

## ORIGINAL RESEARCH ARTICLE

# The Role of Artificial Intelligence in Early Clinical Diagnosis: A Systematic Literature Review of Current Challenges and Future Directions

Parhlad Singh Ahluwalia<sup>1\*</sup><sup>1</sup> Dr. Bhim Rao Ambedkar Law University, Jaipur, Rajasthan, India\*Corresponding Author: Parhlad Singh Ahluwalia | Email: [ahluwalia002@zohomail.in](mailto:ahluwalia002@zohomail.in) | ORCID: 0000-0002-6086-8616

## ABSTRACT

**Background:** Diagnostic error remains a leading and preventable source of patient harm, and delayed detection of cancer, infection, cardiovascular disease and sensory disorders drives much avoidable morbidity. Artificial intelligence (AI), including deep learning, machine learning and, more recently, large language models, is increasingly proposed as a means of detecting disease earlier and more reliably.

**Objective:** To synthesise the evidence on AI in early clinical diagnosis, to appraise the secondary literature that has already reviewed the field, and to identify the technical, clinical, ethical, regulatory, health-system and equity challenges that separate laboratory performance from clinical benefit.

**Methods:** A structured, PRISMA 2020-guided review was conducted. The review covered primary diagnostic accuracy studies, prospective evaluations, randomised trials, systematic reviews, meta-analyses, regulatory analyses and policy documents. Methodological quality was interpreted through the QUADAS-2, QUADAS-AI and PROBAST frameworks. Findings were organised by clinical domain and by analytical perspective.

**Results:** In retrospective, image-based tasks (retinal, dermatological, mammographic, thoracic, histopathological and electrocardiographic), AI systems frequently match or exceed clinicians. Evidence of real-world clinical benefit is much thinner. Related reviews repeatedly report high heterogeneity, high risk of bias, scarce external validation and few prospective or randomised studies. Performance degrades across sites, devices and populations. Documented algorithmic bias, automation bias, limited explainability and unresolved regulatory and liability questions further constrain adoption. Evidence from low- and middle-income settings, including India, suggests that infrastructure, data governance and workforce readiness shape the benefit of AI as much as algorithmic accuracy does.

**Conclusion:** The central barrier is the translational gap between statistical accuracy and demonstrated patient benefit. Closing it requires prospective multi-site validation, standardised reporting, subgroup-level fairness auditing, human-factors-informed deployment, federated and privacy-preserving data collaboration, and governance frameworks designed for adaptive software.

**Keywords:** artificial intelligence; early diagnosis; deep learning; diagnostic accuracy; algorithmic bias; explainability; clinical decision support; PRISMA; health equity

## 1. INTRODUCTION

### 1.1 The diagnostic problem

Diagnosis determines whether a patient receives the right treatment at the right time, so its quality shapes nearly every downstream outcome in medicine. A landmark consensus report concluded that most people will experience at least one diagnostic error in their lifetime, sometimes with devastating consequences [1]. Estimates from three large observational studies of United States adults suggest that roughly one in twenty outpatients experiences a diagnostic error each year, and that about half of these errors are potentially harmful [2]. Analyses of malpractice claims indicate that vascular events, infections and cancers account for around three quarters of serious misdiagnosis-related harm [3]. These are the conditions for which early detection changes prognosis most.

The causes are rarely a simple lack of knowledge. In a classic study of internal medicine cases, cognitive factors contributed to nearly three quarters of diagnostic errors and system-related factors to roughly two thirds, and the two frequently coexisted [4]. Heuristics such as anchoring, premature closure and availability bias help clinicians cope with uncertainty, but they also generate systematic mistakes [5]. A systematic review of cognitive biases in medical decision-making confirmed that such biases are widespread and associated with diagnostic inaccuracy [6]. Fatigue, time pressure, fragmented records and uneven access to specialists compound them. A technology that offers consistent attention, scalable pattern recognition and integration of high-dimensional data therefore addresses a real problem rather than a hypothetical one.

## 1.2 The rise of AI in diagnosis

Over the past decade, advances in deep learning, computing power and digitised health data have moved AI from a research curiosity to a candidate clinical tool. Topol described a convergence of human and artificial intelligence in which algorithms augment clinicians across imaging, pathology, physiological signals and electronic records [7]. Rajkomar and colleagues characterised machine learning as a general-purpose capability for extracting predictive structure from clinical data, one that raises distinctive challenges of validation, interpretability and fairness [8]. Reviews of the field have mapped applications from image interpretation to risk prediction [9,10]. Within the regional literature considered here, the role of AI in diagnostic medicine has also been examined as a component of broader healthcare transformation [11].

Several landmark studies established the credibility of the approach. A deep neural network detected referable diabetic retinopathy in retinal photographs with sensitivity and specificity above 90% at its chosen operating points [12]. A convolutional network classified skin lesions at a level comparable to board-certified dermatologists [13]. An international evaluation of a breast-screening system reported reductions in both false-positive and false-negative reads relative to radiologists in two national datasets [14]. These results carried an implicit promise: if algorithms can find disease earlier and more consistently than people, the burden of delayed diagnosis might be substantially reduced.

## 1.3 The translational gap

The promise has not yet been matched by uniform evidence of clinical benefit. A systematic review of studies comparing deep learning with clinicians found that very few were prospective, that few were tested in real clinical settings, and that many carried a high risk of bias and made claims stronger than their designs supported [15]. A meta-analysis of deep learning in medical imaging found diagnostic accuracy broadly equivalent to that of healthcare professionals, but only a small fraction of studies compared the two groups on the same data [16]. Kelly and colleagues argued that clinical impact depends on far more than accuracy. It also requires generalisability, workflow integration, trust, regulatory clearance and evidence of effect on outcomes [17].

## 1.4 Context and rationale

The question matters most where diagnostic capacity is scarcest. India, for example, faces a double burden of persistent infectious disease and rapidly rising non-communicable disease [18,19], together with marked regional and socioeconomic disparities in access to care, including in states such as Punjab [20]. AI could extend scarce specialist expertise, but it could equally entrench inequity if it is trained on unrepresentative data or deployed without infrastructure. A review that considers technical performance and system context together is therefore warranted.

## 1.5 Objectives and contribution

This review has four objectives. First, it synthesises the evidence on AI performance in early diagnosis across major clinical domains. Second, it appraises and compares the existing secondary literature, a "review of reviews". Third, it analyses the field from complementary technical, clinical, ethical, regulatory, health-system and equity perspectives. Fourth, it derives a set of analytical propositions and a research agenda. The contribution is an integrated reading of why high accuracy so often fails to translate into benefit, and what would change that.

