CURELY AI · RESEARCH
Methodology & Evidence Review · Clinical AI Evaluation Science

When Accuracy Does Not Transfer: A Deployment-Grounded Evaluation Framework for Clinical Artificial Intelligence in Resource-Variable Health Systems

Curely AI Research
Clinical AI Evaluation Group, Curely AI
Correspondence: research@curelyai.com · curelyai.com/research

Article type & disclosure. *This is a methodology paper and structured evidence review. It sets out the evaluation standard Curely AI commits to for its own clinical systems and synthesises external, peer-reviewed evidence on why clinical model performance fails to transfer between settings. It is not a report of a completed Curely trial. Statements describing Curely's internal protocol are labelled as commitments rather than as validated outcome data. The authors declare a competing interest as employees of Curely AI; the framework is presented for scientific scrutiny and is product-neutral in its claims.*

Abstract

Background. Clinical artificial intelligence (AI) models increasingly inform patient care, yet performance measured during development frequently fails to transfer to the settings that adopt them, and the failure is often silent. This risk is greatest for under-resourced health systems, which rarely hold training-representative data or the infrastructure to detect degradation after deployment. **Objective.** To characterise the mechanisms and magnitude of the performance-transfer gap in clinical AI and to specify a reproducible, deployment-grounded evaluation protocol that reduces the risk of undetected failure. **Methods.** We conducted a structured synthesis of peer-reviewed external-validation and human-centred field evidence, graded each source with an explicit evidence-tier scheme, and derived a five-component evaluation framework aligned with consensus AI reporting standards (TRIPOD+AI, DECIDE-AI, SPIRIT-AI, and CONSORT-AI). **Results.** External validation of a widely deployed proprietary sepsis model showed the area under the receiver-operating-characteristic curve (AUC) fall from a developer-reported 0.76–0.83 to 0.63, with 33% sensitivity, a 12% positive predictive value, and roughly two-thirds of septic patients missed at the operational alerting threshold. Field studies of diabetic-retinopathy screening demonstrated that laboratory-calibrated input thresholds and socio-environmental workflow factors, rather than model accuracy alone, governed real-world performance. Three mechanisms — population shift, data-and-workflow shift, and the human-factors gap — account for most observed degradation. **Conclusion.** A reported accuracy figure is a property of a model in a specific setting, not of the model in general. Local pre-deployment validation, sample-size-honest subgroup reporting, silent-mode monitoring, and continuous drift surveillance should be treated as core requirements rather than optional additions. The framework is presented as an evaluation commitment and evidence review, not as outcome data from a completed deployment.

Keywords: *clinical decision support; external validation; distribution shift; model calibration; algorithmic fairness; post-deployment monitoring; healthcare AI governance; low-resource settings.*

1 Introduction

1.1 Background

Machine-learning systems are now embedded in routine clinical infrastructure. They triage emergency presentations, flag deteriorating inpatients, read radiographs and retinal photographs, and surface risk scores inside the electronic health record (EHR) at the point of care. The scientific literature that accompanies these tools typically reports a single headline figure of merit — an accuracy, a sensitivity, or an area under the curve — obtained on a held-out portion of the data used to build the model. That figure is frequently treated, by purchasers and clinicians alike, as an intrinsic property of the model: a number the tool carries with it wherever it is installed.

This assumption is unsafe. A model learns the statistical structure of the population, instruments, and workflows present in its training data. When any of those change — as they invariably do when a model crosses an institutional, geographic, or temporal boundary — the reported figure may no longer describe how the model behaves in the new setting. The gap between reported and realised performance is the central problem this paper addresses.

1.2 The problem of silent failure

A degraded clinical model rarely announces itself. Unlike a mechanical device that stops working, a miscalibrated risk model continues to emit plausible-looking outputs. It may under-alert on the events it was meant to catch while over-alerting elsewhere, fatiguing the clinicians who depend on it. Absent an explicit local check, the degradation is invisible until an independent group performs an external validation — often long after the tool has been deployed at scale. The health systems least able to run that check are precisely those for which the transfer gap is widest, because their populations, documentation practices, and data pipelines diverge most sharply from the North American and European settings in which most clinical AI is developed.

