



From Algorithm to Bedside: Clinically Actionable Artificial Intelligence for Early Disease Detection Using Multimodal Laboratory Biomarkers and Point-of-Care Diagnostics in Resource-Limited Settings

Zia Ullah Shah^{1,*}¹ Department of Medical Laboratory Technology, Kohat University of Science and Technology, Kohat, Khyber Pakhtunkhwa, Pakistan* Corresponding author: bukhariziaullahshah@gmail.com

ARTICLE INFO

Article history:

Received: 25-09-2026

Received in revised form: 28-09-2026

Accepted: 08-10-2026

Available online: 10-10-2026

Keywords:

Artificial Intelligence

Early Diagnosis

Laboratory-Derived Biomarkers

Machine Learning

Point-of-Care Testing

Resource-Constrained Settings

Scoping Review

ABSTRACT

Artificial intelligence (AI) applied to laboratory-derived biomarkers and point-of-care (POC) diagnostics offers substantial potential for early disease detection, but analytical performance does not necessarily translate into clinical actionability, particularly in resource-constrained settings. This scoping review mapped evidence on AI-enabled early disease detection using laboratory biomarkers, POC diagnostics, and multimodal approaches, with emphasis on validation, clinical utility, affordability, equity, and LMIC transferability. Following Joanna Briggs Institute methodology and PRISMA-ScR guidance, relevant primary studies were identified through structured literature searching and citation screening, extracted using a standardized framework, and appraised using PROBAST+AI and QUADAS-2. Of 151 records identified, 86 studies met the inclusion criteria. Discriminative performance was generally high (AUROC frequently 0.80–0.99), but only 29% reported independent external validation, 16% prospective or field validation, and 12% explainability; 5 studies (6%) provided evidence of practical clinical utility or implementation. Multimodal AI represented 14% of studies, while 9 studies (10%) were judged at high risk of bias. Forty-three studies (50%) had explicit LMIC relevance, with the most translationally mature evidence relying on existing diagnostic infrastructure. Overall, substantial translational gaps remain between algorithmic performance and clinically actionable deployment. The proposed Algorithm-to-Bedside Readiness Framework and Resource-Limited AI Diagnostic Implementation Pathway provide structured approaches for advancing validated, equitable, and resource-appropriate diagnostic AI implementation.

1. INTRODUCTION

Delayed diagnosis remains one of the most consequential and preventable contributors to poor outcomes across acute and chronic disease alike. In sepsis, each hour of delayed recognition and appropriate treatment is associated with measurably worse survival, yet clinical recognition based on vital signs and clinician judgment alone is known to lag the underlying pathophysiology [1, 2]. In vector-borne and parasitic diseases such as malaria and dengue, the interval between symptom onset and confirmed diagnosis determines whether treatment or triage occurs before or after the point of irreversible deterioration [3, 4]. In chronic, insidious conditions chronic kidney disease, type 1 diabetes, systemic lupus erythematosus the diagnostic delay is measured not in hours but in months to years, during which asymptomatic organ or immune injury accumulates before a diagnosis is ever made [5]. Across this spectrum of acuity, the shared clinical problem is the same: outcomes improve when detection moves

earlier in the disease course, and the health system's capacity to detect early is bounded by the diagnostic tools available at the point of first contact [6].

Laboratory biomarkers and point-of-care (POC) diagnostics represent the most widely deployable class of tools for closing this gap, precisely because they do not require the imaging infrastructure, specialist interpretation, or lengthy sample-referral pathways that many advanced diagnostic modalities demand. A single complete blood count, a lateral-flow rapid test, or a panel of routine chemistry values can, in principle, be obtained within minutes at nearly any level of a health system, from a tertiary referral hospital to a rural primary-care clinic. The clinical value of these tools has historically been constrained less by their availability than by the interpretive burden they place on already-stretched clinical and laboratory staff: a single CBC report contains dozens of individual values whose joint pattern, not any single value in isolation, often carries the diagnostically relevant signal [7–9].

It is precisely this joint-pattern-recognition problem that artificial intelligence and machine learning are, in principle, well suited to address, and the past decade has seen rapid growth in AI applications across diagnostic medicine from image-based deep learning in radiology and pathology to tabular machine learning applied to electronic health record and laboratory data. Within laboratory medicine specifically, the International Federation of Clinical Chemistry and Laboratory Medicine has recognized the trajectory clearly enough to issue formal working-group recommendations for how machine learning should be developed, validated, and documented when applied to clinical laboratory data itself: an indication that the technology has moved from experimental curiosity to a matter requiring professional-body governance [10].

Multimodal data integration combining laboratory-derived biomarkers with POC diagnostic outputs, clinical variables, and, where available, imaging or genomic data offers a further, largely untested, opportunity: individual data modalities each capture only part of a disease process, and their joint analysis may in principle reveal patterns invisible to any single modality alone. Where this has actually been tested within the available evidence synthesized in this review, the results are encouraging: for example, an integrated genetic, clinical, and lifestyle cardiovascular risk model outperforming single-modality comparators within the same study but such head-to-head multi-modal-versus-single-modality comparisons remain the exception rather than the rule, and the promise of multi-modal AI in diagnostics is, at present, considerably ahead of the available evidence that would be needed to confirm it as a general principle rather than a set of isolated positive examples.

This evidence gap between technical promise and clinical evidence is the central translational problem this review is built to interrogate. A model that achieves an area under the receiver-operating-characteristic curve of 0.95 in a single retrospective cohort has showed that a pattern exists in that dataset; it has not showed that the pattern will hold in a different population, that clinicians will trust and act on its output, that the health system can deliver its output fast enough to matter, or that doing so improves a patient-relevant outcome. High documented performance and clinical usefulness are conceptually and empirically distinct claims, and, as this review's synthesis indicates, the literature at present overwhelmingly documents the former while rarely establishing the latter.

Resource-limited settings sharpen this problem rather than merely adding a further consideration to it. A diagnostic AI system that assumes reliable electricity, broadband connectivity, cloud computing access, or a supply chain for specialized reagents is not a scaled-down version of a solution appropriate for well-resourced settings: it is, in a clinically relevant sense, a different technology, requiring its own evidence base regarding accuracy, robustness, and appropriateness under the actual operating constraints of the health systems in which it would need to function [11]. A model validated exclusively in a tertiary academic hospital in a high-income country carries no inherent presumption of transferability to a rural primary-care clinic in a low-income country, and this review treats that transferability as an empirical question to be classified explicitly for every included study, rather than an assumption to be waived.

Several prior reviews have addressed adjacent pieces of this landscape: AI-enhanced point-of-care diagnostics for infectious disease in resource-constrained settings specifically, AI interpretation of point-of-care ultrasound in low-resource settings, AI in digital pathology, and AI-liquid-biopsy integration for cancer detection and this review's comparison against that existing literature (Table 1) confirms that each addresses a narrower slice of the landscape than the question posed here: narrower by disease (infectious disease only, or oncology only), by modality (imaging only, or liquid biopsy only), or by both. None integrates laboratory-derived biomarkers and POC diagnostics across disease domains while simultaneously applying a structured, evidence-hierarchy-based framework for distinguishing technical feasibility from clinical usefulness and actionability [12].

The gap this review addresses, therefore, is not the absence of any prior synthesis of AI in diagnostics the literature on that broader topic is now large but the absence of a synthesis organized around the specific translational question of what evidence currently supports moving from an analytically promising algorithm to a clinically actionable, validated, affordable, and equitably deployable early-detection system, evaluated consistently across disease domains and explicitly stratified by resource setting. This review's purpose is to answer that question as rigorously as the currently available evidence allows, to state plainly where the evidence is strong and where it is not, and to propose a structured framework grounded in what this evidence base actually shows rather than in aspirational claims about AI's potential for what remains to be established before these systems can be responsibly deployed at the point of care in the settings that most need them.

This review makes four contributions that, to our knowledge, have not previously been combined within a single synthesis. First, it evaluates laboratory-derived biomarkers and point-of-care diagnostics jointly and across disease domains, rather than restricting attention to a single technology (point-of-care testing alone, or liquid biopsy alone) or a single disease category (infectious disease alone, or oncology alone), as each of the six most closely related prior reviews identified in this synthesis does (Table 1). Second, it operationalizes and applies, across all included studies, an explicit seven-level evidence-actionability hierarchy that separates technical feasibility from showed practical clinical utility allowing the review to report, for the first time in this literature to our knowledge, the proportion of AI-diagnostic studies that reach genuine deployment-stage evidence rather than only reporting aggregate discrimination statistics. Third, it classifies every included study's relevance to resource-constrained settings on an explicit five-category transferability scale (LMIC-developed, LMIC-validated, LMIC-deployed, high-income-evidence-potentially-transferable, or uncertain transferability), rather than treating "LMIC-relevant" as a binary label. Fourth, it derives two original conceptual frameworks directly and traceably from patterns observed in the included evidence the Algorithm-to-Bedside Readiness Framework and the Resource-Limited AI Diagnostic Implementation Pathway intended as practical tools for researchers, laboratories, and health systems assessing the translational maturity of a given AI-diagnostic system, rather than as generic recommendations

Table 1. Comparison of this review with the closest related prior reviews identified in the search.

Previous review	Year	Scope	Population/modality focus	LMIC focus	Validation emphasis	Limitation noted	What this review adds
Farquhar et al. (AI-Enhanced POC Diagnostics for Infectious Diseases in Resource-Limited Settings)	2026	POC diagnostics, infectious disease only	POC only	Yes, primary focus	Notes only 6.8% prospective field validation	Infectious disease only; POC only	Adds non-infectious disease domains, laboratory (non-POC) biomarkers, and cross-domain multi-modality analysis
AI-Enhanced POC Ultrasound in Low-Resource Settings (scoping review)	—	POCUS imaging	Imaging only	Yes	29 studies from 1196 records	Imaging modality only, single technology	This review excludes imaging-only studies by design but incorporates POC diagnostics broadly, not just ultrasound
Radiologic al imaging AI in LMICs (scoping review)	—	Radiological imaging	Imaging only	Yes	51 studies, 33 countries	Imaging-only; does not address lab/POC biomarkers	Directly complementary: this review deliberately covers the non-imaging, laboratory/POC space that a radiology-

Previous review	Year	Scope	Population/modality focus	LMIC focus	Validation emphasis	Limitation noted	What this review adds
							focused review does not address
Liquid biopsy + AI systematic review/meta-analysis	2026	Oncology, liquid biopsy	Oncology only	No specific LMIC focus	Meta-analysis of cfDNA-based detection	Single-disease-category (cancer), single-modality (liquid biopsy)	This review situates liquid biopsy as one modality among many, compared across disease domains
AI for strengthening healthcare systems in LMICs (scoping review)	—	Health systems broadly	Not diagnostics-specific	Yes	Notes trust/cost-effectiveness gaps	Not diagnostics-focused	This review provides the diagnostics-specific evidence this broader review does not attempt
Leukemia AI/DL diagnosis (scoping review)	—	Leukemia only	Imaging (blood/marrow smears)	No	Dataset-focused (ALL-IDB, etc.)	Single disease	This review situates leukemia-AI findings within the broader hematological-malignancy and multi-disease evidence base
Racial bias in clinical ML (scoping review)	—	Algorithmic bias, cross-disease	Not disease-specific	No	12 studies on bias-mitigation methods	Methodology-focused, not disease-early-detection-focused	This review identifies the near-total absence of subgroup reporting within disease-specific early-detection studies specifically, as distinct from the general bias-methods literature: a more concrete, actionable gap than the general bias-methods literature addresses

2. METHODS

2.1 Protocol and Registration

The review followed Joanna Briggs Institute (JBI) methodology for scoping reviews and was documented according to the PRISMA extension for Scoping Reviews (PRISMA-ScR) [13]. A review protocol specifying the objectives, research question, eligibility criteria, and search strategy was developed in advance and is available as supplementary material. The protocol was registered on the Open Science Framework (OSF Registries).

2.2 Research Question

This review addressed the following question, framed using the Population–Concept–Context (PCC) structure recommended for scoping reviews: What is the extent, nature, and methodological maturity of the evidence on AI-enabled systems that use laboratory biomarker or point-of-care diagnostic data for early detection of disease, and to what extent does this evidence support deployment in resource-constrained settings? Population: patients or populations evaluated for early detection of disease, diagnosis, screening, or risk stratification in any healthcare context. Concept: application of AI or machine learning to laboratory-derived biomarkers, point-of-care diagnostic outputs, or multi-modal combinations of these. Context: clinical healthcare settings, with particular attention to resource-limited and LMIC settings.

