 Challenge: Generative Loop Artifact

A technical observation: during the fourth round of deliberation, SINESIE-2 entered a generative loop, producing the phrase “This correlation raises five fundamental questions” repeatedly without termination.

> *Full adversarial analysis and resolution for this challenge available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

## 6. Research Hypotheses

The following hypotheses emerge from the convergence analysis and are designed to be independently testable.

### RH-007-01: Proxy-Target Divergence in Clinical AI Validation

**Hypothesis:** There exists a quantifiable divergence between proxy-based performance metrics (AUC on retrospective datasets) and target-based clinical impact measures (patient outcome improvements) for medical AI systems, and this divergence increases as a function of the epistemic distance between the proxy and the target.

> *Experimental design and falsification criteria available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

### RH-007-02: Calibration as Epistemic Validity Indicator

**Hypothesis:** Clinical AI systems with better-calibrated confidence estimates produce larger improvements in patient outcomes than systems with equal discriminative performance (AUC) but poorer calibration, controlling for domain, deployment context, and model architecture.

> *Experimental design and falsification criteria available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

### RH-007-03: Validity Boundary Specification

**Hypothesis:** Clinical AI systems deployed with explicit, documented validity boundaries produce fewer adverse events and higher clinician trust than systems deployed without such boundaries, even when the underlying models are identical.

> *Experimental design and falsification criteria available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

### RH-007-04: Outcome Feedback Loop Efficacy

**Hypothesis:** Clinical AI systems that receive structured outcome feedback achieve measurably higher clinical validity over time than systems operating in open-loop mode, and this improvement cannot be replicated by retraining on larger retrospective datasets alone.

> *Experimental design and falsification criteria available in the complete version.*
> *Request access: intelligence@laboratoriosalexandria.com*

## 7. Frontier Questions

The following questions emerged from the deliberation and represent open directions for which no current answer exists within the Alexandria corpus.

### FQ-01: What Constitutes Sufficient Epistemic Specification for a Clinical AI System?

Every deployed clinical AI system makes implicit epistemic commitments: assumptions about the relationship between its training data and clinical reality, about the validity of its labels, about the generalisability of its performance metrics. Currently, these commitments are undocumented—the model is deployed as a "black box" whose epistemic boundaries are unknown to its operators. The frontier question is whether a practical framework can be developed for *epistemic specification*—a document that, analogous to a drug's prescribing information, enumerates the conditions under which the model's outputs constitute justified beliefs and the conditions under which they do not. Such a framework would transform clinical AI deployment from a statistical exercise into an epistemically grounded practice. The development requires collaboration between computer scientists, epistemologists, clinicians, and regulatory bodies—precisely the kind of cross-domain integration that the Alexandria engine was designed to identify.

### FQ-02: Is There a General Theory of Proxy-to-Target Translation in Medical Evidence?

The proxy problem identified in this brief—evaluation against proxies produces proxy validity, not target validity—is not unique to AI. It appears throughout evidence-based medicine: surrogate endpoints in clinical trials (does lower blood pressure mean fewer strokes?), biomarker-based diagnosis (does PSA level indicate prostate cancer?), and quality metrics in healthcare delivery (does lower readmission rate mean better care?). In each case, a measurable proxy is used to represent an unmeasurable or difficult-to-measure target, and the fidelity of this representation is uncertain. A general theory of proxy-to-target translation would formalise the conditions under which proxy evidence is epistemically justified, the quantitative relationship between proxy fidelity and evidential weight, and the point at which proxy-based reasoning becomes epistemically unjustified. Such a theory does not currently exist in any single discipline—it lives at the intersection of epistemology, statistics, and medical methodology.

### FQ-03: Can Epistemic Validity Be Made Measurable?

> *This frontier question explores the development of a quantitative "epistemic health score" for deployed AI systems. Full analysis available in the complete version.*
>
> *Request access: intelligence@laboratoriosalexandria.com*

### FQ-04: What Happens to Clinical Judgment When AI Mediates the Evidence?

> *This frontier question examines how AI mediation of clinical evidence introduces new forms of epistemic distance. Full analysis available in the complete version.*
>
> *Request access: intelligence@laboratoriosalexandria.com*

## FQ-05: Do Epistemic Limits in Clinical AI Propagate Into Regulatory Frameworks?

*This frontier question addresses whether regulatory frameworks for medical AI (FDA, EU AI Act) are built on epistemically insufficient foundations. Full analysis available in the complete version.*

*Request access: intelligence@laboratoriosalexandria.com*

## 8. Strategic Implications

The findings in this brief have direct operational implications for organisations developing, deploying, purchasing, or regulating medical AI systems.