1.3 Research gap

The methodological literature on clinical prediction models is substantial, but it concentrates on development-time rigour and on one-off external validation. Comparatively little of it is operationalised into a standing protocol that a deploying institution — particularly a resource-variable one — can apply before, during, and after go-live. Reporting guidelines specify what to disclose; they do not, by themselves, constitute a deployment discipline. The gap this paper targets is the absence of a concrete, auditable evaluation protocol that treats a reported accuracy figure as a hypothesis about a specific setting rather than as evidence of fitness for it.

1.4 Objectives

- **O1.** To quantify, from peer-reviewed evidence, the size and character of the performance-transfer gap in deployed clinical AI.
- **O2.** To identify the mechanisms that generate the gap, in a form specific enough to make evaluation targeted rather than ritual.
- **O3.** To specify a five-component evaluation framework, aligned with consensus reporting standards, that a deploying institution can apply and audit.
- **O4.** To define an evidence-grading scheme that the framework applies symmetrically to third-party and first-party claims, including Curely's own.

1.5 Contributions

This paper makes four contributions. First, it consolidates the strongest public evidence that clinical model performance does not travel, and states the effect in absolute clinical terms rather than in summary discrimination statistics alone. Second, it decomposes the transfer gap into three mechanisms — population shift, data-and-workflow shift, and the human-factors gap — and maps each to the evaluation step that addresses it. Third, it operationalises these into a deployment-grounded protocol whose central move is to run new models in *silent mode* until local evidence confirms their behaviour. Fourth, it applies a single evidence-grading discipline to all claims, so that a vendor benchmark — including one of Curely's own — is never presented as established fact.

2 Related Work and Literature Review

2.1 External validation and the transfer gap

The clearest public demonstration that internal performance does not carry to deployment comes from external validations of proprietary EHR-embedded tools. The external validation of a widely implemented proprietary sepsis prediction model across 27,697 patients and 38,455 hospitalizations at a large United States academic centre found a hospitalization-level AUC of 0.63, substantially below the developer-reported range of 0.76 to 0.83 (Wong et al., 2021). More consequentially, at the alerting threshold the hospital used in practice, the model identified only 33% of sepsis cases and carried a positive predictive value of 12%. This work is treated here as *strong evidence* — a large external-validation cohort — and anchors the empirical case in Section 5.

The broader prediction-model literature explains why such gaps recur. Discrimination — the ability to rank patients by risk — and calibration — the agreement between predicted and observed absolute risk — are distinct properties, and a model can retain acceptable discrimination while becoming badly miscalibrated after a change of population. Reporting one without the other conceals exactly the failure mode most dangerous at the bedside, where absolute risk drives action.

2.2 Distribution shift and recurrent local validation

A growing methodological literature argues that external validation, while necessary, is neither sufficient nor stable over time, because patient data shift across time, geography, and facility, producing volatility in the performance of any single fixed model. On this view, generalisability is better pursued through *recurrent local validation* of individual deployed model instances than through a one-time external check (Youssef et al., 2023). We treat this as *emerging evidence* — a methodological argument rather than a completed outcome study — and it directly motivates the continuous-surveillance component of our framework.

2.3 Human-centred deployment studies

Accuracy on curated data is not accuracy in a clinic. A human-centred evaluation of a deep-learning diabetic-retinopathy screening system deployed across clinics in Thailand found that image-quality thresholds calibrated on pristine research images caused the system to reject a substantial share of real-world photographs, returning patients for repeat visits and eroding trust (Beede et al., 2020; *limited evidence*, single field study). A subsequent multisite national prospective screening programme showed that real-world performance was shaped by socio-environmental factors — staffing, lighting, patient flow, and result delivery — not by model accuracy alone (Ruamviboonsuk et al., 2022; *moderate evidence*, prospective interventional cohort). Automation bias compounds the problem in the opposite direction: under time pressure, clinicians

over-trust a confident-looking output and cease to check it independently, converting decision support into unexamined decision-making.