## 2. CONCEPTUAL BACKGROUND

### 2.1 Defining "early" diagnosis

"Early diagnosis" covers three related activities. **Screening** detects disease in asymptomatic populations, as in mammography or retinal examination. **Early case detection** identifies disease at the first symptomatic stage and shortens the diagnostic interval. **Pre-clinical prediction** flags imminent deterioration before overt clinical signs, as in sepsis or acute kidney injury alerts. AI contributes differently to each. In screening it acts as a second reader or triage filter, in case detection as a decision-support aid, and in prediction as a continuous monitor.

### 2.2 Families of AI methods

Classical machine learning relies on hand-crafted features and statistical classifiers. Deep learning, especially convolutional networks, learns features directly from raw images and signals and now dominates medical image analysis [9]. Comprehensive surveys describe its application across radiology, pathology and ophthalmology [10]. Sequence and tabular models process electronic health records and physiological time series. Large language models (LLMs) and multimodal foundation models are the newest family. They generate free text and can in principle integrate images, notes and laboratory values [10]. Each family has distinct failure modes, so evidence cannot be pooled indiscriminately across them.

### 2.3 Why accuracy is not benefit

A diagnostic accuracy study asks whether a test classifies correctly. A clinical-utility study asks whether using the test improves patient outcomes. Between the two lie several conditions: the test must be used at the right point in a pathway, its results must be believed and acted on, the subsequent care must be effective, and overdiagnosis and false reassurance must not offset gains. Much of the AI literature remains at the first step. This distinction organises the analysis that follows.

## 3. METHODS

### 3.1 Design and reporting framework

This review follows the PRISMA 2020 statement, which updates the original 2009 guidance and specifies reporting of eligibility criteria, information sources, selection, appraisal and synthesis [21,22]. Because the field spans heterogeneous designs, from diagnostic accuracy studies to policy analyses, elements of the PRISMA extension for scoping reviews (PRISMA-ScR) were also used to map the breadth of the evidence [23]. Practical guides on applying PRISMA-ScR to emerging-researcher projects informed the charting approach [24,25]. The overall design is therefore a PRISMA-guided structured review with narrative synthesis. Quantitative pooling was not attempted, because the included studies differ in disease, modality, reference standard and metric, and pooling across them would produce misleading summary estimates.

### 3.2 Search strategy and eligibility (protocol)

Table 1 summarises the search architecture and eligibility framework. The strategy is designed to be reproducible across PubMed/MEDLINE, Embase, Scopus, Web of Science, IEEE Xplore and the Cochrane Library, supplemented by regulatory and policy repositories and reference-list chasing.

**Table 1. Review protocol: search architecture, eligibility and appraisal framework**

| Element                 | Specification                                                                                             |
|-------------------------|-----------------------------------------------------------------------------------------------------------|
| Population              | Patients or populations undergoing screening, triage or diagnostic evaluation in any care setting         |
| Intervention/index test | AI-based diagnostic or early-detection systems (machine learning, deep learning, LLMs, multimodal models) |
| Comparator              | Clinician performance, conventional tests, standard care, or no decision support                          |

|                     |                                                                                                                                                                                                                                           |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Outcomes            | Sensitivity, specificity, AUC; workload; time to diagnosis; patient outcomes; bias, safety, explainability, implementation and equity outcomes                                                                                            |
| Study designs       | Diagnostic accuracy studies, prospective evaluations, RCTs, systematic reviews and meta-analyses, scoping reviews, regulatory analyses, policy and ethics papers                                                                          |
| Core concept blocks | ("artificial intelligence" OR "machine learning" OR "deep learning" OR "large language model") AND (diagnosis OR "early detection" OR screening OR triage) AND (validation OR accuracy OR bias OR implementation OR ethics OR regulation) |
| Time window         | 2015 to the present, with earlier foundational sources on diagnostic error                                                                                                                                                                |
| Language            | English                                                                                                                                                                                                                                   |
| Exclusions          | Non-clinical AI; purely technical papers with no clinical evaluation; conference abstracts without full data; duplicate publications                                                                                                      |
| Quality appraisal   | QUADAS-2 and QUADAS-AI for diagnostic accuracy; PROBAST for prediction models; reporting-completeness checks against TRIPOD, TRIPOD+AI, CONSORT-AI and DECIDE-AI                                                                          |
| Synthesis           | Narrative, domain-wise and perspective-wise; no meta-analytic pooling                                                                                                                                                                     |

### 3.3 Selection process

The PRISMA flow has four stages. **Identification** gathers records from databases and supplementary sources. **Screening** removes duplicates and screens titles and abstracts against the criteria in Table 1. **Eligibility** assesses full texts. **Inclusion** defines the final evidence set. In the intended workflow, two reviewers screen independently and resolve disagreements by consensus. No numeric flow counts are reported here, because they must come from an actual database export. The evidence base discussed below consists of landmark primary studies, major systematic reviews and meta-analyses, regulatory analyses and policy documents. It was purposively assembled to represent each clinical domain and each analytical perspective, and it makes no claim to exhaustiveness.

### 3.4 Quality appraisal

Methodological quality was interpreted through established tools. QUADAS-2 evaluates risk of bias and applicability in diagnostic accuracy studies across patient selection, index test, reference standard and flow and timing [26]. QUADAS-AI adapts this framework to AI-centred studies [28]. PROBAST assesses bias in prediction-model studies [27]. Users' guides for reading machine learning articles supplied additional heuristics for judging dataset provenance, validation and clinical relevance [29]. Reporting completeness was judged against TRIPOD for prediction models [30], its AI-specific update TRIPOD+AI [31], CONSORT-AI for trials [32] and DECIDE-AI for early-stage clinical evaluation [33].

### 3.5 Synthesis, ethics and originality

Data were charted by domain, design, comparator, headline metric and principal limitation. Systematic review and meta-analysis methods used in other health domains served as methodological analogues for structuring synthesis [34]. Only published and publicly available materials were used, and the review followed accepted principles of research ethics and transparent reporting [35]. Textual originality was supported by paraphrase-based drafting and similarity-screening practices described in the plagiarism-detection literature [36,37].

## 4. RESULTS

### 4.1 The evidence landscape

Three structural features of the evidence base recur across the literature. The first is a design imbalance. A meta-research survey of studies reporting AI performance on medical images found that only a small proportion (about 6%) performed external validation, and that few used prospective designs [38]. A systematic comparison of AI with clinicians in disease diagnosis likewise concluded that apparent equivalence rested on studies of variable methodological quality [39]. A meta-analysis of deep learning in imaging found very high between-study heterogeneity, which limits the interpretability of pooled accuracy [40].