2.3 Eligibility Criteria

Primary studies that evaluated AI or machine learning for early detection of disease, diagnosis, screening, or clinically relevant risk identification, using laboratory-derived biomarkers, point-of-care diagnostic outputs, or multi-modal combinations of these, in human healthcare contexts, with sufficient methodological detail to extract population characteristics, diagnostic modality, AI method, and at least one performance metric, were considered eligible. Studies were excluded when they were purely technical AI papers without a healthcare application; applied AI only to treatment optimization or administrative workflows; applied AI exclusively to radiological or histopathological imaging without a laboratory or point-of-care component; were editorials or conference abstracts lacking extractable data; or were duplicate publications. Each included study was independently classified for its relevance to LMIC settings rather than assumed, using five categories: LMIC-developed, LMIC-validated, LMIC-deployed, potentially transferable from a high-income-country setting, or of uncertain transferability.

2.4 Operational Definitions

The principal terms used in this review were defined operationally in Table 2, to avoid the ambiguity that affects much of the underlying literature.

Table 2. Definitions and conceptual boundaries used throughout this review.

Term	Operational definition used in this review
Early detection	Identification of disease before the point at which it would otherwise have been clinically recognized through standard care-seeking pathways; encompasses screening, preclinical detection, and early-stage-post-symptom diagnosis as distinct sub-categories (below)
Screening	Testing of asymptomatic or undifferentiated individuals not selected on the basis of a presenting complaint referable to the target condition
Preclinical detection	Identification of biological evidence of disease before any clinical symptom has emerged (e.g., the SLE proteomic model predicting disease before clinical diagnosis)
Early-stage diagnosis after symptom onset	Diagnostic confirmation of a condition that has already produced a presenting complaint, distinguished from risk prediction or screening
Risk prediction	Estimation of future probability of developing a condition, distinct from detecting a condition already present
Multimodal AI	Genuine integration of data from different measurement domains (e.g., laboratory + imaging, or genomic + clinical) within a single model. Multiple biomarkers drawn from the same underlying panel (e.g., several CBC indices) do not constitute multi-modality under this definition
Clinically actionable	An AI output that (i) changes a clinical decision, (ii) is interpretable by the intended clinical user, (iii) uses a threshold with clinical meaning, (iv) is available within a clinically relevant timeframe, and (v) has an established subsequent clinical action: evaluated per this review's 7-level evidence-actionability hierarchy (Figure 2), not inferred from discrimination statistics alone
LMIC-developed	Model trained on data originating in a low- or middle-income-country setting
LMIC-validated	Model (regardless of development setting) with validation performed on independent LMIC data

Term	Operational definition used in this review
LMIC-deployed	System with documented use in routine or pilot clinical service in an LMIC setting
HIC-evidence, potentially transferable	Developed and validated exclusively in high-income-country settings, with no LMIC validation, but using inputs/methods plausibly transferable
Uncertain transferability	Insufficient information to classify, or specialized/costly inputs that make LMIC transfer implausible without considerable adaptation

Note. These definitions were applied consistently during screening and data extraction; studies that did not report enough detail to be classified against a given definition were noted as such rather than assumed to comply with it.

2.5 Information Sources and Search Strategy

A structured search of the literature was undertaken across biomedical and multidisciplinary sources, guided by the search terms and concept groups specified in the review protocol (artificial intelligence, early detection, laboratory-derived biomarkers, point-of-care diagnostics, and resource-constrained settings, in combination), and extended through backward- and forward-citation screening from relevant articles and reviews. Searching was conducted iteratively across about twenty-six disease and technology domains spanning infectious disease, sepsis, cardiometabolic disease, hematological malignancy, and maternal-child health, together with cross-cutting methodological themes including explainability, equity, and cost-effectiveness. Each candidate record was individually checked for existence and, where possible, current peer-review status before inclusion in the working library. Per-database retrieval counts and filters, as specified by PRISMA-S reporting guidance, were not separately logged, and this is addressed in the Limitations.

2.6 Selection of Sources of Evidence

Records were retrieved through structured searches organized by disease and technology themes across about 26 disease and technology domains, combined with backward and forward citation chaining; every candidate record was individually checked for existence and, where possible, current peer-review status before inclusion in the working library ($n = 151$). A further nine methodological and reporting-guideline documents (TRIPOD+AI/PROBAST+AI, FUTURE-AI, the WHO ethics and governance guidance, DECIDE-AI, PRISMA-ScR, QUADAS-2, JBI scoping-review guidance, Arksey and O'Malley's foundational framework, and the IFCC working-group recommendations on machine learning in laboratory medicine) were identified separately as methodological references rather than as disease-specific evidence and are cited accordingly throughout. No duplicate records were identified, as the working library was constructed additively with duplicate-checking applied at the time each record was added.

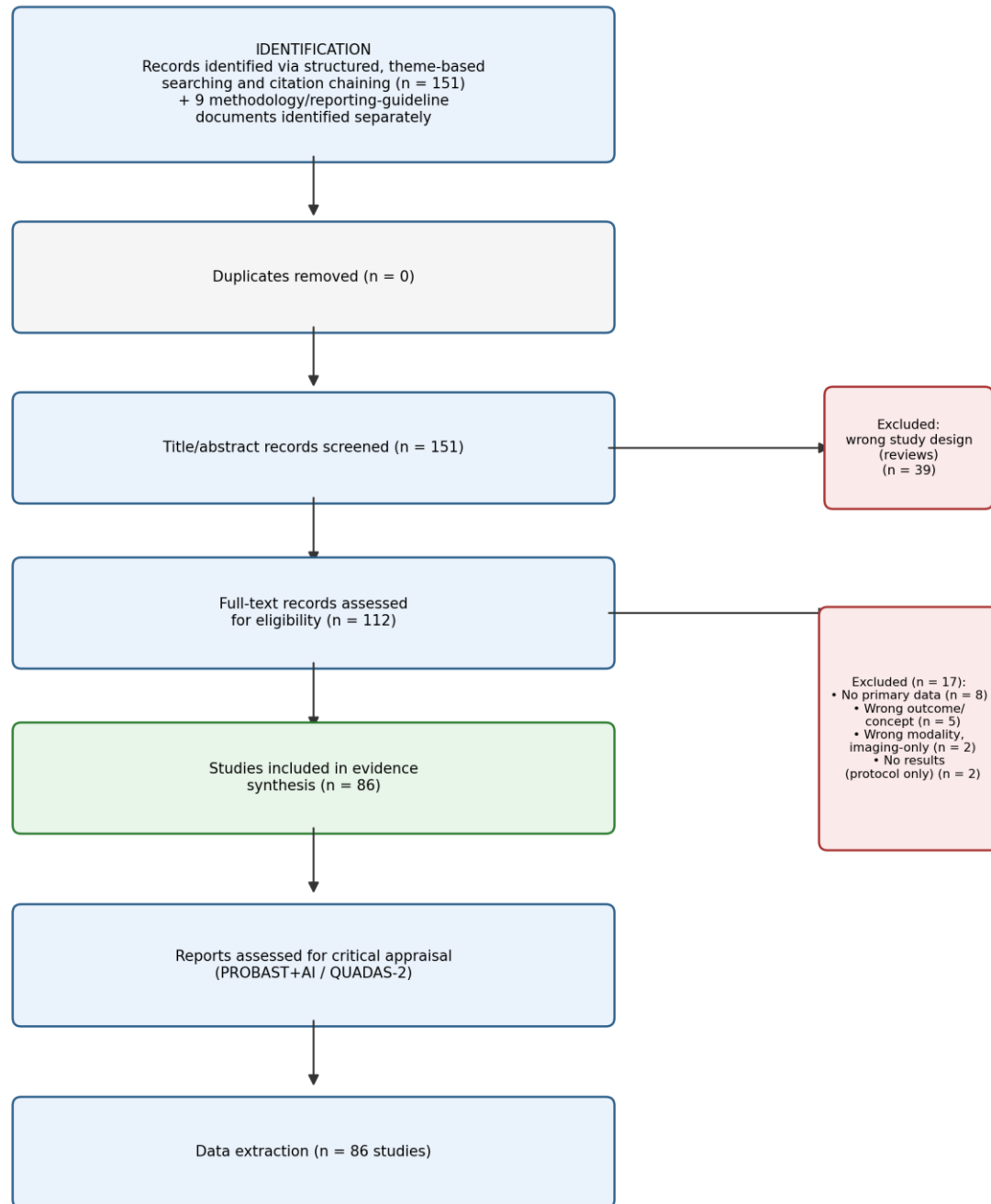
During title and abstract screening, 39 of the 151 records were excluded as reviews, meta-analyses, or narrative syntheses rather than primary research; these were retained separately as a comparator pool for the review-comparison analysis (Table 1) rather than discarded. The remaining 112 records were assessed in full text for eligibility, at which a further 17 were excluded: eight for lacking extractable primary data (conceptual proposals, narrative or perspective pieces, secondary surveys, or background citations not specific to disease early detection), five for addressing the wrong outcome or concept (behavioral or care-seeking prediction rather than diagnostic classification using laboratory or point-of-care data, a non-machine-learning statistical score, or population-level surveillance rather than patient-level diagnosis), two for using an ineligible modality (artificial intelligence applied exclusively to radiological or histopathological imaging without a laboratory or point-of-care component, per the pre-specified eligibility criterion), and two for reporting no results (trial registrations or study protocols without completed findings). After these exclusions, 86 studies remained for the evidence synthesis (Figure 1).

Compositionally, the included studies were concentrated in infectious disease and sepsis-related applications (COVID-19, tuberculosis, malaria, dengue, human immunodeficiency virus, hepatitis, antimicrobial resistance, and sepsis across adult, pediatric, and neonatal populations), which together accounted for about half of the 86 included studies, with the remainder distributed across cardiometabolic and renal disease, hematological malignancy, autoimmune disease, maternal-fetal health, and newborn screening. This distribution reflects the search strategy's deliberate weighting toward disease domains most relevant to the review's central question regarding resource-constrained settings, and should not be read as a representative epidemiological sample of the artificial-intelligence-in-diagnostics literature as a whole.

A structured, itemized log of all 151 records, their screening disposition, and where excluded the specific reason for exclusion, is provided as supplementary material, consistent with PRISMA-ScR's requirement for a documented and reproducible exclusion log.

PRISMA-ScR Flow Diagram

From Algorithm to Bedside: AI for Early Disease Detection Using Multimodal Laboratory Biomarkers and POC Diagnostics in Resource-Limited Settings



Separate pools (not part of n=86):
n = 39 reviews → comparison-with-previous-reviews analysis | n = 9 methodology/framework citations → Methods section

Figure 1. PRISMA-ScR flow diagram of study identification, screening, and selection.

2.7 Data Charting Process and Data Items

Study information was extracted with a standardized data-charting framework capturing bibliographic information, population characteristics, disease and outcome definitions, diagnostic modality, AI/ML method and architecture, validation strategy, documented performance metrics, explainability approach, resource requirements, equity considerations, cost data, and regulatory and ethical considerations. Performance metrics were extracted only as documented by the primary study; no metric was derived, imputed, or converted. A subset of studies identified as bearing most directly on the review's central questions underwent additional full-text verification to confirm exact performance figures, sample sizes, and bibliographic details.

2.8 Critical Appraisal

The included studies were critically assessed using PROBAST+AI for prediction-model development and validation studies (n=76) and QUADAS-2 with PROBAST+AI elements for studies with an index-test-versus-reference-standard diagnostic-accuracy design (n=10). Appraisal assessed participant and data-source selection, predictor and outcome definition, analysis methods, and applicability, with particular attention to overfitting, circular predictor-outcome relationships, small sample sizes, class imbalance, and absence of independent external validation [14, 15].

2.9 Synthesis of Results

Because of considerable clinical and methodological heterogeneity across disease domains, biomarker types, and outcome definitions, a quantitative meta-analysis was not performed, consistent with JBI guidance for scoping reviews. The findings were synthesized narratively across sixteen pre-specified result domains (Results) and classified against a seven-level evidence-actionability hierarchy technical feasibility, internal validity, external validity, prospective validity, practical clinical utility, implementation, and equity and scalability developed for this review to distinguish analytical performance from showed clinical usefulness and actionability (Figure 3) [16, 17].

3. RESULTS

3.1 Characteristics of the Evidence Base

Eighty-six studies met inclusion criteria (Figure 1, PRISMA-ScR flow diagram), drawn from a broader pool of 151 identified records after exclusion of 39 reviews/meta-analyses (retained separately for comparison in Table 1) and 17 records excluded for wrong study design, wrong outcome, wrong modality, or absence of documented results. The evidence base was disease-heterogeneous by design: no single condition accounted for more than 10% of included studies (COVID-19-related laboratory prediction, n=9, was the largest single cluster; dengue/typhoid, n=8; tuberculosis/malaria point-of-care applications, n=7; hematological/CBC-based applications, n=7). This heterogeneity is consistent with the review's scoping (rather than disease-specific systematic) design and precluded meta-analysis, in line with the pre-specified methodological framework.