2.4 Consensus reporting standards

The field has responded to a documented history of incomplete reporting with consensus guidelines: TRIPOD+AI for the development and validation of prediction models (Collins et al., 2024); DECIDE-AI for early-stage live clinical evaluation of AI decision-support systems (Vasey et al., 2022); and the SPIRIT-AI and CONSORT-AI extensions for the protocols and reports of AI clinical trials (Cruz Rivera et al., 2020; Liu et al., 2020). These are *strong evidence* in the sense of formal consensus guidance, and they make a claim auditable — but adherence to a reporting standard is not, by itself, evidence of clinical benefit.

2.5 Positioning of this work

Prior work establishes that transfer gaps exist, explains their statistical origin, documents their human-factors dimension, and standardises disclosure. What remains under-specified is a single, auditable protocol that binds these threads into a deployment discipline usable by institutions with limited validation capacity. This paper supplies that protocol and applies a uniform evidence grade to every claim within it.

3 Mechanisms of Performance-Transfer Failure

Three mechanisms account for most of the observed transfer gap. Naming them precisely is what makes the evaluation framework in Section 4 targeted rather than ceremonial: each mechanism has a corresponding safeguard.

3.1 Population shift

A model encodes the base rates, comorbidity patterns, and outcome definitions of its training population. Deployed into a population with a different age structure, disease prevalence, or documentation habit, its calibration drifts even when its ranking ability is preserved. This is why models trained overwhelmingly on North American and European cohorts warrant particular scrutiny before use in African and other under-represented health systems, where both the epidemiology and the recorded data differ in ways the model never observed.

3.2 Data and workflow shift

Models assume that inputs arrive in the same form, at the same time, from the same instruments as in training. In practice, image quality, device characteristics, sampling timing, and documentation completeness all vary between the laboratory and the clinic. The retinopathy field evidence illustrates the point: a system accurate on research images rejected clinic photographs taken under ordinary conditions, so that the operative failure was upstream of the classifier entirely.

3.3 The human-factors gap

A correct output helps only if it reaches a clinician who can act on it within the available time. Real-world screening performance is governed by staffing, lighting, patient flow, and how results are delivered, and automation bias runs both ways — an output that is trusted uncritically is as hazardous as one that is ignored. Model accuracy is a necessary but not sufficient condition for clinical benefit.

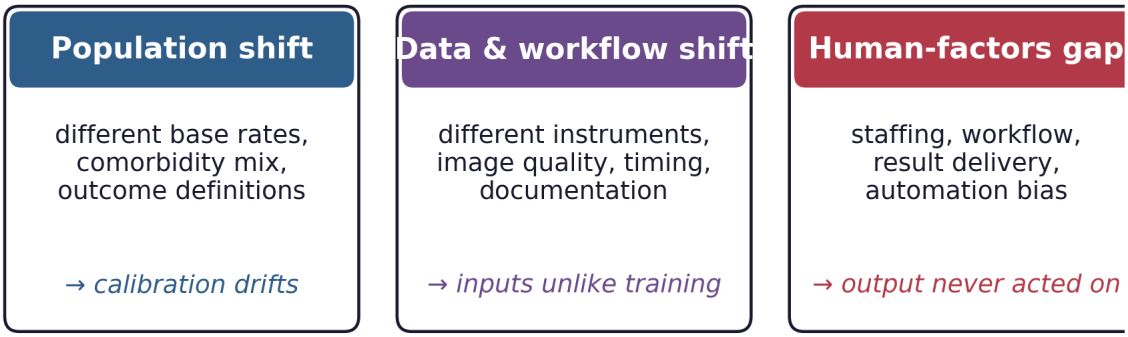

Figure 1. *The three mechanisms of performance-transfer failure. Each acts at a different layer of the deployment stack — the population the model reasons about, the data and workflow that feed it, and the human system that must act on its output — and each degrades realised performance independently of the others.*