The second feature is regulatory visibility. Comparative analyses of AI-based devices cleared in the United States and Europe show a steady growth in approvals, concentrated in radiology, but also uneven transparency about

the clinical evidence supporting clearance [41]. A public database of FDA-cleared AI-based devices documented the same concentration and the limited availability of performance data [42].

The third feature is the dominance of retrospective, single-modality, image-centric work. Prospective evaluations, outcome trials and studies of equity and implementation are far less common. These patterns frame the interpretation of every domain-specific finding below.

#### 4.2 Related research reviews: their results and outputs

A review of this field must situate itself against the secondary literature. Table 2 compares the principal related reviews. Their conclusions are strikingly consistent. Accuracy is high in retrospective testing, evidence quality is low, and generalisability and implementation remain under-examined.

**Table 2. Related reviews, their outputs and analytical implications**

| Ref. | Review focus                                                                  | Principal results/outputs                                                                                                                                           | Key limitation identified                                                | Implication for this review                             |
|------|-------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------|
| [16] | Deep learning vs. clinicians in imaging (systematic review and meta-analysis) | Pooled accuracy of deep learning broadly equivalent to clinicians (e.g., sensitivity about 87% vs. 86%; specificity about 93% vs. 91% in like-for-like comparisons) | Few studies compared AI and clinicians on identical data; poor reporting | Equivalence claims need like-for-like designs           |
| [15] | AI vs. clinicians: design, reporting, claims                                  | Few prospective studies and few real-world tests; high risk of bias common; claims often overstated                                                                 | Scarcity of randomised evidence                                          | Treat headline claims with caution                      |
| [38] | Design characteristics of imaging AI studies                                  | Only a small share used external validation                                                                                                                         | Narrow validation                                                        | Generalisability is the exposed weak point              |
| [40] | Diagnostic accuracy of deep learning in imaging (meta-analysis)               | Good pooled accuracy in several specialties; very high heterogeneity                                                                                                | Heterogeneity limits clinical interpretation                             | Pooled estimates are context-dependent                  |
| [39] | AI vs. clinicians in disease diagnosis                                        | Comparable performance, with concerns about quality and comparability                                                                                               | Variable methodology                                                     | Standardised comparison protocols needed                |
| [43] | Machine learning for COVID-19 on chest imaging                                | None of the reviewed models judged of potential clinical use due to methodological flaws                                                                            | Bias, duplication, poor reporting                                        | Crisis-driven research can mass-produce unusable models |
| [44] | Prediction models for COVID-19 diagnosis/prognosis                            | Nearly all models at high or unclear risk of bias                                                                                                                   | Weak validation and calibration                                          | Peer review alone does not guarantee model validity     |
| [17] | Key challenges for clinical impact                                            | Impact depends on generalisation, workflow, trust, regulation and evidence of benefit                                                                               | Conceptual rather than empirical                                         | Provides the framework for Section 5                    |
| [45] | AI and patient safety (scoping review)                                        | AI offers safety opportunities, with risks from error and over-reliance                                                                                             | Heterogeneous evidence                                                   | Safety must be studied as a system property             |
| [46] | AI for healthcare in low- and middle-income countries (scoping review)        | Many applications at early or pilot stages; limited real-world evaluation                                                                                           | Sparse implementation data                                               | Evidence from the Global South is thin                  |
| [47] | Large language models in medicine                                             | Broad potential in documentation and decision                                                                                                                       | Rapidly changing field                                                   | LLM evidence is provisional                             |

|      |                                                                   |                                                                                      |                                             |                                                  |
|------|-------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------|--------------------------------------------------|
|      |                                                                   | support; concerns over accuracy, bias and privacy                                    |                                             |                                                  |
| [48] | Deep learning for healthcare: opportunities and challenges        | Identified data heterogeneity, interpretability and privacy as barriers              | Largely technical                           | Barriers are cross-cutting                       |
| [49] | Deep learning in medical image analysis (survey)                  | Mapped task types and architectures across imaging fields                            | Performance reported mostly retrospectively | Confirms imaging as the maturest domain          |
| [50] | AI in cancer imaging                                              | Described clinical applications and challenges, including validation and integration | Limited prospective data                    | Oncology is a high-value, high-risk application  |
| [51] | AI in radiology                                                   | Outlined the opportunities for detection, characterisation and monitoring            | Translational barriers                      | Workflow integration is decisive                 |
| [52] | Digital transformation in hospital management (systematic review) | Examined organisational enablers and barriers of digital health adoption             | Context-specific evidence                   | Adoption depends on management as well as models |
| [11] | AI in diagnostic medicine (regional review)                       | Discussed the diagnostic role of AI within healthcare transformation                 | Conceptual breadth                          | Supports the Global South framing                |
| [53] | AI-driven prediction of antimicrobial resistance                  | Considered machine learning with genomic epidemiology for stewardship                | Emerging evidence                           | Illustrates diagnosis beyond imaging             |

*Reading the table.* Four cross-cutting conclusions emerge. First, systematic reviews agree that accuracy is promising but under-validated [15,16,38,40]. Second, crisis-era reviews expose a failure mode in which large volumes of research are produced without clinical usability [43,44]. Third, conceptual reviews relocate the problem from the algorithm to the system [17,45]. Fourth, evidence from lower-resource settings is sparse and largely early-stage [46]. The current review adds domain-level comparison and a cross-perspective analysis that most prior reviews treat separately.

### 4.3 Domain-wise evidence

**4.3.1 Ophthalmology.** Diabetic retinopathy screening is the most mature application. The early validation of a deep learning algorithm on retinal photographs reported high sensitivity and specificity in two independent datasets [12]. A multiethnic validation reported sensitivity of about 90% and specificity of about 92% for referable disease and extended detection to related eye conditions [54]. An algorithm trained on optical coherence tomography scans produced referral recommendations with an error rate of about 5.5%, matching or exceeding expert performance [55]. A pivotal prospective trial in primary care led to the first autonomous AI diagnostic system authorised for this indication, with sensitivity of about 87% and specificity of about 91% [56]. Validation in India reported sensitivity of roughly 89 to 92% and specificity of roughly 92 to 95% against manual grading across two sites [57]. This domain also exposes the deployment gap. A human-centred study in Thai clinics found that poor image quality and workflow friction caused a considerable share of images (roughly one fifth) to be ungradable, frustrating nurses and patients despite strong laboratory performance [58]. An Australian primary-care evaluation showed feasibility of AI-assisted screening outside specialist settings [59].