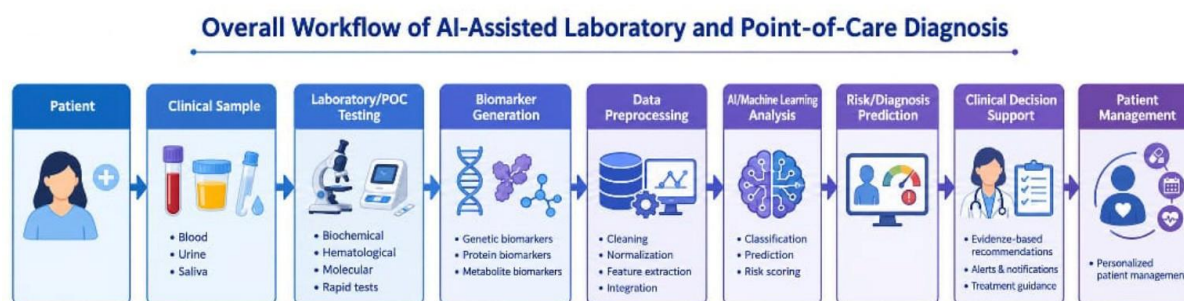


Figure 2. Overall workflow of AI-assisted laboratory and point-of-care diagnosis.

The workflow links sampling, laboratory/POC testing, biomarker generation, data processing, AI/ML analysis, risk prediction, decision support, and patient management. (ai generated image)

Publication years spanned from 2016 to 2026, with a pronounced concentration in 2024–2026 (about two-thirds of included studies), reflecting the rapidly evolving nature of this literature already anticipated in the protocol. Seven of the 86 included studies (8%) remain preprints without confirmed peer review at the time of this review (medRxiv and arXiv); these are explicitly labeled as such throughout the synthesis rather than treated as equivalent to peer-reviewed evidence.

Geographically, the available evidence was split between high-income-country-generated studies and a considerable and, in this particular corpus, better-represented-than-might-be-expected contingent of studies developed in or explicitly validated for low- and middle-income-country settings. Forty-three of 86 studies (50%) carried explicit LMIC relevance (developed in, validated in, or deliberately designed for a resource-limited context); thirteen were unambiguously LMIC-developed and -tested end to end (e.g., Zimbabwe, Bangladesh, Nepal, Ethiopia, Vietnam, Pakistan, Nigeria, Sri Lanka, Colombia, Rwanda, Uganda). This should not be read as representative of the field's overall geographic balance the literature search's search strategy deliberately weighted several themes (neonatal sepsis, dengue, malaria, sickle cell disease, malnutrition) toward LMIC contexts precisely because the review's central question required that evidence, and the resulting proportion is a function of search design as much as of the literature's true composition.

Table 3. Characteristics of the 86 included studies.

Characteristic	Category	n	%
Publication status	Peer-reviewed	79	92%
	Preprint (not yet peer-reviewed)	7	8%
Publication period	2016–2023	~26	30%
	2024–2026	~60	70%
Geographic/resource relevance	LMIC-developed (end-to-end)	13	15%
	Any explicit LMIC relevance (developed, validated, or purpose-designed)	43	50%
	HIC-only, no stated LMIC relevance	43	50%
Disease domain cluster	Infectious disease/sepsis (incl. COVID-19, dengue, TB, malaria, HIV, hepatitis, AMR)	~45	52%
	Non-communicable/chronic disease (CVD, CKD, diabetes, stroke, autoimmune)	~25	29%
	Maternal-child health, hematological malignancy, newborn screening	~16	19%
Modality type	Single-modality	74	86%
	Multimodal	12	14%
AI architecture (where classifiable)	Classical ML (RF/XGBoost/LR/SVM/ensemble)	19	22%*
	Deep learning (CNN/RNN/attention)	15	17%*
	Both compared within study	3	3%*
	Not clearly classifiable from available notes	49	57%*

Note. Percentages for AI-architecture classification are of the full N=86; the large "not classifiable" category reflects a named extraction-completeness limitation (Section 3.6) rather than an absence of describable methodology in the underlying studies.

3.2 Disease Domains Targeted for Early Detection

Included studies clustered into broad disease categories: infectious disease and sepsis (adult, neonatal, and pediatric sepsis; COVID-19; dengue/typhoid; tuberculosis; malaria; HIV; hepatitis B/C; antimicrobial resistance prediction) accounted for the largest share (about 45 of 86 studies), followed by non-

communicable and chronic disease applications (cardiovascular disease, chronic kidney disease, diabetes, stroke, autoimmune disease) at about 25 studies, and a smaller cluster of maternal-child health, hematological malignancy, and newborn-screening applications (about 16 studies). Oncology-focused multi-modal liquid-biopsy studies were comparatively underrepresented in the final included set ($n=2$) relative to their prominence in the wider literature search (a considerably larger number of oncology-AI records were identified but excluded at screening, the majority being reviews, narrative commentaries, or imaging-only studies falling outside the pre-specified eligibility criteria) a pattern that itself illustrates the risk of conflating a heavily published subfield with a well-validated one.

3.3 Laboratory Biomarkers Used by AI

The modalities most frequently paired with AI/ML across included studies were: routine hematological parameters (complete blood count and its derived indices neutrophil-to-lymphocyte ratio, monocyte distribution width, RDW), used across sepsis, cardiovascular, COVID-19, dengue, and hematological-malignancy applications; routine clinical chemistry (creatinine, liver enzymes, albumin, coagulation parameters); inflammatory/immunological markers (CRP, procalcitonin, D-dimer); and, in a smaller but analytically prominent subset, molecular/genomic data (whole-genome sequencing for antimicrobial resistance, gene-expression panels for stroke and severe dengue, blood transcriptomics for pediatric diabetes, serum proteomics for lupus). A clear and clinically relevant pattern emerged: studies restricted to routine, low-cost, widely available laboratory parameters (CBC-derived indices in particular) were both common and, per the critical appraisal (Section 3.5), more plausibly transportable to resource-constrained settings than studies relying on specialized panels (multi-gene expression signatures, whole-genome sequencing, multi-omic proteomic/metabolomic profiling), which whatever their analytical performance face a considerably higher infrastructure barrier to translation.

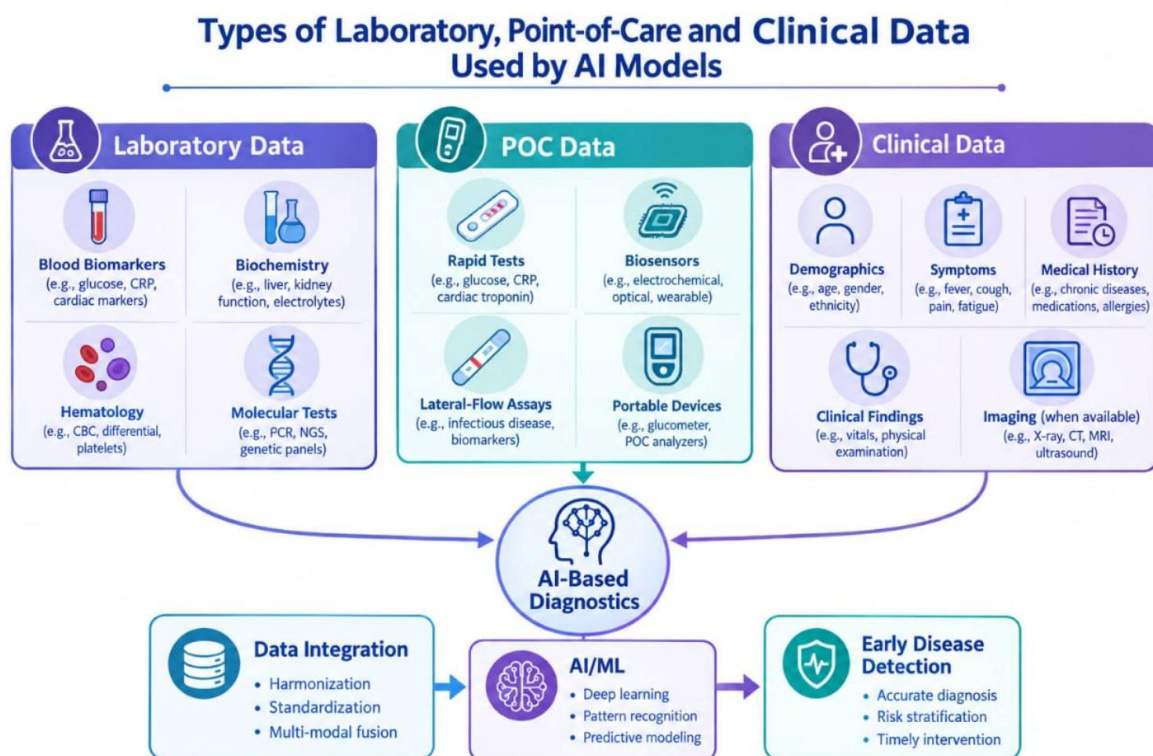


Figure 3. Laboratory, point-of-care, and clinical data sources used by AI models.

AI-based diagnostics may combine laboratory measurements, POC outputs, and relevant clinical information after harmonization and standardization..(ai generated image)

Table 4. Laboratory biomarker categories and representative AI applications identified in this review.

Biomarker category	Representative applications in this review	LMIC feasibility (this review's judgment)
Routine hematology (CBC and derived indices)	Sepsis detection (MDW+CBC), AMI classification (SHAP on CBC),	High routine, low-cost, widely available

Biomarker category	Representative applications in this review	LMIC feasibility (this review's judgment)
	anemia/leukemia screening, heart failure risk (CBC-derived, externally validated), dengue confirmation (single CBC exam)	
Routine clinical chemistry	Liver fibrosis (routine labs only, hepatitis B), CKD prognosis, stroke severity (routine blood/urine)	High to moderate
Inflammatory/coagulation markers	Sepsis (CRP/PCT/MR-proADM/lactate), COVID-19 mortality (D-dimer/platelets/neutrophil/lymphocyte), stroke (D-dimer, homocysteine)	Moderate some assays less universally available
Molecular/genomic (WGS, gene-expression panels)	Antimicrobial resistance prediction (WGS, MALDI-TOF), stroke gene-signature, severe dengue 8-gene panel, childhood-diabetes transcriptomics	Low near-term sequencing infrastructure-dependent
Proteomic/multi-omic	SLE serum proteomics, NAFLD stacked multi-modal (genetic+clinical), lung cancer multi-omics liquid biopsy	Low specialized, costly
Zero-/minimal-cost clinical predictors (no lab test)	Pre-eclampsia zero-cost screening (clinical characteristics only), CKD Bangladesh minimal non-pathology feature subset	Highest explicitly designed for no-lab-access settings

Note. Categories are illustrative rather than exhaustive and are not mutually exclusive; several included studies used biomarkers spanning more than one category.

3.4 Point-of-Care Diagnostic Technologies

POC-specific evidence clustered around three technology families: (i) smartphone- or low-cost-microscope-coupled image classification for microscopy-dependent diagnoses (malaria thick/thin blood smears, sickle cell morphology), where AI is used to augment or partially replace manual microscopy; (ii) lateral-flow/rapid-test image interpretation via smartphone camera and deep learning (HIV rapid tests, general LFA reading), directly targeting the well-documented human-interpretation-variability problem in rapid diagnostic testing; and (iii) paper-based/microfluidic molecular assays coupled to on-device or cloud AI interpretation (the Uganda malaria DNA-diagnostic system) [18]. Genuinely field-deployed POC-AI evidence (as opposed to laboratory bench validation of POC-adjacent technology) remained rare the Uganda malaria system, the South African HIV rapid-test classifier, and the UK CURIAL-Rapide laboratory-free triage tool were the clearest examples of POC-AI systems evaluated, at least in part, in real operational use rather than solely on retrospectively collected specimens [18–20]. Such devices increasingly sit within cloud- and IoT-connected smart-healthcare systems [21].

Table 5. Point-of-care technologies paired with AI in the included evidence.