4 Methodology: A Deployment-Grounded Evaluation Framework

4.1 Design principle

The framework rests on a single principle: a vendor accuracy claim — *including Curely's own internal benchmarks* — is a hypothesis to be tested in the deployment setting, not evidence of fitness for it. The protocol has five components, described below as commitments; results are reported against them per deployment rather than asserted in advance. Figure 2 shows the pipeline and its feedback loop.

4.2 Component 1 — Local validation before deployment

External validation on another institution's data is necessary but not sufficient. Before a model informs care at a site, it is validated on data representative of that site, and discrimination and calibration are reported separately, because a model can rank patients acceptably while being systematically miscalibrated on absolute risk. Where the evidence base for recurrent local revalidation over time is emerging, the framework follows it (Youssef et al., 2023).

4.3 Component 2 — Subgroup reporting with real sample sizes

Aggregate accuracy conceals subgroup failure. Performance is reported broken down by the axes that matter clinically and demographically for the population in question, with the sample size stated in each subgroup — a fairness claim resting on forty patients is not a claim. Where a subgroup is too small to evaluate, that is disclosed rather than absorbed into a reassuring aggregate.

4.4 Component 3 — Adherence to consensus reporting standards

Model documentation and evaluation are aligned with the established guidelines: TRIPOD+AI for development and validation, DECIDE-AI for early-stage live clinical evaluation, and SPIRIT-AI and CONSORT-AI for AI trial protocols and reports (Collins et al., 2024; Vasey et al., 2022; Cruz Rivera et al., 2020; Liu et al., 2020). These

guidelines exist because the field has a documented history of incomplete reporting, and following them is the least costly way to make a claim auditable.

4.5 Component 4 — Silent-mode monitoring before the model touches decisions

New models run in shadow: predictions are generated, logged, and compared against observed outcomes, but are not shown to clinicians, until local evidence confirms the model performs as claimed. This is the step that would have surfaced the sepsis-model shortfall before any alert fatigued any clinician.

4.6 Component 5 — Continuous post-deployment surveillance for drift

A validated model is not permanently valid. Populations, documentation practices, and upstream data pipelines change, and calibration drifts with them. Deployed-model performance is monitored on an ongoing basis, with thresholds that trigger revalidation or rollback. This matters most where no external regulator performs post-market surveillance — which describes the majority of the settings this framework is built for.