**4.3.2 Dermatology.** The 2017 skin-lesion classifier was trained on over 129,000 clinical images spanning more than 2,000 diseases and compared favourably with 21 dermatologists [13]. A study of dermoscopic melanoma recognition found that a convolutional network outperformed most of 58 dermatologists, with the largest gap relative to less experienced clinicians [60]. The most informative result for clinical practice, however, is human-computer collaboration. In a large reader study, combined human-AI performance exceeded either alone, but incorrect AI suggestions degraded the accuracy of less experienced readers more than that of experts [61].

**4.3.3 Breast screening.** The international evaluation reported absolute reductions of 5.7% in false positives and 9.4% in false negatives in the UK dataset, and 1.2% and 2.7% in the US dataset, relative to radiologists, along with a simulated reduction in second-reader workload [14]. A multi-dataset reader comparison found stand-alone

AI performance comparable to that of an average radiologist [62]. A randomised controlled screening trial subsequently reported that AI-supported reading reduced screen-reading workload by about 44% in its safety analysis, without evidence of inferior safety [63]. This is among the strongest prospective signals in the field, and its benefit is chiefly efficiency rather than discovery of new disease.

**4.3.4 Thoracic imaging.** A three-dimensional deep learning system for low-dose CT lung cancer screening achieved an AUC of about 94% and reduced false positives and false negatives in settings without prior imaging [64]. A chest-radiograph algorithm matched radiologists on most of fourteen pathologies, outperformed them on one and underperformed on three [65]. In tuberculosis, a multi-site evaluation of three commercial systems in Nepal and Cameroon reported AUCs between 0.92 and 0.94 [66], and the World Health Organization subsequently endorsed computer-aided detection for screening in people aged fifteen and over [67]. A cautionary study showed that pneumonia models performed less well on data from outside their training hospitals, partly because they learned site-specific cues [68].

**4.3.5 Pathology.** A challenge benchmark for lymph-node metastasis detection found that the best algorithms exceeded a panel of pathologists working under time constraints (AUC about 0.99 versus about 0.88) [69]. A clinical-grade system trained on tens of thousands of slides with only slide-level labels reached AUCs above 0.98 for several cancer types [70]. These results suggest pathology AI is promising for triage and quality control, though workflow digitisation is a prerequisite.

**4.3.6 Cardiology and physiological signals.** A deep neural network classified twelve rhythm classes from ambulatory ECGs with an F1 score exceeding the average cardiologist's [71]. An AI-enabled ECG detected left ventricular systolic dysfunction with an AUC of about 0.93, hinting at opportunistic screening from an inexpensive test [72]. Retinal photographs have been used to predict cardiovascular risk factors, although discriminative performance for cardiac events was modest (AUC about 0.70) [73].

**4.3.7 Early-warning systems.** In acute care, a recurrent network predicted a substantial share of in-hospital acute kidney injury up to 48 hours ahead, including most cases requiring dialysis, at the cost of roughly two false alerts per true alert [74]. By contrast, an external validation of a widely deployed proprietary sepsis model found weak discrimination (AUC about 0.63), detection of only about a third of sepsis cases and alerts for a substantial share of all hospitalised patients [75]. A different sepsis system, evaluated prospectively across multiple hospitals, was associated with a relative reduction in mortality of about 19% among patients whose alerts were confirmed by clinicians within three hours [76]. Implementation research on a deep-learning sepsis system emphasised the sociotechnical work of embedding alerts in clinical routine [77]. Together these studies show that early-warning AI can succeed or fail on the same task depending on validation rigour and workflow design.

**4.3.8 Symptom checkers, LLMs and probabilistic reasoning.** An audit of 23 symptom checkers found that the correct diagnosis was listed first in only about 34% of cases and within the top twenty in about 58% [78]. Later evaluations confirmed substantial variation across apps [79]. Large language models represent a step change in apparent capability. A medically tuned model reached about 68% on USMLE-style questions [80], and a general-purpose chatbot performed at or near the passing threshold on licensing examination items [81]. Exam performance is not clinical reasoning, however. In a randomised trial, access to an LLM did not significantly improve physicians' diagnostic reasoning relative to conventional resources, although the model alone scored higher than either physician group [82]. This shows that capability does not automatically convert into clinician benefit. LLMs have also been shown to reproduce race-based medical misconceptions [83]. Causal machine learning offers an alternative to purely associational models, with a counterfactual diagnostic algorithm ranking within the top quarter of a panel of doctors on vignettes [84]. A paediatric system built from electronic records performed comparably to experienced physicians in several diagnostic categories [85].

**4.3.9 Multimodal and foundation models.** Reviews of multimodal biomedical AI emphasise the integration of imaging, genomics, notes and sensors as the next frontier [86]. The "generalist medical AI" concept proposes foundation models able to perform varied tasks with minimal task-specific labels [87]. Overviews of AI in health and medicine place these models within a broader trajectory from narrow tools to flexible assistants [88]. Evidence of clinical benefit for such systems is, as yet, almost entirely prospective in aspiration rather than in fact.

#### 4.4 Comparative performance across landmark studies

**Table 3. Comparative synthesis of landmark evidence on AI in early diagnosis**

| Domain         | Ref. | Task/design                                       | Headline result                                                                  | Comparator                 | Evidence level               | Main caveat                                        |
|----------------|------|---------------------------------------------------|----------------------------------------------------------------------------------|----------------------------|------------------------------|----------------------------------------------------|
| Retina         | [12] | Referable DR; retrospective validation            | Sensitivity about 97% at high-sensitivity point; specificity about 93%           | Ophthalmologist reference  | Retrospective                | Curated datasets                                   |
| Retina         | [56] | Autonomous DR detection; prospective primary care | Sensitivity about 87%; specificity about 91%                                     | Reading-centre reference   | Prospective                  | Image-quality dependence [58]                      |
| Retina (India) | [57] | DR vs. manual grading, two sites                  | Sensitivity about 89–92%; specificity about 92–95%                               | Manual graders             | Retrospective, multisite     | Workflow integration untested                      |
| Skin           | [13] | Lesion classification                             | Matched 21 dermatologists                                                        | Dermatologists             | Retrospective                | Limited skin-tone diversity                        |
| Skin           | [61] | Human-AI collaboration                            | Combined accuracy exceeded either alone                                          | Clinicians                 | Reader study                 | Incorrect AI advice harmed novices                 |
| Breast         | [14] | Screening mammography                             | Absolute reductions in false positives (1.2–5.7%) and false negatives (2.7–9.4%) | Radiologists               | Retrospective, two countries | Prospective effect unknown                         |
| Breast         | [63] | AI-supported reading; randomised                  | Screen-reading workload reduced by about 44%                                     | Standard double reading    | RCT (safety analysis)        | Outcome data still maturing                        |
| Lung           | [64] | LDCT screening                                    | AUC about 0.94                                                                   | Radiologists               | Retrospective                | Prior-imaging dependence                           |
| Chest X-ray    | [65] | 14 pathologies                                    | Matched radiologists on most labels                                              | Radiologists               | Retrospective                | Label noise                                        |
| TB             | [66] | CAD for TB, multi-site                            | AUC 0.92–0.94                                                                    | Reference standard         | Multi-site retrospective     | Population specificity                             |
| Pathology      | [69] | Lymph-node metastasis                             | AUC about 0.99 vs. about 0.88                                                    | Pathologists, time-limited | Challenge benchmark          | Time-constrained comparator                        |
| Cardiology     | [72] | ECG for LV dysfunction                            | AUC about 0.93                                                                   | Echocardiography           | Retrospective                | Population enrichment                              |
| Acute care     | [74] | AKI prediction up to 48 h                         | Predicted most dialysis-requiring cases                                          | Clinical course            | Retrospective                | About 2 false alerts per true alert; skewed cohort |