POC technology	Applications identified	Validation maturity reached (this review)
Smartphone-camera + deep learning image classification	Malaria thick/thin blood smear reading, HIV lateral-flow rapid-test reading, sickle cell morphology screening	Level 3–6 (HIV reader: field pilot deployment; malaria: mixed, mostly Level 2–4)
Paper-based/microfluidic molecular assay + AI	Uganda malaria DNA diagnostic (blockchain-secured)	Level 4–6 (field-validated, Uganda)
Fluorescence/portable digital microscopy	<i>P. falciparum</i> detection (Tanzania)	Level 3–4
Laboratory-free clinical-variable triage + POC analyser	CURIAL-Rapide COVID-19 triage	Level 6 (routine deployment)
AI-assisted chest radiograph screening + molecular confirmation	Romania national TB screening program	Level 5–6 (national program scale)
Low-cost automated microscopy vs. reference-standard comparison	Sickle cell disease/beta-thalassemia multi-test comparison (Nepal/Canada)	Level 3–4 (multi-site comparative)

Note. Validation maturity is expressed using the seven-level evidence-actionability hierarchy defined in Section 3.11 and Figure 3.

3.5 Single-Modality Versus Multimodal AI

Applying this review's strict pre-specified definition of multi-modality (genuine integration of data types from different measurement domains e.g., laboratory plus imaging, or genomic plus clinical rather than multiple biomarkers drawn from the same underlying test panel), only 12 of 86 included studies (14%) were genuinely multi-modal; the remaining 74 (86%) were single-modality, most frequently laboratory-only. Among the 12 multi-modal studies, several explicitly tested whether multi-modal integration improved on single-modality performance rather than assuming it: the UK Biobank cardiovascular study (integrated genetic, clinical, and lifestyle data, AUC 0.85, higher than single-modality comparators documented within the same study) and the lung-cancer liquid-biopsy multi-omics study (multi-omic fusion improved discrimination relative to any single modality alone). This overall pattern multi-modal integration providing a real, quantified improvement over single-modality baselines in every study that actually tested the comparison is encouraging, but the small number of studies that performed this comparison (as opposed to simply asserting multi-modal superiority without a single-modality baseline) means the finding should be treated as suggestive rather than established. This review's caution against assuming multi-modal AI is automatically superior is corroborated: it is superior where tested, but it is tested infrequently [22, 23].

3.6 AI Architectures and Modelling Strategies

Classical machine-learning algorithms (random forest, XGBoost, gradient boosting, logistic regression, support vector machines, and related ensemble/tree-based methods) were used, alone or as the best-performing model among several compared, in at least 19 of 86 studies where architecture could be clearly classified; deep-learning architectures (convolutional neural networks for image-based tasks, recurrent/attention-based architectures for sequence or time-series data) appeared in at least 15 studies, concentrated heavily in image-based applications (microscopy, blood-smear, and radiograph classification) and blood-based deep phenotyping (the Gaussian-process/RNN sepsis model). A further 3 studies explicitly compared classical ML against deep learning within the same study; in the remaining about half of studies, the specific architecture could not be confidently classified from the extracted notes without full-text methods-section review, a genuine extraction-completeness limitation acknowledged here rather than resolved by inference. Feature-selection and interpretability-oriented preprocessing (LASSO, Boruta, SHAP-guided feature reduction) appeared as an explicit methodological step in a minority of studies, most consistently in the LMIC-oriented studies that were deliberately designing for minimal, low-cost feature sets (e.g., the Bangladesh CKD screening study's explicit comparison of full versus minimal non-pathology feature subsets). Neural computation is following a similar path in clinical image diagnostics [24].

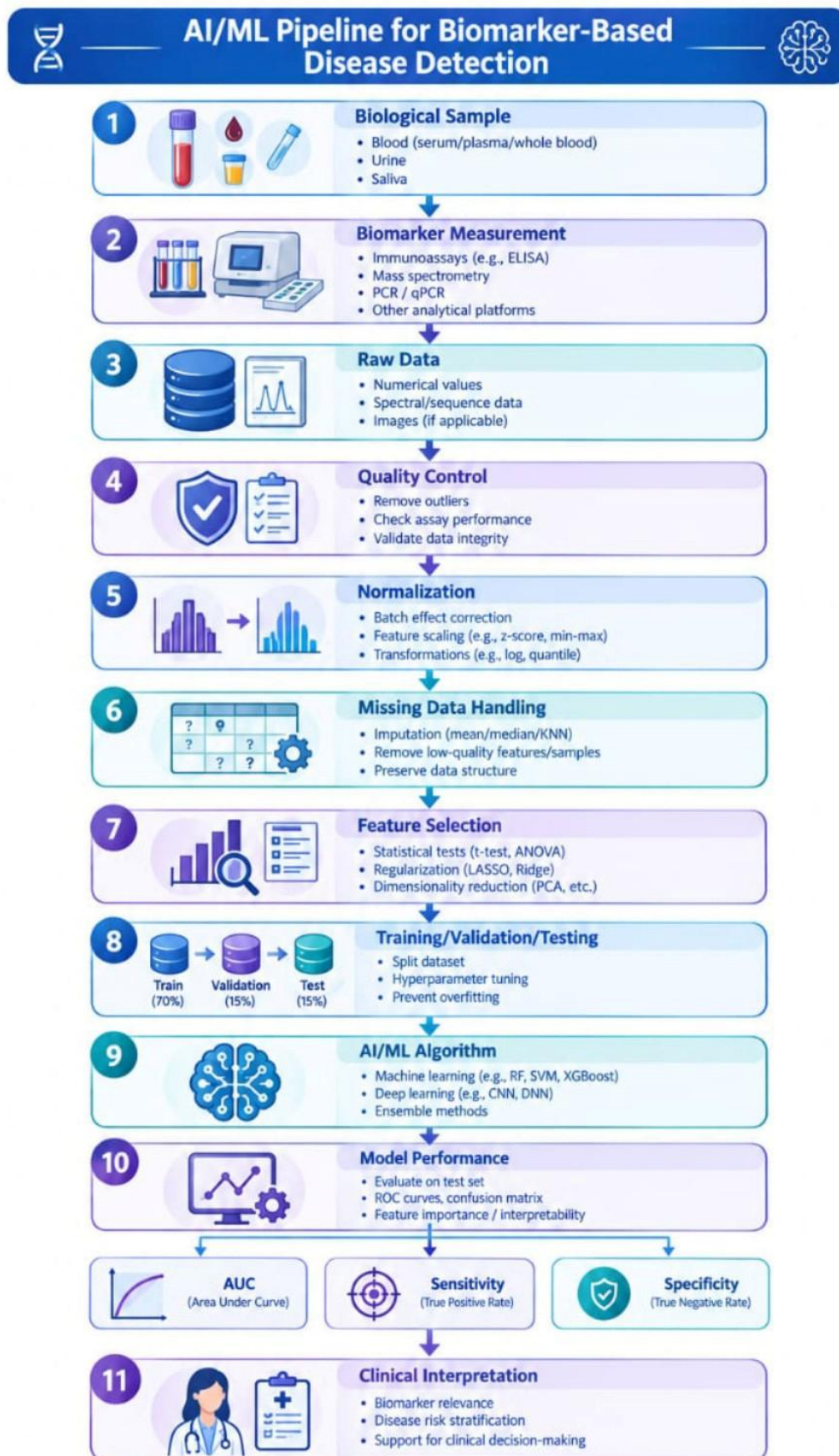


Figure 4. AI/ML pipeline for biomarker-based disease detection.

The pipeline covers sampling, biomarker measurement, quality control, preprocessing, feature selection, model development, evaluation, and clinical interpretation..(ai generated image)

3.7 Diagnostic and Predictive Performance

Across the included studies, documented performance was generally high when measured on the study's own internal or held-out test data: AUROC/AUC values in the 0.80–0.99 range were common across nearly every disease domain, and several studies documented sensitivity or specificity figures above 95%. This apparent consistency of high performance is precisely the pattern this review's interpretive framework warns against over-relying upon. Just documented metrics are presented here; no metric was derived, imputed, or converted between measures. A recurring and clinically clinically relevant asymmetry appeared in several studies: high specificity paired with comparatively modest sensitivity (the Hong Kong multi-wave COVID-19 study: AUC up to 95.8% but sensitivity only 55.5–77.8%), or high sensitivity paired with very low positive predictive value in a low-prevalence population (the Colombia severe-dengue model: sensitivity 86.4% but PPV only 20.9%). Both patterns illustrate that a single headline discrimination statistic (AUC) can obscure clinically consequential trade-offs that determine whether a model's output is actually actionable at the point of care directly illustrating this review's central distinction between technical performance and practical clinical utility [4, 25].

Table 6. Validation categories and representative documented performance ranges across the 86 included studies.

Validation category	n (of 86)	%	Representative AUC/performance range where documented
External/independent validation	25	29%	0.71 (external, decayed) to 0.97 (well-preserved)
Prospective/field validation	14	16%	Sens 87.5%–97.8%, spec 85.4%–100% (deployment-stage examples)
Internal validation only (explicit single-cohort)	11	13%	AUC/accuracy frequently >0.85, unverified beyond development data
Validation approach unclear/not documented	36	42%	Not comparable: reporting gap itself is the finding

Note. Performance ranges are documented exactly as stated in the source studies; no metric was derived, imputed, or converted between measures. The complete per-study extraction is available as a supplementary data file.

3.8 Internal Versus External Validation

The critical-appraisal findings showed that, only 25 of 86 studies (29%) documented external or independent-cohort validation, and a further 14 (16%) documented prospective or field validation; the remaining 47 studies (55%) documented internal validation only or did not clearly report their validation strategy. Where independent external validation was performed, performance frequently though not universally declined relative to internal/development-cohort figures (Section 3.7 above and the critical-appraisal synthesis, Pattern 2), with declines ranging from modest (KNHANES oculoimics model, essentially preserved) to considerable (Brazil-to-USA cardiovascular model, AUC 0.871→0.786; MALDI-TOF antimicrobial-resistance model, AUROC decline of 0.10–0.25 over an 18-month location/time shift; CKD-MBD biomarker model, AUC 0.95→0.74 from training to prospective evaluation) [26–29].

3.9 Prospective and Real-World Evaluation

Prospective evaluation in actual or intended clinical settings as distinct from retrospective validation on a held-out split of already-collected data was identified in 14 studies (16%), and of these, only a small subset progressed to actual clinical deployment or silent/shadow implementation testing: the CURIAL-Rapide laboratory-free COVID-19 triage tool (independent external validation across three NHS trusts, prospective evaluation at Oxford, and live deployment at the John Radcliffe Hospital); the deep-learning HIV rapid-test reader (pilot field deployment across 60 fieldworkers in rural KwaZulu-Natal); the Uganda smartphone malaria DNA-diagnostic system (field-validated in Tororo district schools); the Romania national tuberculosis AI-CXR screening program (prospective national screening, 2019–2023); and a pediatric emergency-department sepsis antibiotic-acceleration model tested via silent implementation before live use. These five studies represent the clearest Level 5–6 evidence (practical clinical

utility/implementation) in the entire corpus a strikingly small fraction (about 6% of included studies) of an evidence base numbering in the dozens [18–20, 30, 31].

3.10 Explainability and Human-AI Interaction

Ten of 86 studies (12%) documented an explainability method (SHAP in the majority of cases, with LIME and inherently interpretable architectures decision trees, nomograms in a minority), a proportion notably lower than the prominence of explainability as a dedicated research theme might suggest (the literature search identified a considerable separate XAI-in-diagnostics literature, comprising review and meta-analysis papers about explainability methods in clinical decision support generally, rather than explainability embedded within disease-specific early-detection studies). Where explainability was documented, it was used predominantly for feature-importance identification and biomarker discovery (e.g., SHAP-identified CBC features for AMI classification, SHAP-identified minimal feature subsets for LMIC-appropriate CKD screening) rather than for prospectively tested clinician-facing decision support. No included study documented a formal evaluation of whether clinicians' trust, understanding, or decision-making changed as a result of an explanation, a gap directly relevant to DECIDE-AI's human-factors emphasis [32].

3.11 Clinical Actionability

Applying the seven-level evidence-actionability hierarchy proposed in this review across the 86 included studies: 13 studies (15%) sat at Level 1 (technical feasibility only), 29 (34%) at Level 2 (internal validity), 29 (34%) at Level 3 (external validity), 10 (12%) at Level 4 (prospective validity), 3 (3%) at Level 5 (showed practical clinical utility), and 2 (2%) at Level 6 (implementation in a routine workflow). No included study presented evidence corresponding to Level 7 (showed equity and scalability across resource-constrained settings as a formally evaluated outcome, as opposed to a stated aspiration or design intent). This distribution has the pyramid shape one would expect of a translationally immature field: attrition is steep and roughly monotonic from Level 1 through Level 6 and is presented visually in Figure 3. Framed against this review's central thesis, only 5 of 86 studies (about 6%) provide evidence that would support a claim of genuine clinical usefulness and actionability as operationally defined in this review (a showed change in clinical decision-making or workflow, evaluated prospectively or in deployment); the great majority of the literature, however methodologically sound within its own terms, indicates that an algorithm can detect or predict, not that a health system can safely act on that prediction.