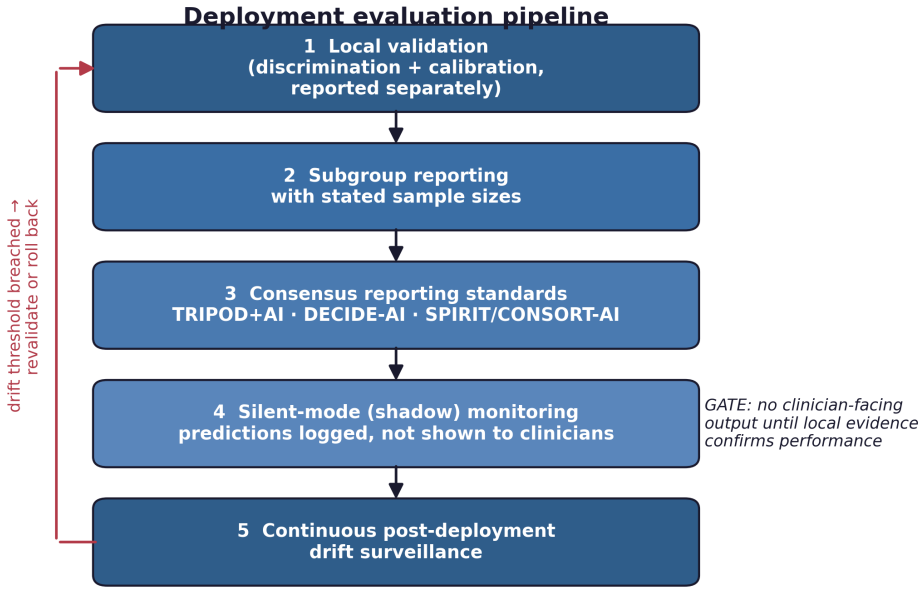

Figure 2. *The five-component deployment evaluation pipeline. A model advances only when local evidence confirms performance; the clinician-facing gate (right) prevents any output from influencing care during silent-mode monitoring. The red feedback path returns a drifting model to revalidation or rollback, making evaluation a standing loop rather than a one-time event.*

4.7 A symmetric evidence-grading scheme

Every empirical claim carries an evidence tier, and the same discipline is applied to first-party statements. A Curely internal benchmark is, by this scheme, vendor-tier evidence until an independent or prospective study raises it. Grading one's own results honestly is the only way the scheme retains meaning when applied to others (Table 3).

5 Evidence Synthesis

A note on framing. This section synthesises external, peer-reviewed evidence to demonstrate the transfer gap and to test the framework's mechanisms against real cases. It does *not* report outcome data from a completed Curely deployment; no such claim is made or implied.

5.1 Case 1 — a proprietary sepsis model at scale

The sepsis-model validation is the flagship case because the tool was embedded in one of the most widely used EHR systems and deployed at hundreds of hospitals before an independent group published its external validation. Table 1 juxtaposes the developer-reported figures with the validated ones. The AUC drop from 0.76–0.83 to 0.63 understates the clinical problem; the operational reality is that the alert caught one sepsis case in three, missed 1,709 of 2,552 septic patients, and fired on 18% of all hospitalizations at a 12% positive predictive value. A tool marketed as early warning was missing most of the events it targeted while generating enough false alarms to fatigue its users. Figure 3 renders the contrast.

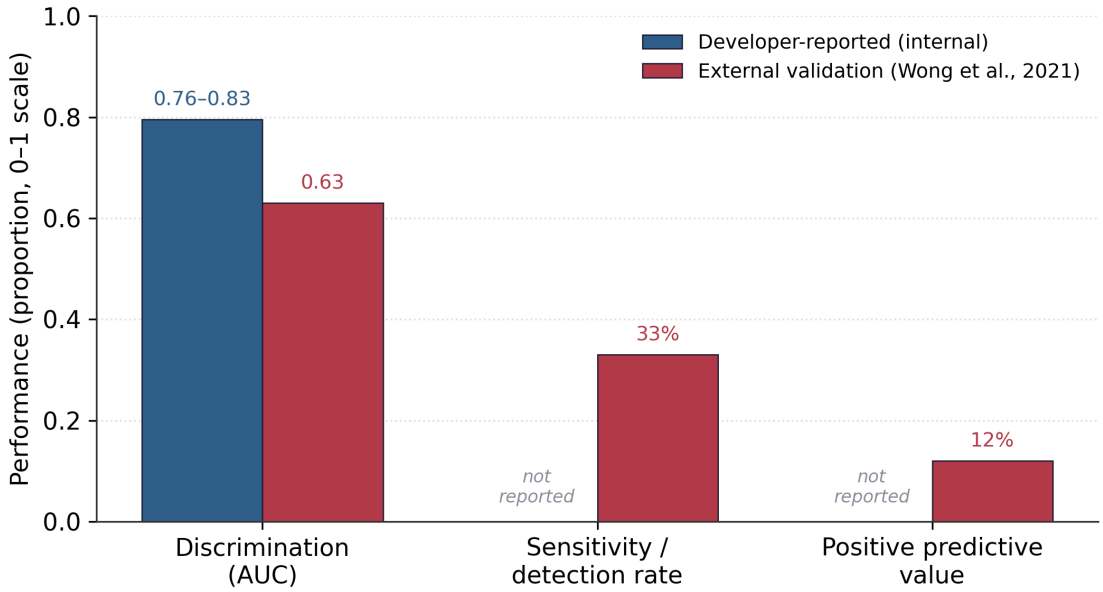

Figure 3. *Developer-reported versus externally validated performance of a widely deployed proprietary sepsis model (Wong et al., 2021). The internal AUC range (blue) and the external result (red) diverge; the operational sensitivity and positive predictive value — which the internal report did not surface — reveal the clinical shortfall a single AUC obscures.*