|            |      |                                   |                                                           |                             |                        |                                |
|------------|------|-----------------------------------|-----------------------------------------------------------|-----------------------------|------------------------|--------------------------------|
| Acute care | [75] | Sepsis model external validation  | AUC about 0.63; about one-third sensitivity               | Clinician-recognised sepsis | External retrospective | Poor generalisation            |
| Acute care | [76] | Sepsis early warning; prospective | About 19% relative mortality reduction (confirmed alerts) | Usual care                  | Prospective multisite  | Observational design           |
| Triage     | [78] | Symptom checkers                  | Correct first diagnosis about 34%                         | Vignette truth              | Audit                  | Vignette-based                 |
| LLM        | [82] | Physician diagnostic reasoning    | No significant gain with LLM access                       | Conventional resources      | RCT                    | Small sample; single task type |

*Interpretation.* Three patterns stand out in Table 3. First, performance is highest and best replicated in well-defined, image-based classification with a stable reference standard. Second, studies at the prospective or randomised level tend to report efficiency or process gains (workload, time) more readily than diagnostic gains. Third, the same task can yield opposite outcomes under different validation conditions, as the two sepsis studies show [75,76]. This variation across validation conditions is arguably the most important comparative finding of the review.

## 5. ANALYSIS FROM MULTIPLE PERSPECTIVES

### 5.1 The technical perspective: generalisation, robustness and explainability

**Distribution shift.** Models trained on one hospital's data learn the quirks of its scanners, protocols and populations. The pneumonia study demonstrated this concretely [68]. Reviews that find high heterogeneity and sparse external validation [38,40] are measuring the same underlying weakness. Models that appear to learn pathology may be exploiting confounders such as equipment markers or prevalence differences between sites.

**Adversarial vulnerability.** Medical machine learning systems can be manipulated through small, imperceptible perturbations, and the financial incentives in healthcare create realistic motives for such attacks [89]. Robustness testing remains uncommon in clinical AI papers.

**Explainability.** Proponents of explainability argue that clinicians and patients need reasons as well as outputs, and a multidisciplinary analysis identified technological, legal, clinical and patient-centred grounds for this [90]. Sceptics note that current post-hoc explanation techniques may give false comfort. They highlight regions of an image without guaranteeing that the model's reasoning is faithful or clinically meaningful [91]. A further view holds that, for high-stakes decisions, inherently interpretable models should be preferred over explaining black boxes [92]. These positions are not reconcilable by technique alone. The right balance depends on the task, the risk and the clinician's ability to verify the output independently.

### 5.2 The clinical perspective: human-AI interaction and workforce

**Automation bias and over-reliance.** An experiment on clinical decision aids found that participants were influenced by advice irrespective of its purported source, and that low-quality advice degraded decisions [93]. A study of mammography reading showed that incorrect AI suggestions reduced radiologists' accuracy sharply, most severely for less experienced readers [94]. This echoes the dermatology finding that novices are most harmed by wrong AI advice [61]. Because AI mistakes are least likely to be caught by the people who most need support, AI can widen rather than narrow skill gaps.

**The clinician's role.** Commentators argue that the proper aim is not to replace physicians but to free them for the human elements of care, which algorithms cannot supply [95]. Clinical decision support in the AI era must be designed to preserve clinician judgement, provide understandable output and avoid alert fatigue [96].

**Workforce and organisational factors.** Systems-level analyses of burnout and bureaucracy in healthcare professionals show that excess administrative burden is a driver of distress [97]. AI that adds alerts or documentation overhead will compound this, whereas AI that removes low-value tasks may help. Leadership style shapes how technology is adopted in hospitals [98], and crisis-leadership scholarship stresses organisational learning from failure [99]. Both are relevant to how institutions respond when AI systems err.

### 5.3 The ethical perspective: bias, privacy and accountability

**Algorithmic bias.** Bias can enter through unrepresentative data, label choice and deployment context. A widely cited analysis of a commercial risk algorithm found that, at the same risk score, Black patients were considerably sicker than white patients, because the algorithm used healthcare cost as a proxy for need, and correcting this would have substantially increased the share of Black patients flagged for extra care [100]. In diagnostics, a large study of chest-radiograph classifiers found systematically higher underdiagnosis rates among female, younger, Black, Hispanic and lower-income patients [101]. Electronic health record data carry biases arising from differential access, documentation and missingness [102]. Frameworks for clinical safety stress that bias is a patient-safety issue, not merely a fairness concern [103]. Others argue that AI could also reduce disparities if designed with equity as an explicit objective [104,105]. Broader analyses of algorithmic bias and social stratification in digital societies reach a similar conclusion, namely that algorithms inherit and can amplify existing social hierarchies [106].

**Privacy and surveillance.** Large-scale health data aggregation challenges traditional notions of consent and de-identification, and existing privacy law fits poorly with data-hungry machine learning [107]. Wider scholarship on digital surveillance and sovereignty in the Global South warns that data infrastructures can be repurposed for control, which makes governance of health data a political as well as technical issue [108]. Ethical analyses of implementing machine learning in healthcare identify consent, transparency, fairness and the duty of care as core concerns [109].

**Accountability.** When an AI-assisted diagnosis is wrong, responsibility is distributed among developer, institution and clinician. Current liability doctrines were not designed for adaptive, opaque software, and the resulting ambiguity may encourage either defensive over-reliance on AI or defensive disregard of it [111].