3.12 Resource Requirements and Feasibility

Explicit, quantified resource-requirement reporting (compute, connectivity, hardware cost, technician training time) was uncommon across the included evidence, and no included study documented a formal benchmark of computational or hardware feasibility, such as on-device inference time on low-cost hardware, despite such benchmarking being readily achievable and directly relevant to resource-constrained settings. The more common pattern across the broader evidence base was implicit resource-appropriateness: studies using only routine, already-collected laboratory parameters (Section 3.3) are de facto more resource-feasible even where the authors do not frame the work in explicit feasibility terms set against a considerable minority of studies (multi-omic, whole-genome-sequencing, and specialized-proteomic approaches) whose resource requirements would, on their face, exclude most resource-limited laboratory settings for the foreseeable future regardless of analytical performance.

Table 7. Resource-limited implementation considerations identified across the included evidence.

Consideration	Finding	Illustrative study
Locally available data as design starting point	Most translationally mature LMIC evidence built around already-deployed diagnostic infrastructure (rapid tests, routine CBC, existing microscopy)	Uganda malaria; Zimbabwe Neotree sepsis
Edge/offline computation feasibility	Not directly tested by any included study; identified as a priority for future LMIC-oriented AI-diagnostics research	Evidence gap (Section 3.16, Table 8)
Minimal/zero-cost feature design	Explicit comparison of full vs. minimal feature subsets; minimal subset often near-equivalent	Bangladesh CKD screening; pre-eclampsia zero-cost screening

Consideration	Finding	Illustrative study
Test heterogeneity within "low-cost" category	Not all low-cost tests equivalent across phenotypes: some miss specific conditions entirely	Sickle cell/beta-thalassemia multi-test comparison (0% sensitivity for trait in some tests)
Local retraining needs	Cross-site/cross-time performance decay implies local recalibration is likely necessary, not optional	MALDI-TOF AMR study (authors' own conclusion)
Digital-infrastructure precursor requirement	Digitization of records/data flow often precedes any AI application	Neotree (Zimbabwe); digital pathology feasibility study (Brazil)

Note. Illustrative examples are drawn from the studies discussed in Section 3.12 and Discussion Section 4.9.

3.13 Equity and Population Generalizability

Direct empirical evidence of algorithmic bias or population-transferability limits within the disease-specific evidence base itself was found in the cross-country/cross-population validation studies already discussed (Sections 3.7–3.8): the Nigeria pneumonia CXR transferability failure and the Brazil-to-USA cardiovascular performance decline are the clearest disease-specific illustrations. Broader equity-methodology evidence, including landmark demonstrations that clinical algorithms can encode racial bias through biased proxy variables, was identified but treated as background/discussion literature rather than disease-specific primary evidence, since it addresses algorithmic bias as a general phenomenon rather than early-disease-detection specifically. The near-total absence of studies reporting demographic subgroup performance breakdowns (by ethnicity, sex, or socioeconomic status) within the disease-specific early-detection literature itself as opposed to the separate equity-methodology literature is a substantive evidence gap in its own right [27, 33, 34].

3.14 Cost and Health-System Considerations

Formal health-economic evaluation was documented in only a small number of included studies. The clearest example the China diabetic-retinopathy AI screening program economic evaluation across 251,535 participants provides direct empirical support for this review's central thesis at the health-economics level: the most diagnostically accurate AI configuration (sensitivity 93.3%, specificity 87.7%) was, in that specific analysis, also the most cost-saving configuration, but the same analysis showed that this need not hold in general several less-accurate configurations in the same modeled parameter space would have been more cost-saving under different assumptions, and several higher-sensitivity configurations gained additional quality-adjusted life-years only at considerably higher incremental cost. This is among the strongest pieces of evidence in the entire corpus that "most accurate" and "most cost-effective" are empirically separable questions requiring their own dedicated analysis, not a corollary automatically following from diagnostic performance [35].

3.15 Ethical, Regulatory and Governance Issues

Explicit engagement with ethical, regulatory, or governance dimensions (data privacy, consent, algorithmic accountability, regulatory pathway) within the primary evidence base itself was rare; where it appeared, it was almost always in the framing/discussion sections of papers describing field deployments (the CURIAL-Rapid and HIV rapid-test studies both discuss real-world deployment governance considerations) rather than as a formally evaluated study outcome. This mirrors the broader pattern already noted for explainability (Section 3.10) and equity (Section 3.13): ethics, governance, and explainability all exist as considerable, mature literatures in their own right (reflected in the WHO AI-ethics guidance, FUTURE-AI, and DECIDE-AI frameworks that inform this review's methodology) but are not yet routinely integrated as evaluated components within the disease-specific early-detection studies that constitute the bulk of the primary evidence base [20, 36, 37].

3.16 Evidence Gaps

Synthesizing across Sections 3.1–3.15, five evidence gaps are most consequential for the translational question this review set out to answer:

1. The external/prospective evaluation gap (Sections 3.8–3.9): the great majority of the available evidence (~55%) has not been tested beyond a single internal cohort, and even the minority that has been externally validated shows inconsistent, currently unpredictable transportability.

2. The multi-modality-evaluation gap (Section 3.5): multi-modal AI is uncommon (14% of studies) and rigorous single-vs-multi-modal comparison rarer still, despite multi-modal integration showing a consistent, if thinly evidenced, advantage where actually tested.

3. The explainability-integration gap (Section 3.10): explainability methods exist as a mature separate literature but are rarely embedded, and never prospectively human-factors-tested, within disease-specific early-detection studies.

4. The equity-and-subgroup-reporting gap (Section 3.13): near-total absence of demographic subgroup performance reporting within the disease-specific literature itself.

5. The cost-effectiveness gap (Section 3.14): formal economic evaluation is rare, and the one clear example in this corpus indicates that accuracy and cost-effectiveness are separable questions meaning the absence of economic evaluation elsewhere in the corpus is not a minor omission but leaves the single most policy-relevant question for resource-limited deployment largely unanswered.

Together, these gaps define the translational distance between "the algorithm can detect" and "the health system can safely and affordably act" that motivates the Discussion and the proposed Algorithm-to-Bedside Readiness Framework (Section 5 and Figure 3).

Table 8. Evidence gaps identified in this review and corresponding research priorities.

Evidence gap	Current state (this review)	Research priority
External/prospective evaluation	29% external, 16% prospective	Multi-site consortia; funding for multi-year prospective designs
Multimodal-vs-single-modality testing	14% multi-modal; direct comparison rarer still	Mandate single-modality baseline reporting in multi-modal-AI studies
Explainability-clinical-integration	12% report XAI; none test human-factors impact	Joint XAI + DECIDE-AI-style human-factors evaluation as standard design
Demographic subgroup/equity reporting	Near-absent in disease-specific studies	TRIPOD+AI/PROBAST+AI subgroup reporting as a submission requirement
Health-economic evaluation	1 clear example in full corpus	Mandatory cost-effectiveness component for LMIC-targeted AI-diagnostic proposals
Resource-requirement quantification	Quantified in no included study	Standard reporting of compute/connectivity/hardware assumptions
Workflow-integration evidence	2/86 studies	Silent/shadow-implementation phase as a standard pre-deployment step
Post-deployment monitoring	0/86 studies report this	Standards for AI-diagnostic post-market surveillance appropriate to LMIC regulatory capacity

Note. Research priorities follow directly from the gaps described in Section 3.16 and are elaborated further in Discussion Section 4.13.

The included studies were critically assessed using PROBAST+AI (n=76 prediction-model development and validation studies) and QUADAS-2 with PROBAST+AI elements (n=10 studies with an index-test-versus-reference-standard diagnostic-accuracy design, comprising MALDI-TOF antimicrobial-resistance prediction, malaria microscopy, HIV rapid-test image reading, tuberculosis chest-radiograph screening against GeneXpert confirmation, sickle cell disease test panels, and dengue complete-blood-count-based confirmation). Full per-study appraisal judgments are provided as supplementary material; this subsection presents the pattern-level synthesis across the corpus.

Overall, 9 studies (10%) were judged to be at high risk of bias, 41 (48%) raised some concerns, and 36 (42%) were judged low-to-moderate risk (Table 9). No study was judged unambiguously low risk across every domain; even studies in the low-to-moderate category typically had at least one domain most often calibration reporting or blinding of outcome assessment that could not be confirmed from the information available.

Validation status was the dominant determinant of appraisal outcome. External or independent validation was performed in 25 studies (29%), prospective or field validation in 14 (16%), internal validation only in 11 (13%), and validation approach was not clearly documented in the remaining 36 (42%). Collapsing the latter two categories, about 55% of included studies (47/86) presented no evidence beyond a single internal cohort or an unclearly documented validation strategy, a pattern consistent with the overlapping antimicrobial-resistance and cardiovascular-disease meta-analyses identified during screening, several of which similarly called for greater independent external validation.

Where independent external validation was performed, model performance frequently declined relative to internal or development-cohort estimates. Illustrative examples include a cardiovascular risk model developed in Brazil and externally validated in the United States (area under the curve 0.871 internally versus 0.786 externally), a MALDI-TOF antimicrobial-resistance model whose discriminative performance declined by 0.10–0.25 points over an 18-month period when applied to data from an unrelated location and time, a chronic kidney disease biomarker model whose performance declined from 0.95 in training to 0.74 in prospective evaluation, and a pediatric pneumonia chest-radiograph model whose accuracy declined considerably when a model trained in the United States was applied to Nigerian data [27–29, 33]. External validation did not uniformly result in declining performance: a cardiovascular risk model incorporating oculomics data and a nonalcoholic fatty liver disease model validated across the United Kingdom and China both showed well-preserved external performance. No single study characteristic reliably predicted which pattern a given model would follow, although models restricted to routine, low-cost, and widely standardized inputs appeared more likely to transport successfully than models relying on specialized assays or population-specific electronic health record features [26, 38].

Four recurring methodological concerns accounted for most high-risk-of-bias classifications. First, four studies documented near-perfect or "perfect" classification performance without independent external validation, including a childhood-diabetes transcriptomic classifier, a *Pseudomonas aeruginosa* antimicrobial-resistance model based on whole-genome sequencing, and two leukemia classification models; near-ceiling accuracy in a single retrospective cohort without independent external validation is treated throughout this review as a signal warranting caution rather than as an indicator of superior performance [39–42]. Second, one chronic kidney disease prognosis study used serum creatinine as both a predictor and, via estimated glomerular filtration rate, part of the diagnostic definition of the outcome being predicted a circularity explicitly identified and avoided by a separate, more methodologically robust study addressing the same condition in a low-resource setting] [5, 43]. Third, two studies presented very small discovery-stage samples (a sepsis biomarker study with a total sample of 33, and a stroke gene-signature study with a combined discovery and validation sample of 63) as though generalizable. Fourth, a small number of additional concerns, including a highly imbalanced case-control ratio in one discovery-stage stroke study, accounted for the remaining high-risk classifications [44–46].

Explainability methods (SHAP, LIME, or an inherently interpretable architecture) were documented in 10 of 86 studies (12%), a smaller proportion than the prominence of explainability as a research theme might suggest. Explainability methods identified during the literature search were concentrated in a separate body of methodological and review literature addressing clinical decision support generally, rather than embedded within the disease-specific early-detection studies that constitute the bulk of the available evidence reviewed here.

Calibration assessment and explicit handling of missing data, both components of the PROBAST+AI analysis domain, could not be systematically assessed across the included studies from the information available without full-text methods-section review of all 86 papers, and are addressed as a field-wide reporting gap in the Discussion rather than scored on a per-study basis.

Table 9. Summary of risk-of-bias judgments across the 86 included studies (PROBAST+AI / QUADAS-2).

Overall judgment	n	%	Dominant contributing factors
HIGH risk of bias	9	10%	Near-perfect/perfect performance claims without independent external validation (4); circular reasoning/information leakage (1); very small samples (2); imbalanced case/control ratio (2)
SOME CONCERNS	41	48%	Predominantly: validation strategy unclear or internal-only, with no other major red flag
LOW–MODERATE risk of bias	36	42%	External or prospective evaluation performed, no major additional concern flagged; calibration/missing-data/blinding reporting

Overall judgment	n	%	Dominant contributing factors
			could not be independently confirmed for any study in this category (a broader reporting gap discussed further below)

Note. Full per-study judgments, including the specific concerns contributing to each rating, are provided in the supplementary critical-appraisal data file.