### 8.1 Near-Term (6–12 months)

**For pharma and medtech R&D directors:** Re-evaluate the internal evaluation framework for AI-assisted diagnostic and therapeutic tools. If the primary success metric is retrospective AUC, the evaluation is measuring proxy alignment, not clinical validity. Supplement AUC with calibration analysis (expected calibration error), distributional robustness testing (performance under systematic covariate shift), and explicit validity boundary documentation. These additions do not require new infrastructure—they require a change in what is measured and what is reported.

**For healthtech CTOs:** Begin implementing epistemic metadata in AI system outputs. When a model provides a clinical recommendation, the output should include not only the confidence score but also a validity boundary statement: the conditions under which this recommendation is epistemically justified. This is analogous to the shift from point estimates to confidence intervals in statistics—it provides the clinician with the information needed to exercise appropriate epistemic judgment. The parallel to AIB-2026-009 is direct: computational systems that evaluate their own validity produce contradictions unless they include explicit boundary conditions in their outputs.

**For clinical AI researchers:** Publish epistemic audits alongside performance reports. For each model, document: (1) the proxy-target relationship (what does the evaluation metric actually measure, and how does this relate to the clinical outcome of interest?), (2) the label provenance (who produced the labels, under what conditions, and with what inter-annotator agreement?), and (3) the validity boundaries (for which patient populations, imaging systems, and clinical contexts are the model's outputs epistemically justified?). These audits, which currently do not exist as a standard publication format, would transform the medical AI literature from a collection of performance claims into a body of epistemically grounded evidence.

### 8.2 Medium-Term (12–24 months)

*This section provides strategic recommendations for regulators (EU AI Act integration), hospital systems (outcome feedback loops), and health economics (proxy-target fidelity as investment parameter). Full analysis available in the complete version.*

*Request access: intelligence@laboratoriosalexandria.com*

### 8.3 Long-Term (24+ months)

*This section analyses the long-term epistemic maturation trajectory for medical AI and presents the thematic triangle (AIB-007 + 009 + 011) constituting the Life Sciences × Epistemology Cluster (LS-E). Full analysis available in the complete version.*

*Request access: intelligence@laboratoriosalexandria.com*

## 9. Source Provenance

### PRIMARY SOURCE (VAULT-VERIFIED)

**Christopher Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg S. Corrado, Dominic King.** “Key challenges for delivering clinical impact with artificial intelligence.” Google Health / DeepMind.

`Vault ID:` `doc_480f2cd51c0547c7`

`Correlation ID:` `2c59b14542c01a49`

`Cross-domain log:` `ID 479`

`Thesis:` `THESIS-20260610-043 | Session FORO-20260610-b53646`

`Detection pathway:` `EUREKA → cross_domain_log → Foro Epistémico → ALETHEIA deliberation`

### CLAIMS FROM DELIBERATION (SINESIE-GENERATED, PENDING VERIFICATION)

*Five SINESIE-generated claims documented for methodological transparency, including fabricated references, unverified numerical claims, and generative loop artifacts. Full provenance audit available in the complete version.*

*Request access: intelligence@laboratoriosalexandria.com*

## 10. Related Analyses — Available Upon Request

The Alexandria intelligence engine continuously surfaces cross-domain connections across 360,000+ scientific records. The following analyses are related to the themes explored in this brief and are available as commissioned intelligence products.

**AIB-2026-009** | The Self-Evaluation Paradox in Computational Systems | *Computer Science × Epistemology*

*When a computational system evaluates its own validity, it produces structural contradictions. The clinical manifestation: AI models validated against labels produced by the same epistemic framework systematically overestimate their own accuracy.*

**AIB-2026-011** | The RLHF Category Error | *AI/ML × Ethics*

*RLHF commits a structural category error by collapsing multidimensional values into a scalar reward signal. The clinical consequence: when medical AI is trained with human feedback, the clinician's rich epistemic framework—experience, diagnostic intuition, patient context—is destroyed in the reward signal.*

**AIB-2026-008** | Epistemic Limits in Materials Characterization | *Materials Science × Epistemology*

*The limits of characterisation in materials science have a direct analogue in medicine: diagnostic instruments alter what they measure. The observer effect in clinical testing—patient reactivity to procedures, confirmation bias in imaging, iatrogenic effects of monitoring—mirrors the characterisation paradox at human scale.*

**COMMISSION THIS ANALYSIS**

intelligence@laboratoriosalexandria.com



**FULL VERSION AVAILABLE**

This demo edition contains selected sections of AIB-2026-007.
The full analysis includes complete adversarial findings, experimental designs, falsifiability conditions, frontier questions, and source provenance audit.

**intelligence@laboratoriosalexandria.com**