### 5.4 The regulatory and legal perspective

Regulation is moving from device-centric to lifecycle-centric models. The World Health Organization's guidance sets out ethical principles including protecting autonomy, promoting well-being, ensuring transparency, fostering accountability, ensuring inclusiveness and promoting sustainable AI [110]. The European Union's AI Act classifies many medical AI systems as high-risk and imposes obligations of risk management, data governance, transparency and human oversight [112]. Legal analyses highlight unresolved issues around software as a medical device, adaptive algorithms and cross-border data flows [111]. Health-legislation scholarship argues that statutory frameworks are central to strengthening national health systems and must anticipate technological change [113]. A roadmap for responsible machine learning in healthcare adds that regulation should be coupled with deployment-time monitoring, because performance drifts as populations and practice change [114]. The reporting standards in Section 3 (TRIPOD+AI, CONSORT-AI, DECIDE-AI) function as soft regulation by raising the evidential floor for publication [31–33].

### 5.5 The health-system and economic perspective

AI succeeds or fails inside organisations. Digital-transformation research in hospital management identifies leadership, infrastructure, staff capability and change management as determinants of success [52]. Machine learning is also being applied to hospital operations, including facility management, which suggests that diagnostic AI will compete with and complement other digital priorities for scarce capital [115]. Evidence-based environmental design and educational interventions have been argued to improve healthcare outcomes [116], a reminder that diagnosis is one input among many. Infection-control scholarship likewise frames patient safety as a bundle of system strategies [117]. Because diagnostic AI changes workflow, it requires structured organisational learning and workforce development in the same way as other major change programmes [118]. Trauma-informed and patient-centred cultures matter too, since they shape whether patients accept and benefit from technological care [119]. Analyses of patient engagement in AI-driven systems argue that communication, transparency and shared decision-making must be redesigned alongside the technology [120].

The economic case remains under-evidenced. Efficiency gains, such as the screen-reading reduction in breast screening [63], are tangible, but few studies report cost-effectiveness, downstream costs of false positives or the cost of monitoring and maintaining models. These gaps are a priority for health-technology assessment.

## 5.6 The equity and Global South perspective

The promise and the risk are both greatest in low-resource settings. Analyses of AI in resource-poor contexts argue that it could help where specialists are scarce, but only if it fits local infrastructure, data realities and disease profiles [121]. Reviews of applications in rural areas of developing countries describe both opportunities and obstacles, including connectivity and training [122]. A scoping review of low- and middle-income evidence found many pilots and few rigorous evaluations [46]. India's national AI strategy identified healthcare as a priority domain, reflecting both need and ambition [125].

Digital health equity scholarship shows that access and outcomes in telemedicine vary systematically across socioeconomic groups [123]. This suggests that diagnostic AI delivered through digital channels may inherit the same gradient. Work on smart cities and spatial justice makes the parallel point that technological upgrades can concentrate benefit in already advantaged areas [124]. Socio-legal analyses of health disparities in Punjab show how structural inequities and legislative gaps shape care [20], and epidemiological studies of India document a heavy and shifting disease burden that algorithms trained elsewhere may not reflect [18,19]. Social stigma, such as that documented for cancer among students in Punjab, can also delay presentation regardless of how good the diagnostic test is [126]. In such settings, an accurate algorithm addresses only one link in a long chain from symptom recognition to treatment.

The tuberculosis and diabetic-retinopathy experiences are encouraging, because they show AI-assisted screening validated and endorsed in lower-resource settings [57,66,67]. The Thai clinic experience, however, shows that pre-deployment context assessment is non-negotiable [58].

## 6. SYNTHESIS: ANALYTICAL PROPOSITIONS

To make the review's reasoning citable and testable, we state its principal conclusions as propositions derived from the evidence above.

**Proposition 1 (The validation-gradient proposition).** Reported diagnostic performance of AI declines as validation moves from internal, to external, to prospective, to randomised designs, and the decline is steepest where models exploit site-specific cues. *Support:* [38,40,68,75,76].

**Proposition 2 (The efficiency-before-discovery proposition).** The earliest robust prospective benefits of diagnostic AI are in workload and process efficiency rather than in improved detection of disease. *Support:* [14,63,82].

**Proposition 3 (The novice-vulnerability proposition).** Wrong AI advice harms less experienced clinicians most, so unsupervised deployment can widen skill-related disparities in care. *Support:* [61,94].

**Proposition 4 (The proxy-and-subgroup proposition).** Population-level accuracy conceals subgroup failure, and bias often originates in the choice of labels and proxies rather than in algorithmic design alone. *Support:* [100,101,102,103].

**Proposition 5 (The context-dependence proposition).** The clinical value of a diagnostic algorithm depends on the surrounding pathway: image quality, referral capacity, treatment availability, stigma and trust. *Support:* [58,120,126].

**Proposition 6 (The governance-lag proposition).** Regulation and liability frameworks lag behind adaptive and generative systems, and reporting standards currently perform much of the quality-assurance work. *Support:* [31–33,110–112,114].

## 7. FUTURE DIRECTIONS

### 7.1 From retrospective accuracy to pragmatic trials

The priority is evidence of effect on outcomes. Systematic reviews of randomised trials of machine learning interventions show the trial literature is small, mostly single-centre and inconsistent in reporting participant characteristics [128]. Pragmatic, multi-site, stepped-wedge or cluster-randomised trials that measure time to diagnosis, stage at diagnosis, treatment and mortality are required. Early-stage human-factors studies reported under DECIDE-AI should precede them, so that failures of usability do not masquerade as failures of efficacy [33].

### 7.2 Standardised reporting and appraisal

Wider adoption of TRIPOD+AI, CONSORT-AI, DECIDE-AI and QUADAS-AI would make studies comparable and appraisable [28,31–33]. Journals, funders and regulators could make these checklists conditions of submission or approval. Living systematic reviews, as pioneered during COVID-19, could keep syntheses current in a field that changes monthly [44].

### 7.3 Privacy-preserving collaboration

Federated learning trains shared models without moving patient data and is promising for rare diseases and underrepresented populations [126]. A multi-institutional COVID-19 study using this approach reported gains in average discrimination (about 16%) and generalisability (about 38%) over locally trained models [127]. Federated approaches do not remove the need for governance, auditing and incentives for participation, and the surveillance concerns raised earlier still apply [108].

### 7.4 Fairness by design and continuous auditing

Fairness should be assessed before deployment and monitored afterwards, with subgroup reporting by sex, age, ethnicity, geography and socioeconomic status as a minimum standard [101,105]. Dataset documentation, label audits and participatory design with affected communities are further safeguards. Where social stratification is the underlying driver, technical fixes must be paired with structural ones [106].