4. DISCUSSION

4.1 Principal Findings

The evidence synthesis comprised 86 studies applying AI/ML to laboratory biomarker- and/or point-of-care-diagnostic-based early detection of disease, spanning infectious disease, sepsis, cardiovascular and metabolic disease, hematological malignancy, maternal-fetal health, and newborn screening. Reported analytical performance was, with striking consistency, high (AUROC/AUC frequently 0.80–0.99). Evidence supporting genuine clinical usefulness and actionability was not: only 29% of studies documented independent external validation, only 16% documented prospective or field validation, only 12% documented any explainability method, and only about 6% (5 of 86 studies) provided evidence corresponding to Level 5 or 6 of this review's evidence-actionability hierarchy showed practical clinical utility or real-world implementation. Multimodal AI, while theoretically motivated and empirically promising in the small number of studies that tested it directly, remained uncommon (14%). Formal cost-effectiveness evaluation was rarer still, and the one clear example identified showed that diagnostic accuracy and cost-effectiveness are empirically separable questions, not a single question answered by the same analysis.

4.2 Why High Accuracy Is Not Sufficient

The evidence synthesized here makes this review's central thesis difficult to avoid: a high AUROC in a retrospective, single-center, internally validated cohort is evidence of an association within that dataset, and nothing more, until independently showed otherwise. Several of the highest-performing models in this corpus (near-perfect or "perfect" classification claims in the childhood-diabetes transcriptomic study and two leukemia multi-omics/self-attention studies) carried no independent external validation at all the studies most likely to be cited for their impressive headline statistics are, on the appraisal conducted here, among the least well-supported in terms of generalizability. This is not a criticism of the individual investigators, most of whom operated within normal disciplinary conventions for early-stage AI method papers; it is a structural observation about what the current publication and evaluation ecosystem for diagnostic AI research rewards, and it is precisely the pattern this review's evidence hierarchy is designed to make visible rather than obscure behind an aggregate "AI shows promise" narrative.

4.3 The Validation Gap

Fifty-five percent of included studies presented no evidence beyond a single internal cohort. Where independent external validation was performed, the direction of the finding was informative but not uniform: considerable performance decay was observed in a cross-country cardiovascular model (Brazil to USA), in an antimicrobial-resistance MALDI-TOF model tested across an 18-month location/time shift, and in a chronic-kidney-disease biomarker model moving from training to prospective evaluation while a South Korean oculomics cardiovascular model and a UK-to-China nonalcoholic-fatty-liver-disease model both showed well-preserved external performance. No single study characteristic in this corpus reliably predicted which pattern a given model would follow, though a plausible signal did emerge: models restricted to routine, standardized, widely available inputs (basic CBC-derived indices, vital signs) appeared more likely to transport than models built on population-specific electronic-health-record features or specialized assays whose measurement characteristics may vary meaningfully across laboratories and populations. This observation should be treated as hypothesis-generating rather than established, given the small number of studies in this corpus that performed genuine cross-population independent external validation with enough methodological detail to compare directly.

4.4 The Multimodal Evidence Gap

Just 12 of 86 studies met this review's strict definition of multi-modal integration, and fewer still directly compared multi-modal against single-modality performance within the same study design. Every study in this corpus that did perform that direct comparison found an advantage for the multi-modal approach a striking, if statistically fragile, degree of consistency given the small sample of qualifying studies. This creates an asymmetric evidentiary situation: the theoretical case for multi-modal AI is strong, the empirical case where actually tested is also strong, but the number of studies willing or able to test it directly (as opposed to building a multi-modal system without a single-modality baseline for comparison) is small enough that the finding cannot yet be generalized with confidence. Given the infrastructure and data-integration burden multi-modal systems typically impose precisely the burden most acute in resource-constrained settings the absence of larger-scale, prospectively designed multi-modal-versus-single-modality trials represents one of the more consequential gaps identified in this review.

4.5 The Laboratory-to-Algorithm Interface

A pattern visible across several disease themes in this corpus concerns not the AI methodology itself but the upstream question of which laboratory or POC signal is fed into the model in the first place. The clearest illustration is the recurring circular-reasoning risk identified in chronic kidney disease prediction, where serum creatinine (and its transformation, eGFR) was used by one study as both a predictor and, implicitly, part of the diagnostic definition of the outcome being predicted a problem explicitly and directly addressed by a separate, better-designed CKD study in the same corpus, which deliberately constructed a "non-pathology-test" feature subset to avoid exactly this circularity. This is a laboratory-medicine-specific methodological hazard that a purely computer-science-oriented review of "AI in healthcare" would be less likely to surface, and it argues for laboratory professionals' direct involvement in AI model design not merely as data providers, but as co-designers who understand which input-outcome relationships are definitionally circular versus genuinely predictive.

4.6 Poc Diagnostics and AI in Resource-Limited Settings

The most mature evidence relevant to LMIC settings in this corpus the Uganda smartphone malaria DNA-diagnostic system, the South African HIV rapid-test image classifier, the Zimbabwe Neotree neonatal sepsis tool, the Bangladesh/South Asia CKD screening model, and the Nepal/Canada sickle cell disease multi-test comparison shares a common design feature: each was built around a data modality already deployable in the target setting (a rapid test, a routine blood draw, a low-cost microscope) rather than requiring new specialized infrastructure [5, 8, 9, 11, 18, 19]. This is an clinically relevant corrective to a design pattern visible elsewhere in the broader AI-diagnostics literature, in which a system is developed using whatever data modality is convenient in a well-resourced academic setting and LMIC applicability is asserted rather than designed for from the outset. The sickle cell disease comparative study offers a further, more specific lesson: not every "low-cost" test performs equivalently across all clinically relevant phenotypes several of the evaluated low-cost tests had zero sensitivity for beta-thalassemia trait despite excellent performance for severe sickle cell disease, a distinction with direct clinical consequences that a single aggregate "low-cost test works well" summary would have obscured.

4.7 Explainability and Clinical Trust

Explainability methods (predominantly SHAP) were applied in only 10 of 86 studies, and in every one of those cases, explainability was used as an analytical or biomarker-discovery tool for the investigators rather than as a prospectively tested clinician-facing decision-support feature. No study in this corpus documented measuring whether an explanation changed a clinician's trust, understanding, or behavior the human-factors question DECIDE-AI is specifically designed to evaluate. This suggests that "explainable AI" and "AI that clinicians can appropriately trust and act on" are currently two separate research programs rather than one, and that the considerable, mature XAI methods literature identified in this review's search (a body of review and meta-analysis papers about explainability techniques generally) has not yet been meaningfully integrated into the disease-specific early-detection literature that would need to adopt it for explainability to have any bearing on clinical usefulness and actionability.

4.8 Bias, Fairness and Population Transferability

Direct, disease-specific evidence of population-transferability limits was identified in this corpus (the Nigeria pneumonia chest-radiograph transferability failure when a US-trained model was applied to

Nigerian data; the Brazil-to-USA cardiovascular performance decline), but systematic demographic subgroup performance reporting (by ethnicity, sex, age strata, or socioeconomic status) was essentially absent from the disease-specific early-detection studies themselves [27, 33]. The broader algorithmic-fairness literature including landmark demonstrations that healthcare algorithms can encode racial bias through biased proxy variables provides part of the conceptual and methodological toolkit for addressing this gap, but that toolkit has evidently not yet been widely applied within the specific literature synthesized here. Given that resource-constrained settings are, almost by definition, the settings least represented in the training data of most AI systems developed to date, this absence of subgroup-level transferability evidence is a first-order equity concern, not a secondary methodological nicety [34].

4.9 Infrastructure and Affordability

Few studies quantified their resource requirements explicitly to the point of absence: no included study documented a formal benchmark of computational or hardware feasibility, of the kind that would directly test whether a proposed system could run on the infrastructure available in a resource-limited setting rather than merely using low-cost input data. The more common situation across the broader evidence base was implicit rather than evaluated resource-appropriateness: many studies used only routine, low-cost laboratory inputs without framing this choice explicitly as a feasibility design decision, while a considerable minority (multi-omic, whole-genome-sequencing, and specialized-proteomic studies) used approaches whose infrastructure requirements would exclude most resource-limited laboratory settings regardless of analytical merit. The absence of a distinguishing vocabulary in most primary studies between "this uses cheap inputs" and "we evaluated whether this is deployable" is itself a gap the field's reporting conventions should address.

4.10 Clinical Workflow Integration

Just two studies in this corpus (CURIAL-Rapide's deployment at the John Radcliffe Hospital, and the pediatric emergency-department sepsis model's silent-implementation phase) provided direct evidence of integration into a live clinical workflow, and even these represent single-site experiences rather than evidence of workflow integration replicated across settings. The remainder of the corpus, however strong its underlying analytical performance, provides no direct evidence regarding how a given model's output would be received, interpreted, or acted upon within the time pressures and existing decision-support ecosystem of a real clinical service. This is consistent with the broader validation-gap finding (Section 4.3) but represents a conceptually distinct evidence gap: workflow integration failure can occur even for a model with excellent, externally validated analytical performance, if its output arrives too late, in the wrong format, or without a clear associated clinical action precisely the actionability criteria this review's framework is built around [20, 31].

4.11 Regulatory and Ethical Challenges

Regulatory and governance issues were infrequently evaluated formally within the primary evidence base and, where present, was confined to the discussion sections of the few deployment-stage studies rather than treated as an evaluated study outcome. This is unsurprising given that most included studies represent pre-regulatory, developmental-stage research, but it does mean that this review's evidence base cannot speak with authority to how well the WHO's ethics and governance principles, or FUTURE-AI's fairness/universality/traceability/usability/robustness/explainability framework, are actually being operationalized in practice only that the disease-specific literature has not yet reached the stage of development at which such operationalization would typically be documented [36, 37].

4.12 From Prediction to Clinical Action

Evidence density falls sharply between algorithmic prediction and routine clinical use. The few studies reaching advanced actionability stages were designed as deployment or implementation programs rather than model-development studies ending with performance metrics. The main translational gap is therefore evidence that a model improves timely, safe, affordable, and equitable clinical action.

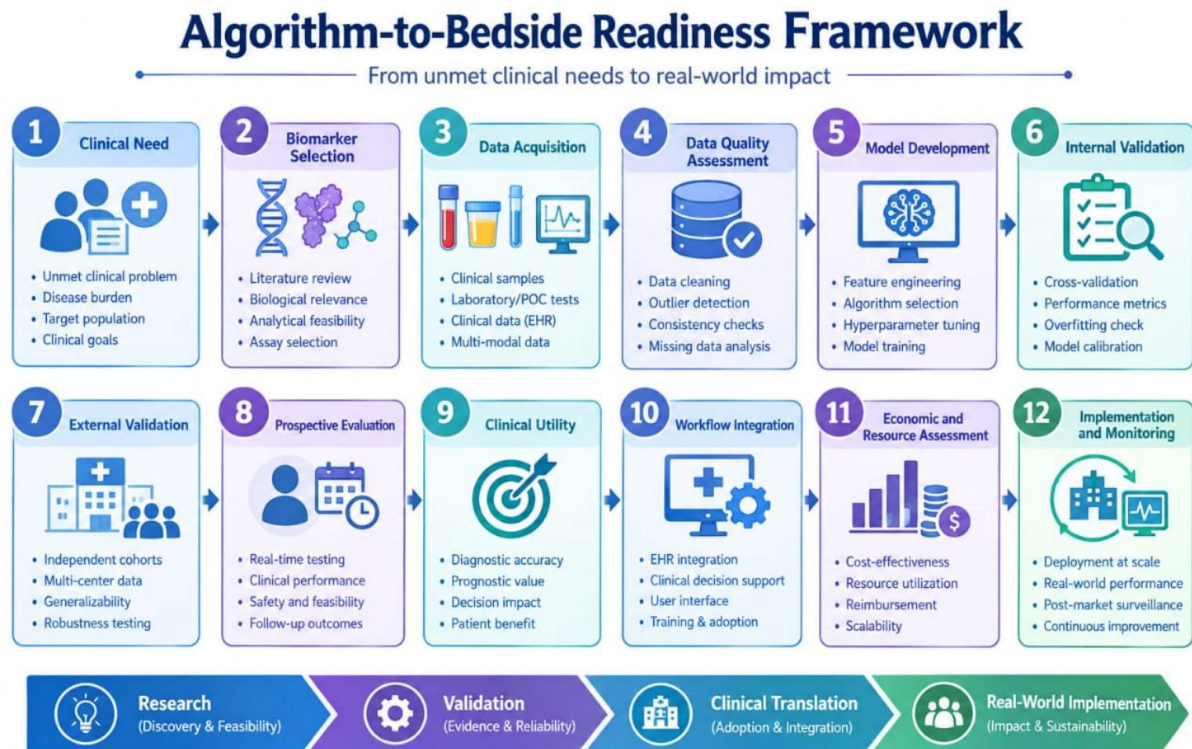


Figure 5. Algorithm-to-bedside readiness framework.