The lesson is not that one vendor built a poor model. It is that impressive internal numbers can accompany a model into wide deployment while being substantially wrong for the adopting setting, with no one aware until the external validation is run — a risk that compounds when the buyer has limited capacity to run its own check.

5.2 Case 2 — retinopathy screening from laboratory to clinic

The retinopathy evidence isolates the data-and-workflow and human-factors mechanisms. A system accurate on research images rejected a substantial share of clinic photographs because its quality threshold was calibrated on pristine inputs (Beede et al., 2020), and a national prospective programme showed realised performance turning on staffing, lighting, patient flow, and result delivery rather than on classifier accuracy (Ruamviboonsuk et al., 2022). Neither failure is visible in a development-time accuracy figure; both are visible in silent-mode monitoring on real inputs and in continuous surveillance after go-live.

5.3 Cross-case synthesis

Mapping the two cases onto the three mechanisms and the five framework components yields Table 2. Each mechanism has a corresponding safeguard, and in each case the safeguard would have surfaced the failure before it reached a patient: local validation and separate calibration reporting for the sepsis case; silent-mode monitoring on clinic images and continuous surveillance for the retinopathy cases. The synthesis supports the framework's central claim — that a reported accuracy figure describes a model in a setting, and that the setting must be evaluated, not assumed.

Table 1. *Epic Sepsis Model — developer-reported versus externally validated performance (Wong et al., 2021). Absolute clinical figures, not the summary AUC alone, reveal the operational shortfall.*

Discrimination (AUC)	0.76 – 0.83	**0.63**
Sensitivity at operational alert threshold	Not reported	33% (843 of 2,552 detected)
Septic patients missed	—	**1,709 of 2,552**
Alert burden (hospitalizations flagged)	—	18% of all hospitalizations
Positive predictive value	Not reported	**12%**
Validation cohort	Developer internal	27,697 patients / 38,455 hospitalizations; single U.S. academic centre

Table 2. *Cross-case synthesis: each failure mechanism, an illustrative peer-reviewed source, how it manifests in deployment, and the framework component that primarily mitigates it.*

Population shift	Sepsis model external validation (Wong et al., 2021)	Calibration and detection collapse on a new population; two-thirds of events missed	Local pre-deployment validation; separate calibration reporting (C1)
Data & workflow shift	Retinopathy clinic field study (Beede et al., 2020)	Clinic images rejected by lab-tuned quality gate; repeat visits, lost trust	Silent-mode monitoring on real inputs before go-live (C4)
Human-factors gap	National screening programme (Ruamviboonsuk et al., 2022)	Socio-environmental factors and automation bias govern realised benefit	Continuous surveillance; DECIDE-AI live evaluation (C3, C5)

Table 3. *The evidence-grading scheme, applied symmetrically to third-party and first-party claims. A Curely internal benchmark is vendor-tier until an independent or prospective study raises it.*

Strong	Systematic reviews, multiple randomised trials, or formal regulatory / consensus guidance	Sepsis external validation; TRIPOD+AI, DECIDE-AI, CONSORT-AI
Moderate	A single well-designed trial or a large prospective cohort	National prospective retinopathy screening programme
Limited	A single observational study, a small sample, or a preprint	Single human-centred retinopathy field study
Emerging	Early pilots or first-in-class results not yet replicated	Recurrent-local-validation methodological argument
Vendor / anecdotal	A company benchmark or single case; never presented as established fact	Any internal accuracy benchmark, including Curely's own, prior to independent study

6 Discussion

6.1 Significance

The most useful habit a deploying institution can build is to stop treating a reported accuracy figure as a property of the model and start treating it as a property of the model in a specific setting. The questions that follow — where the number came from, on which population, at what sample size, and whether it was validated anywhere other than where it was built — are not obstacles to adoption. They are how adoption is done without harming patients.