### 7.5 Human-centred deployment

Future systems should present calibrated uncertainty, support disagreement rather than silent override, and adapt their level of assistance to clinician experience, given the novice-vulnerability evidence [61,94]. Training programmes should address AI literacy and automation bias. Organisational learning mechanisms, including incident reporting, feedback loops and leadership engagement, should be established before go-live [99,118].

### 7.6 Multimodal, causal and generative models

Multimodal integration may improve early detection where single-modality signals are weak [86], and causal approaches may improve robustness to distribution shift [84]. Foundation models and LLMs require especially careful evaluation, including bias audits, hallucination testing and prospective trials with clinical endpoints [47,82,83,87]. Specialist applications such as AI-driven prediction of antimicrobial resistance illustrate how AI diagnosis can extend to laboratory and genomic data and support stewardship [53].

### 7.7 Equity-focused implementation research in low-resource settings

India and similar settings need implementation studies that measure not only accuracy but also uptake, referral completion, cost and patient experience. These should be co-designed with frontline workers and linked to national digital health architecture and legislation [113,125]. Investment in connectivity, data standards and workforce training is a precondition for any benefit [123].

## 8. STRENGTHS AND LIMITATIONS

**Strengths.** The review integrates primary evidence, secondary reviews, regulatory analysis and equity scholarship in a single framework. It applies consistent appraisal lenses (QUADAS-2, QUADAS-AI, PROBAST) and offers explicit, testable propositions.

**Limitations.** First, the evidence base was purposively assembled rather than retrieved by an executed, documented search, so selection bias is possible and numeric PRISMA flow data are not reported. Second, heterogeneity precluded meta-analytic pooling, so comparisons in Tables 2 and 3 are narrative and should not be read as pooled estimates. Third, the field moves quickly, especially for large language models, and some findings will date rapidly. Fourth, publication bias probably inflates reported performance. Fifth, evidence from low-resource settings and from non-English literature is underrepresented. Sixth, regional sources drawn from the supplied reference list were used for contextual and conceptual support, and their findings were not independently verified here. Finally, any future update should add dual independent screening and a registered protocol.

## 9. CONCLUSION

AI has demonstrated, in well-defined and largely retrospective tasks, that it can match and sometimes exceed clinicians in detecting disease from images, signals and records. The strongest prospective signals so far concern efficiency, as in screening workload, and selected early-warning applications. The weakest signals concern whether algorithms improve outcomes for patients, particularly across institutions and populations. Related reviews converge on this diagnosis: high heterogeneity, high risk of bias, scant external validation and limited reporting quality. The perspectives analysed here show that the obstacles are not only technical. Automation bias, subgroup underperformance, opaque reasoning, regulatory lag, organisational unreadiness and structural inequity each can turn a statistically excellent algorithm into a clinically marginal or even harmful one.

The agenda is therefore clear. Evidence should move from retrospective accuracy to pragmatic trials with patient-centred outcomes. Reporting and appraisal should be standardised. Fairness should be audited continuously and by subgroup. Deployment should be designed around human factors and organisational learning. Privacy-preserving collaboration should widen the data base. Implementation research in low-resource settings should be funded as a priority rather than an afterthought. If these steps are taken, AI can plausibly shorten the diagnostic interval, extend scarce expertise and reduce preventable harm. If they are not, the field risks producing a large body of accurate algorithms with little demonstrable effect on patients' lives.

## REFERENCES

1. Balogh EP, Miller BT, Ball JR, editors. Improving diagnosis in health care. Washington (DC): National Academies Press; 2015.
2. Singh H, Meyer AND, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ Qual Saf.* 2014;23(9):727-31.
3. Newman-Toker DE, Schaffer AC, Yu-Moe CW, Nassery N, Saber Tehrani AS, Clemens GD, et al. Serious misdiagnosis-related harms in malpractice claims: the "Big Three" – vascular events, infections, and cancers. *Diagnosis (Berl).* 2019;6(3):227-40.
4. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med.* 2005;165(13):1493-9.
5. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med.* 2003;78(8):775-80.
6. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inform Decis Mak.* 2016;16(1):138.
7. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44-56.
8. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380(14):1347-58.
9. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng.* 2018;2(10):719-31.
10. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230-43.
11. Ahluwalia PS. Artificial intelligence in healthcare: the role of artificial intelligence in diagnostic medicine. *Siddhanta's Int J Adv Res Arts Humanit.* 2025:164-75.

12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-10.
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-8.
14. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94.
15. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689.
16. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1(6):e271-97.
17. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
18. Ahluwalia PS. Disease burden in India: a comprehensive analysis. *Siddhanta's Int J Adv Res Arts Humanit*. 2025;179-235.
19. Ahluwalia PS, John S. Disease pattern/trends in India: a comprehensive analysis of epidemiological transitions and public health challenges (2020-2025). *Siddhanta's Int J Adv Res Arts Humanit*. 2025;20-51.
20. Singh D, Ahluwalia PS. A socio-legal study on health disparities in Punjab: structural inequities, legislative gaps, and pathways to equitable healthcare. *Shodh Prakashan*. 2026;1-14.
21. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
22. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009;339:b2535.
23. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467-73.
24. Ahluwalia PS. Enhancing systematic transparency through PRISMA-ScR: a scoping review approach for emerging researchers. *Siddhanta's Int J Adv Res Arts Humanit*. 2025;1-27.
25. Ahluwalia PS. Structured exploration through PRISMA-ScR: a transparent method for mapping research landscapes. *Siddhanta's Int J Adv Res Arts Humanit*. 2025;137-67.
26. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-36.
27. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-8.
28. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med*. 2021;27(10):1663-5.
29. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019;322(18):1806-16.
30. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
31. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
32. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-74.
33. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(5):924-33.
34. Elliason EK, Ahluwalia PS. Prevalence and predictors of trauma-related mental health outcomes among refugees and internally displaced persons in Africa during conflict and crisis: a systematic review and meta-analysis. *Interdiscip J Afr Alliance Res Advocacy Innov*. 2025;1(2).
35. Ahluwalia PS. Ethical issues in social science research. *Shodh Patra Int J Multidiscip Stud*. 2024;1-8.
36. Ahluwalia PS. Plagiarism: detection techniques and tools. *Siddhanta's Int J Adv Res Arts Humanit*. 2024;1-11.
37. Ahluwalia PS. The role of artificial intelligence in combating plagiarism in scholarly publishing. *Shodh Prakashan J Humanit Soc Sci*. 2025;26-30.
38. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol*. 2019;20(3):405-10.
39. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR Med Inform*. 2019;7(3):e10010.

40. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med.* 2021;4(1):65.
41. Muehlmatter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health.* 2021;3(3):e195–203.
42. Benjamins S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med.* 2020;3:118.
43. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell.* 2021;3(3):199–217.
44. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ.* 2020;369:m1328.
45. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med.* 2021;4(1):54.
46. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *NPJ Digit Med.* 2022;5(1):162.
47. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930–40.
48. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform.* 2018;19(6):1236–46.
49. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60–88.
50. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin.* 2019;69(2):127–57.
51. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018;18(8):500–10.
52. Ahluwalia PS. Assessing the impact of digital transformation on hospital management: a systematic review. *Int J Healthc Hosp Manag Stud.* 2025:1–16.
53. Ahluwalia PS. Artificial intelligence-driven prediction of antimicrobial resistance patterns: integrating machine learning with genomic epidemiology for precision antibiotic stewardship. *J Integr Res.* 2025:97–122.
54. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017;318(22):2211–23.
55. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* 2018;24(9):1342–50.
56. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med.* 2018;1:39.
57. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol.* 2019;137(9):987–93.
58. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* New York: ACM; 2020. p. 1–12.
59. Kanagasalingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney ML, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. *JAMA Netw Open.* 2018;1(5):e182665.
60. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol.* 2018;29(8):1836–42.
61. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. *Nat Med.* 2020;26(8):1229–34.
62. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst.* 2019;111(9):916–22.
63. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* 2023;24(8):936–44.
64. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med.* 2019;25(6):954–61.
65. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* 2018;15(11):e1002686.

66. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci Rep*. 2019;9(1):15000.
67. World Health Organization. WHO consolidated guidelines on tuberculosis. Module 2: screening – systematic screening for tuberculosis disease. Geneva: World Health Organization; 2021.
68. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15(11):e1002683.
69. Bejnordi BE, Veta M, van Diest PJ, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199-210.
70. Campanella G, Hanna MG, Geneslaw L, Mirafior A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301-9.
71. Hannun AY, Rajpurkar P, Haghighpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25(1):65-9.
72. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25(1):70-4.
73. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-64.
74. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116-9.
75. Wong A, Oates E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181(8):1065-70.
76. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med*. 2022;28(7):1455-60.
77. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform*. 2020;8(7):e15182.
78. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*. 2015;351:h3480.
79. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. *BMJ Open*. 2020;10(12):e040269.
80. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-80.
81. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198.
82. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2440969.
83. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6(1):195.
84. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun*. 2020;11(1):3923.
85. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25(3):433-8.
86. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28(9):1773-84.
87. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259-65.
88. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-8.
89. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. 2019;363(6433):1287-9.
90. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310.
91. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-50.
92. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.
93. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. 2021;4(1):31.

94. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*. 2023;307(4):e222176.
95. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. 2018;319(1):19-20.
96. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA*. 2018;320(21):2199-200.
97. Ahluwalia PS. Burnout and bureaucracy: a systems-level analysis of mental health challenges in healthcare professionals. [Date unknown].
98. Ahluwalia PS. Healthcare leadership and management: effective leadership styles in hospital administration. *Siddhanta's Int J Adv Res Arts Humanit*. 2025:69-82.
99. Ahluwalia PS. Crisis leadership in healthcare: organizational learning from pandemic failures. *Siddhanta's Int J Adv Res Arts Humanit*. 2025:136-46.
100. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-53.
101. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27(12):2176-82.
102. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544-7.
103. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-7.
104. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics*. 2019;21(2):E167-79.
105. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-72.
106. Ahluwalia PS. Algorithmic bias and social stratification: how AI shapes inequality in digital societies. *Siddhanta's Int J Adv Res Arts Humanit*. 2025:28-37.
107. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. *Nat Med*. 2019;25(1):37-43.
108. Ahluwalia PS. The rise of techno-authoritarianism: digital surveillance and sovereignty in the Global South. *Siddhanta's Int J Adv Res Arts Humanit*. 2025:34-54.
109. Char DS, Shah NH, Magnus D. Implementing machine learning in health care – addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-3.
110. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021.
111. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, editors. *Artificial intelligence in healthcare*. London: Academic Press; 2020. p. 295-336.
112. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union*. 2024.
113. Ahluwalia PS, John S. The role of health legislation in strengthening national health systems: a policy-legal perspective. *Interdiscip J Anal Educ Sci Commer Manag Law*. 2025:1-20.
114. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25(9):1337-40.
115. Singh P. Integrating machine learning for smart facility management in healthcare hospitality environments. *Shodh Patra Int J Multidiscip Stud*. 2025:82-92.
116. Ahluwalia PS. Transforming healthcare outcomes through evidence-based environmental design and educational interventions. *Siddhanta's Int J Adv Res Arts Humanit*. 2025:279-318.
117. Ahluwalia PS. Improving patient safety: strategies for reducing hospital-acquired infections. *Siddhanta's Int J Adv Res Arts Humanit*. 2024:126-45.
118. Ahluwalia PS, John S, Husain I, Elliason E. Organization learning & development: a comprehensive analysis of strategic human capital development in Indian corporate ecosystem. *Interdiscip J Afr Alliance Res Advocacy Innov*. 2025;1(3).
119. Ahluwalia PS, John S, Elliason EK. Embedding trauma-informed care into healthcare leadership: a paradigm shift in organizational culture and patient outcomes. *Interdiscip J Afr Alliance Res Advocacy Innov*. 2025;1(2).
120. Ahluwalia PS. Redesigning patient engagement in AI-driven healthcare systems: an analysis. *Siddhanta's Int J Adv Res Arts Humanit*. 2025:203-17.
121. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health*. 2018;3(4):e000798.
122. Guo J, Li B. The application of medical artificial intelligence technology in rural areas of developing countries. *Health Equity*. 2018;2(1):174-81.
123. Ahluwalia PS. Digital health equity: evaluating access and outcomes in telemedicine across socioeconomic groups. *Siddhanta's Int J Adv Res Arts Humanit*. 2025:84-92.

124. Ahluwalia PS. Smart cities and social divide: spatial justice in the age of urban tech. *Siddhanta's Int J Adv Res Arts Humanit.* 2025;185-202.
125. NITI Aayog. National strategy for artificial intelligence: #AIforAll. New Delhi: Government of India; 2018.
126. Elliason EK, Singh K, Singh AP. Breaking the silence: understanding cancer-related stigma among tertiary students in Punjab. *Siddhanta's Int J Adv Res Arts Humanit.* 2025;1-24.
127. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med.* 2020;3:119.
128. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw Open.* 2022;5(9):e2233946.
129. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med.* 2021;27(10):1735-43.