The framework progresses from clinical need and model development through validation, clinical utility, workflow integration, economic assessment, and monitored implementation..(ai generated image)

Table 10. Translation barriers identified between algorithm development and clinical deployment.

Barrier	Evidence from this review	Readiness-framework stage most affected
Absence of independent external validation	55% of studies never tested beyond one internal cohort	Stage 4
Unpredictable cross-population/cross-site performance decay	Brazil→USA CVD (0.871→0.786); MALDI-TOF AMR (−0.10 to −0.25 AUROC over 18 months); Nigeria pneumonia CXR (US-trained, Nigeria-tested)	Stage 4
Overfitting / near-ceiling performance without external check	4 of 10 HIGH-risk studies	Stages 2–3
Circular predictor-outcome design	CKD creatinine/eGFR case	Stage 2
Missing decision-impact evidence	>90% of corpus reports discrimination only, not decision change	Stage 6
Absent workflow-integration evidence	Just 2/86 studies (CURIAL-Rapide; pediatric ED sepsis)	Stage 7
Unquantified resource requirements	Explicit hardware/connectivity benchmarking found in no included study	Stage 8
Absent subgroup/equity reporting	Near-total absence in disease-specific studies	Stage 9
Rare economic evaluation	1 clear health-economic evaluation in full corpus (China DR screening)	– (cross-cutting)

4.13 Research Priorities

Four priorities emerge: prospective multi-site external validation; direct comparison of multimodal models with simpler single-modality baselines; integration of explainability with human-factors evaluation; and prospective health-economic assessment for systems intended for resource-limited deployment.

4.14 Proposed Implementation Pathway

The proposed Resource-Limited AI Diagnostic Implementation Pathway responds to the LMIC evidence: the most translationally mature systems were designed around locally available diagnostic infrastructure. External technologies should therefore be treated as hypotheses requiring local validation rather than solutions transferred unchanged.

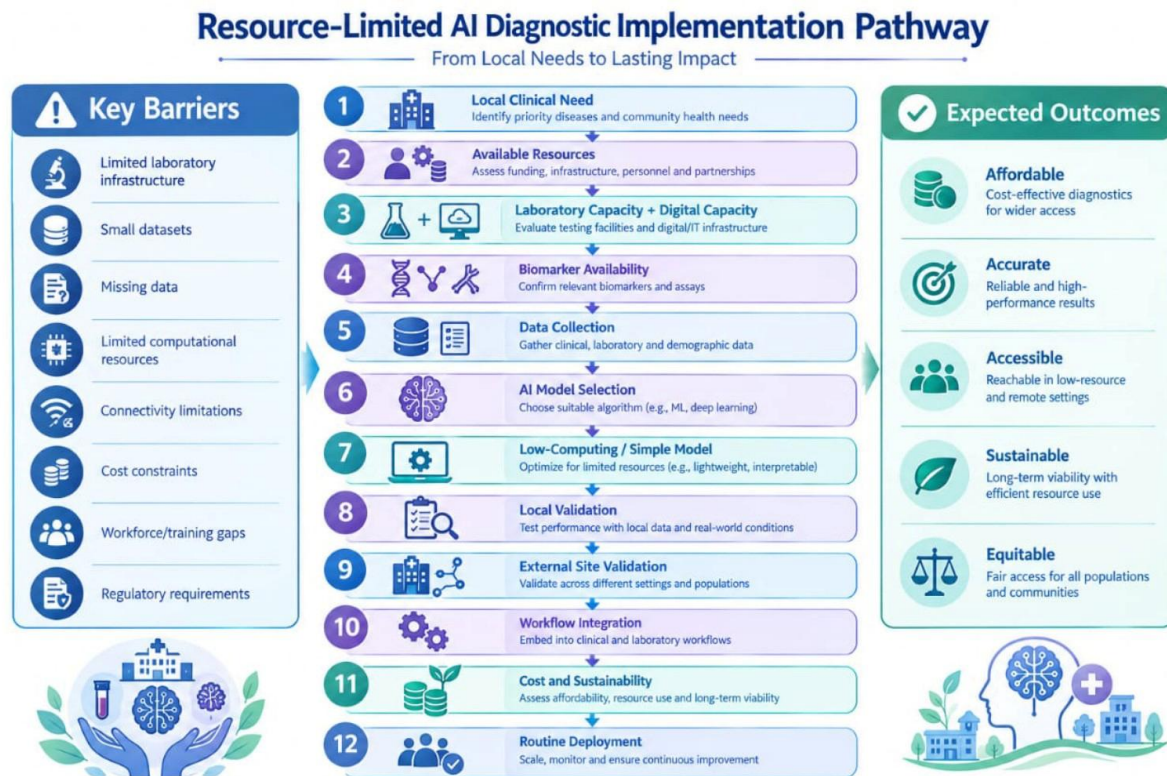


Figure 6. Resource-limited AI diagnostic implementation pathway.

The pathway emphasizes local need, available resources, data collection, model selection, local and external validation, workflow integration, sustainability, and routine deployment..(ai generated image)

4.15 Implications for Laboratories, Clinicians, Researchers, Policymakers and Technology Developers

For laboratories, AI development should involve laboratory professionals from the outset. Clinicians need external-validation and clinical-actionability evidence in addition to accuracy. Researchers should make prospective evaluation, validation, and economic assessment routine components. Health-system leaders should evaluate imported tools locally and support locally led validation and implementation.

Stage	Evidence required	Common failure observed in this review	Minimum benchmark	Research priority
1. Analytical feasibility	Model discriminates target in a dataset	Near-perfect performance claimed with tiny/unrepresentative samples (e.g., n=33 discovery-stage sepsis biomarker study [43])	Report on a held-out split at minimum	Pre-registration of feature set before outcome-labeling to prevent circular design
2. Model development	Documented feature selection, handling of class imbalance	Circular reasoning (predictor definitionally overlapping outcome, as in CKD-creatinine	Explicit check that predictors are not definitionally part of the outcome	Standardized checklist for laboratory-medicine-specific leakage risks

Stage	Evidence required	Common failure observed in this review	Minimum benchmark	Research priority
3. Internal validation	Cross-validation or held-out test set	case [43]) Reported without external or temporal check (55% of this corpus)	k-fold CV or temporal split, minimum	Journals/reviewers to require explicit internal-validation reporting
4. External validation	Independent-population or independent-site test	Performance decay undetected because never tested (Brazil→USA [27], MALDI-TOF [28] cases show decay IS detectable when tested)	At least one geographically or temporally independent test set	Multi-site consortia for shared external test sets by disease domain
5. Prospective clinical evaluation	Evaluation on prospectively, not retrospectively, collected data	Retrospective-only design presented as clinically validated	Formal prospective cohort design	Funding structures that support the multi-year timeline prospective studies require
6. Clinical utility	Demonstrated change in a clinical decision or outcome	Accuracy documented without any decision-impact measure (dominant pattern, >90% of corpus)	Decision-impact or time-to-action outcome measured	Trials measuring decision change, not only discrimination
7. Workflow integration	Evidence of function within routine clinical workflow	Just 2/86 studies in this corpus (CURIAL-Rapide, pediatric ED sepsis)	Silent/shadow implementation phase before live use	Standard adoption of silent-implementation testing as an intermediate step
8. Resource feasibility	Quantified compute/connectivity/hardware/training requirements	Not directly quantified by any included study; inferred only from input-data cost (Section 3.12)	Explicit statement of hardware/connectivity assumptions	Mandatory resource-requirement reporting in AI-diagnostics methods sections
9. Equity assessment	Subgroup performance reporting	Almost entirely absent in the disease-specific corpus (Section 3.13)	Performance stratified by at least one relevant demographic/geographic axis	Subgroup reporting as a standard TRIPOD+AI/PROBAST+AI expectation
10. Regulatory approval	Documented regulatory pathway/status	Not documented for any pre-deployment study in this corpus	Explicit regulatory-status statement	Clear LMIC-appropriate regulatory pathways for AI-based IVDs
11. Implementation	Deployment in routine service, beyond a pilot	Present in 2/86 studies (CURIAL-Rapide [20]; Romania TB screening at national-program scale [30])	Sustained use beyond the study period	Post-deployment funding models, not only development grants
12. Post-deployment monitoring	Ongoing performance/drift monitoring after deployment	Not documented for any study in this corpus	A defined monitoring/re-validation schedule	Standards for AI-diagnostic post-market surveillance in low-resource settings

5. CONCLUSIONS

AI applied to laboratory biomarkers and point-of-care (poc) diagnostics demonstrates substantial potential for early disease detection, including in resource-limited settings. however, translation from algorithmic performance to clinical actionability remains limited. among 86 included studies, only 25 (29%) reported

external or independent validation, 14 (16%) prospective or field validation, 10 (12%) explainability, and 5 (approximately 6%) evidence of practical clinical utility or implementation. no study formally evaluated equity and scalability across resource-constrained settings.

Multimodal integration was identified in 12 studies (14%), indicating emerging potential for combining complementary laboratory and poc information, although direct comparisons with single-modality approaches remained limited. in resource-limited settings, translational feasibility further depends on local diagnostic infrastructure, workforce, connectivity, computational requirements, affordability, and workflow compatibility.

The principal research priority is therefore to move beyond analytical performance toward externally validated, prospectively evaluated, clinically interpretable, and contextually deployable ai systems. the algorithm-to-bedside readiness framework and resource-limited ai diagnostic implementation pathway proposed in this review provide structured approaches for this transition. ultimately, clinically actionable diagnostic ai will require alignment of multimodal laboratory and poc data with robust validation, clinical workflow, health-system capacity, and resource-appropriate implementation.

6. DECLARATIONS

6.1 Funding

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

6.2 Competing Interests

The author declares no competing interests.

6.3 Ethics Approval and Consent to Participate

Not applicable. This is a scoping review of previously published literature; no original human or animal data were collected.

6.4 Consent for Publication

Not applicable.

6.5 Acknowledgments

Not applicable.

REFERENCES

- [1] Bedoya AD; Futoma J; Clement ME; Corey K; Brajer N; Lin A; Simons MG; Gao M; Nichols M; Balu S; Heller K; Sendak M; O'Brien C (2020). Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA Open* 3(2):252-260. 10.1093/jamiaopen/ooaa006
- [2] Ristori MV; Ruffini F; Spoto S; Cammarata R; La Vaccara V; Bani L; Caputo D; Soda P; Guarrasi V; Angeletti S (2026). Machine Learning Models for Sepsis: From Early Detection to Short- and Long-Term Prognosis. *International Journal of Molecular Sciences* 27(6):2721. PMC13026196
- [3] Nguyen RN; Lam HT; Phan HV; Bui NQ (2025). Machine Learning Nomogram for Predicting Dengue Shock Syndrome in Pediatric Patients With Dengue Fever in Vietnam. *Cureus*. 10.7759/cureus.81819 / PMID 40337565
- [4] Liu YE; Saul S; Rao AM; Robinson ML; Agudelo Rojas OL; Sanz AM; et al. (2022). An 8-gene machine learning model improves clinical prediction of severe dengue progression. *Genome Medicine*. 10.1186/s13073-022-01034-w
- [5] Kabir MA; Munira S; Azad DT; Ikram SM; Sarker MHR; Hanifi SMA (2026). Community-based early-stage chronic kidney disease screening using explainable machine learning for low-resource settings. *International Journal of Medical Informatics*. 10.1016/j.ijmedinf.2026.106435 / Vol 214 / PubMed 41946122
- [6] Hocaoglu M; Das J; Sawalha AH (2026). Identifying Systemic Lupus Erythematosus From Serum Proteomic Profiles Using Machine Learning and Genetic Risk Stratification. *Arthritis & Rheumatology*. 10.1002/art.70156
- [7] Amjad H; Hussain Z; Hasan M; Ul Hassan M (2025). Machine learning-based models for screening of anemia and leukemia using features of complete blood count reports. *Scientific Reports*. 10.1038/s41598-025-21279-w
- [8] Shrestha P; Lohse H; Bhatla C; McCartney H; Alzaki A; Sandhu N; Oli PK; Chaudhary S; Amid A; Onell R; Au N; Merkeley H; Kapoor V; Pande R; Stoeber B (2026). Morphology-based screening for sickle cell disease using