6.2 Clinical implications

For the clinician, the framework reframes an AI output as a claim with a provenance rather than an oracle. Silent-mode monitoring protects the therapeutic relationship during the period of greatest uncertainty, when a new tool has not yet earned local trust, and separate calibration reporting ensures that the absolute risks driving clinical action are the ones that have been checked. Guarding against automation bias is as much a part of safe deployment as guarding against model error.

6.3 Technical implications

For the developer and the deploying data team, the framework converts evaluation from a milestone into a standing service. Discrimination and calibration are monitored as separate signals; subgroup performance is tracked with explicit denominators; and drift thresholds are defined in advance with pre-agreed revalidation or rollback actions. The recurrent-local-validation view (Youssef et al., 2023) implies that the evaluation infrastructure, not the frozen model artefact, is the durable deliverable.

6.4 Comparison with prior frameworks

Consensus reporting guidelines specify what to disclose at defined milestones — development (TRIPOD+AI), early live evaluation (DECIDE-AI), and trials (SPIRIT-AI, CONSORT-AI). The present framework is complementary rather than competing: it binds these disclosures into a continuous deployment loop and adds two elements the guidelines do not operationalise — a mandatory silent-mode gate before any clinician-facing use, and a symmetric evidence grade that constrains first-party claims as strictly as third-party ones.

6.5 Application within Curely AI

Curely applies this framework to its own clinical systems — spanning clinical decision support, remote patient monitoring, and hospital-workflow intelligence — because the health systems it serves have the least room to absorb a model that fails quietly. Consistent with the grading scheme, any internal Curely benchmark is treated as vendor-tier evidence until an independent or prospective study raises it, and results are reported against the five components per deployment. Advanced healthcare intelligence is worth extending to under-resourced systems only if it is intelligence that has been checked in those systems.

6.6 Limitations

This paper describes a protocol and reviews external evidence; it does not report outcome data from a completed Curely deployment, and should not be read as such. The strongest cited evidence — the sepsis-model validation — comes from a single large United States academic centre, and its specific numbers no more transfer to other settings than the model itself did, which is precisely the point. The retinopathy evidence is drawn from a small number of programmes and should be read as illustrative of a mechanism rather than as a precise effect size. The reporting guidelines the framework commits to are consensus standards, not

guarantees of clinical benefit. The framework reduces, but does not eliminate, the risk that a model works in a paper and fails at a bedside; any party claiming to have eliminated that risk should be asked for its external-validation data.

7 Conclusion and Future Work

Clinical AI performance is not portable. A model validated in one population, health system, and data pipeline may fail — silently — in another, and the failure is most likely and least detectable in the resource-variable settings that stand to gain most from these tools. This paper has quantified the transfer gap from peer-reviewed evidence, decomposed it into population, data-and-workflow, and human-factors mechanisms, and specified a five-component evaluation framework — local validation, sample-size-honest subgroup reporting, consensus-standard adherence, silent-mode monitoring, and continuous drift surveillance — that treats a reported accuracy figure as a hypothesis about a setting rather than a property of a model.

Future work should move the framework from commitment to evidence. Priorities include prospective, DECIDE-AI-aligned evaluations of silent-mode deployments in under-represented health systems; empirical calibration of drift thresholds that balance sensitivity to degradation against alert fatigue; and independent, third-party validation of first-party benchmarks so that vendor-tier claims can be raised to a higher evidence tier on their merits. The framework's value is not that it removes risk, but that it makes risk visible early enough to act on — which is the minimum standard responsible deployment should meet.