- low-cost automated microscopy and machine learning. *npj Digital Public Health* 1:25. 10.1038/s44482-026-00028-4
- [9] Shrestha P; Lohse H; Bhatla C; McCartney H; Alzaki A; Sandhu N; Oli PK; Chaudhary S; Amid A; Onell R; Au N; Merkeley H; Kapoor V; Pande R; Stoeber B (2025). Evaluation of low-cost techniques to detect sickle cell disease and beta-thalassemia: an open-label international multicentre study. *Lancet Regional Health - Southeast Asia* 35:100571. 10.1016/j.lansea.2025.100571 / PMID 40230447
 - [10] Master SR; Badrick TC; Bietenbeck A; Haymond S (2023). Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group. *Clinical Chemistry*. 10.1093/clinchem/hvad055 / Clin Chem 2023;69(7):690-698
 - [11] Lowther E; Khan N; Cortina-Borja M; Chimhini GL; Neal SR; Mangiza M; Fitzgerald F; Heys M; Chimbhuya S (2025). Predicting risk of early-onset sepsis in low-resource neonatal units using routine healthcare data: development and evaluation of multivariable statistical and machine learning models. *BMJ Paediatrics Open*. 10.1136/bmjpo-2025-003617
 - [12] Farquhar H (2026). AI-Enhanced Point-of-Care Diagnostics for Infectious Diseases in Resource-Limited Settings: A Scoping Review. *Tropical Medicine & International Health* 31(8):966-996. 10.1111/tmi.70170
 - [13] Tricco AC; Lillie E; Zarin W; O'Brien KK; Colquhoun H; Levac D; Moher D; Peters MDJ; Horsley T; Weeks L; Hempel S; Akl EA; Chang C; McGowan J; Stewart L; Hartling L; Aldcroft A; Wilson MG; Garritty C; Lewin S; Godfrey CM; Macdonald MT; Langlois EV; Soares-Weiser K; Moriarty J; Clifford T; Tuncalp O; Straus SE (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*. 10.7326/M18-0850 / PMID 30178033
 - [14] Collins GS; Dhiman P; Navarro CLA; Ma J; Hooft L; Reitsma JB; Logullo P; Beam AL; Peng LH; Van Calster B; van Smeden M; Riley RD; Moons KGM (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. PMC8273461 / PMID 34244270
 - [15] Whiting PF; Rutjes AWS; Westwood ME; Mallett S; Deeks JJ; Reitsma JB; et al. (QUADAS-2 Group) (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*. PMID 22007046 / 10.7326/0003-4819-155-8
 - [16] Peters MDJ; Godfrey CM; Khalil H; McInerney P; Parker D; Soares CB (2015). Guidance for conducting systematic scoping reviews / Scoping reviews chapter. *International Journal of Evidence-Based Healthcare* 13(3):141-146. PMID 26134548
 - [17] Arksey H; O'Malley L (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 8(1):19-32. 10.1080/1364557032000119616
 - [18] Guo X; Khalid MA; Domingos I; Michala AL; Adriko M; Rowell C; Ajambo D; Garrett A; Kar S; Yan X; Reboud J; Tukahebwa EM; Cooper J (2021). Smartphone-based DNA diagnostics for malaria detection using deep learning for local decision support and blockchain technology for security. *Nature Electronics* 4(8):615-624. 10.1038/s41928-021-00612-x
 - [19] Turbe V; Herbst C; Mngomezulu T; Meshkinfamfard S; Dlamini N; Mhlongo T; Smit T; Cherepanova V; Shimada K; Budd J; Arsenov N; Gray S; Pillay D; Herbst K; Shahmanesh M; McKendry RA (2021). Deep learning of HIV field-based rapid tests. *Nature Medicine*. 10.1038/s41591-021-01384-9 / Vol 27, pp.1165-1170
 - [20] Soltan AAS; Yang J; Pattanshetty R; Novak A; Yang Y; Rohanian O; Beer S; Soltan MA; Thickett DR; Fairhead R; Zhu T; Eyre DW; Clifton DA; et al. (2022). Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening. *Lancet Digital Health*. 10.1016/S2589-7500(21)00272-7 / PMID 35279399
 - [21] Pandey, J. K., Das, S., Vats, P., Dhabliya, D., Jaiswal, S., Manikandan, K., & Pandey, D. (2024). The implications of cloud computing, IoT, and wearable robotics for smart healthcare and agriculture solutions. In N. Sindhvani, R. Anand, A. George, & D. Pandey (Eds.), *Robotics and automation in Industry 4.0* (pp. 26–45). CRC Press. <https://doi.org/10.1201/9781003317456>
 - [22] Yang Q; Lu Y; Jiang X; Hu H; Lu X; Guo C (2026). Enhanced cardiovascular disease risk prediction using integrated machine learning models: a study from the UK Biobank cohort. *Open Heart* 13(1):e003827. 10.1136/openhrt-2025-003827 / PMID 41720602
 - [23] Kwon HJ; Park UH; Goh CJ; Park D; Lim YG; Lee IK; Do WJ; Lee KJ; Kim H; Yun SY; Joo J; Min NY; Lee S; Um SW; Lee MS (2023). Enhancing Lung Cancer Classification through Integration of Liquid Biopsy Multi-Omics Data with Machine Learning Techniques. *Cancers* 15(18):4556. 10.3390/cancers15184556
 - [24] Pandey, J. K., Verma, S. K., Kumar, J., Perwej, Y., Jha, S. K., Panchal, B. Y., Ferdouse, R., Sindhu, V., Banerjee, S., Tiwari, M., Badal, R., Mandal, P., & Baghel, J. S. (2026). Transformative role of advanced neural computation in clinical image diagnostics: A review of key concepts and applications. *Seminars in Ultrasound, CT and MRI*. Advance online publication. <https://doi.org/10.1053/j.sult.2026.06.010>

- [25] Du R; Tsougenis ED; Ho JWK; Chan JKY; Chiu KWH; Fang BXH; Ng MY; Leung ST; Lo CSY; Wong HYF; Lam HYS; Chiu LFJ; So TY; Wong KT; Wong YCI; Yu K; Yeung YC; Chik T; Pang JWK; Wai AKC; Kuo MD; Lam TPW; Khong PL; Cheung NT; Vardhanabhuti V (2021). Machine learning application for the prediction of SARS-CoV-2 infection using blood tests and chest radiograph. *Scientific Reports* 11:14250. 10.1038/s41598-021-93719-2 / PMID 34244563
- [26] Tong C; Luo N; Zhang K; et al. (2024). Machine-learning-based models to predict cardiovascular risk using oculomics and clinic variables in KNHANES. *BioData Mining* 17:11. 10.1186/s13040-024-00363-3
- [27] Shimizu GY; Schrempf M; Romao EA; Jauk S; Kramer D; Rainer PP; Cardeal da Costa JA; de Azevedo-Marques JM; Scarpelini S; Suzuki KMF; Cesar HV; de Azevedo-Marques PM (2024). Machine learning-based risk prediction for major adverse cardiovascular events in a Brazilian hospital: Development external validation and interpretability. *PLOS ONE* 19(10):e0311719. 10.1371/journal.pone.0311719 / PMID 39392843
- [28] Wiesmann N; Enders D; Westendorf A; Koch R; Schaumburg F (2025). Prediction of antimicrobial resistance from MALDI-TOF mass spectra using machine learning: a validation study. *Journal of Clinical Microbiology*. 10.1128/jcm.01186-25
- [29] Li Y; Lou Y; Liu M; Chen S; Tan P; Li X; Sun H; Kong W; Zhang S; Shao X (2024). Machine learning based biomarker discovery for chronic kidney disease-mineral and bone disorder (CKD-MBD). *BMC Medical Informatics and Decision Making* 24:37. 10.1186/s12911-024-02421-6
- [30] Mahler B; Stoichita A; Baiceanu D; Panciu TC; Dendrino D; Mihai M; Bobocea R; Ibraim E; Balteanu M; Popescu O; Burecu MO; Mosteanu IM; Veronese V; Matache R; Munteanu I; Popa C; Dragomir JA (2025). Artificial intelligence: a useful tool in active tuberculosis screening among vulnerable groups in Romania - advantages and limitations. *Frontiers in Public Health* 13:1433450. 10.3389/fpubh.2025.1433450 / PMID 39991692
- [31] Cao K; Braykov N; McCarter A; Kandaswamy S; Orenstein EW; Ray E; Carter R; Gleeson MB; et al. (2025). Development and Validation of an Artificial Intelligence Predictive Model to Accelerate Antibiotic Therapy for Critically Ill Children with Sepsis in the Pediatric ED with Pediatric ICU Disposition. *medRxiv* (preprint). 10.1101/2025.03.25.25324127 / PMID 40196256
- [32] DECIDE-AI Steering Group (Delphi consensus - multistakeholder) (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*. PMID 35584845 / naturemedicine 28:924-933
- [33] Togunwa TO; Babatunde AO; Fatade OE; Olatunji R; Ogbole G; Falade A; Ghasemi H (2025). Detection of pneumonia in children through chest radiographs using artificial intelligence in a low-resource setting: A pilot study. *PLOS Digital Health*. 10.1371/journal.pdig.0000713
- [34] Obermeyer Z; Powers B; Vogeli C; Mullainathan S (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 366(6464):447-453
- [35] Wang YX; Liu K; Hu YY; Luo L; Shi D; et al. (2024). Economic evaluation for medical artificial intelligence: accuracy vs. cost-effectiveness in a diabetic retinopathy screening case. *npj Digital Medicine* 7:43. 10.1038/s41746-024-01032-9
- [36] World Health Organization (2021). Ethics and governance of artificial intelligence for health: WHO guidance. WHO (Geneva). ISBN 978-92-4-002920-0 (electronic) / 978-92-4-002921-7 (print)
- [37] Lekadir K; Frangi AF; Porras AR; Glocker B; Cintas C; Langlotz CP; et al. (117-118 consortium members) (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 10.1136/bmj-2024-081554
- [38] Xiao L; Zeng L; Wang J; Hong C; Zhang Z; Wu C; Cui H; Li Y; Li R; Liang S; Deng Q; Li W; Zou X; Ma P; Liu L (2025). Development and Validation of Machine Learning-Based Marker for Early Detection and Prognosis Stratification of Nonalcoholic Fatty Liver Disease. *Advanced Science*. 10.1002/adv.202410527
- [39] Huang X; Ouyang D; Xie W; Zhuang H; Gao S; Liu P; Guo L (2025). Development and validation of machine learning-based diagnostic models using blood transcriptomics for early childhood diabetes prediction. *Frontiers in Medicine*. 10.3389/fmed.2025.1636214
- [40] Noman SM; Shafiq M; Yuan Y; Zeng M; Xie Q; Li X; Zeeshan M; Arshad J; Amentie MD; Jiao X (2023). Machine Learning Techniques for Antimicrobial Resistance Prediction of *Pseudomonas Aeruginosa* from Whole Genome Sequence Data. *Computational Intelligence and Neuroscience* 2023:5236168. 10.1155/2023/5236168
- [41] Maruf Md.; Haque Md. Mahbulul; Paul B (2025). Deep Learning with Self-Attention and Enhanced Preprocessing for Precise Diagnosis of Acute Lymphoblastic Leukemia from Bone Marrow Smears in Hemato-Oncology. *arXiv* (preprint). arxiv.org/abs/2508.17216
- [42] Abbasi EY; Deng Z; Ali Q; Khan A; Shaikh A; Al Reshan MS; Sulaiman A; Alshahrani H (2024). A machine learning and deep learning-based integrated multi-omics technique for leukemia prediction. *Heliyon*. 10.1016/j.heliyon.2024.e25369

- [43] Dubey Y; Mange P; Barapatre Y; Sable B; Palsodkar P; Umate R (2023). Unlocking Precision Medicine for Prognosis of Chronic Kidney Disease Using Machine Learning. *Diagnostics* 13(19):3151. 10.3390/diagnostics13193151
- [44] Li H; Jiang S; Chen Z; Yao Y; Chen M; Hu Y (2025). Identification and validation of core biomarkers for sepsis: a comprehensive analysis using bioinformatics and machine learning. *Frontiers in Immunology* 16:1700704. 10.3389/fimmu.2025.1700704
- [45] O'Connell GC; Petrone AB; Treadway MB; Tennant CS; Lucke-Wold N; Chantler PD; Barr TL (2016). Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischaemic stroke. *npj Genomic Medicine* 1:16038. 10.1038/npjgenmed.2016.38 / PMID 29263821
- [46] Zhang J; Yuan T; Wei S; Feng Z; Li B; Huang H (2022). New strategy for clinical etiologic diagnosis of acute ischemic stroke and blood biomarker discovery based on machine learning. *RSC Advances* 12(23):14716-14723. 10.1039/d2ra02022j