ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1515

Integrating Artificial Intelligence into Healthcare to Improve Quality of Care

Shihab Ahmed Hameed[1]; Afnan Sh. Ahmed[2]; Akram M. Zeki[3]

[1]Artificial Intelligence and Robotics Engineering Department, Albayan University, Iraq
[2]Ministry of Health, Iraq
[3]Faculty of ICT, International Islamic University Malaysia - IIUM, Malaysia

Publication Date: 2026/11/04

Abstract: Artificial intelligence (AI) is transitioning from accuracy studies on historical data to prospective, real-world testing of performance in everyday care, but whether this will lead to better quality of care is debated. We review latest peer-reviewed evidence (2023-26) by six quality domains-effectiveness, safety, efficiency, patient-centeredness, cost and equity across set of validated articles. As expectation for use in screening, the strongest evidence is from a RCT on mammography-screening (n=105,934 women) where AI increased mean cancer detection by 29% and decreased work rate by 43.6% without increased false positives. For large language models (LLMs), results are mixed: LLM-aided clinicians outperform unassisted clinicians (relative top-1 accuracy 1.13) in a pooled analysis; a 2026 cluster-randomized controlled trial on Kenyan primary care found an LLM-supported clinical decision tool safe but not reducing treatment failures; an uncontrolled before-after study showed reduced clinician burnout from 51.9% to 38.8% with ambient AI tools in EMR documentation; and economic analyses are generally positive but usually omit capital and infrastructure costs. We argue that AI improves quality when prospectively validated, integrated into clinicians' workflows, governed post-deployment, audited for equity, and propose a five-pronged framework for AI integration.

Keywords: *Artificial Intelligence, Healthcare Quality, Clinical Decision Support, Large Language Models, Evidence Synthesis, Health Equity, Digital Health.*

How to Cite: Shihab Ahmed Hameed; Afnan Sh. Ahmed; Akram M. Zeki (2026) Integrating Artificial Intelligence into Healthcare to Improve Quality of Care. *International Journal of Innovative Science and Research Technology*, 11(10), 2663-2670. https://doi.org/10.38124/ijisrt/26oct1515

I. INTRODUCTION

All health systems are under pressure from increasing demand, workforce capacity constraints and persistent variation in quality of care. Diagnostic error, delayed triage and administrative burden are not theoretical challenges and all impact negatively on safety, timeliness and efficiency of care. AI- ML models for imaging and risk prediction, and more recently large language models (LLMs) is being promoted as a solution. That was demonstrated in the growth in published literature and a series of review-of-reviews that have documented use of AI in most clinical disciplines while acknowledging variable reporting standards [1].

Enthusiasm, however, is not quality improvement. A systematic review of AI for quality of hospital services in Saudi Arabia included only twelve eligible studies, but concluded there were improvements in diagnostic accuracy, patient management and operational efficiency [2]. A more extensive JAMA systematic review of LLM use found that only around 5% of evaluations involved real patient-care data [3]. Thus, the key questions to policy-makers and clinicians remain whether application in care pathways offers tangible quality improvements, and in which circumstances.

This brief tackle this question by systematically interrogating the latest peer-reviewed evidence available in the field today.

The paper is driven by four research questions:

- RQ1 – What is the strongest available evidence that AI supports higher care-effectiveness and higher safety;
- RQ2 – Does AI support higher care-efficiency, clinician-experience and financial value;
- RQ3 – What are the associated equity and reliability challenges of integrating AI; and
- RQ4 – What factors determine successful integration in practice?

The innovations from this research are an evidence matrix linking AI to care domains, design-based grading of the evidence and a five-pillar framework for responsible integration.

Scope and definitions. Throughout, "AI" means machine-learning systems-deep-learning image analysis, predictive models and generative large language models-with the purpose of assisting or replacing a clinical or

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1515

administrative task. "Quality" is defined broadly as per the multi-domain definition below (figure 1), rather than using a single metric like accuracy, because a tool that is accurate but unsafe, inequitable or unaffordable cannot improve care. Retrospective accuracy studies are generally given less weight than (prospective or pragmatic) randomized and other pragmatic evaluations, and negative and neutral results are reported alongside positive outcomes to inform evidence-based and proportionate policy.

II. CONCEPTUAL FRAMEWORK AND METHODOLOGY

➢ *Domains of Quality*

We define a range of operational domains of care quality, more common in quality-improvement literature than research practice: effectiveness (detection, diagnosis and outcomes), safety and reliability, timeliness and efficiency, patient-centeredness and clinical burnout, economic value and equity. For each paper we identified the domain or domains it is most relevant to (Fig. 1). (Left side of figure) describes the domain specific what AI can do; (right) defines the context under which those qualities are realized.

➢ *Source Confirmation and Identification*

Indexing, abstracts and descriptors of peer-reviewed articles from the following sources were searched in October 2026: npj Digital Medicine, Nature Medicine, The Lancet family, JAMA Network Open, BMC Health Services Research, JMIR AI and Frontiers in Digital Health. Synergistic search queries by combination of artificial intelligence, healthcare quality, clinical decision support, large language models, randomized trial, ambient documentation, screening and fairness were used. Inclusion criteria were peer-reviewed publication 2023-2026, human clinical or health-system endpoints or a scoping or systematic review of the same, and a digital-object identifier. Trial registrations, preprints, press releases and editorial commentary were excluded.

➢ *Grading and Interpreting the Evidence*

Subsequently, studies were classified by design (table I): for randomized trials, 'Grade A,' for systematic reviews, meta-analyses, and scoping reviews, 'Grade B,' and for uncontrolled or quality-improvement designs, 'Grade C'. Results were synthesized narratively across each of the three domains of interests (clinical management, organization of care, and patient perspective), accounting for effect direction, size, comparator, context, bias risk, and conflicts of interest. This is an analytical synthesis for a specific purpose rather than a prespecified systematic review; see Section VI.

➢ *Comparability of Results*

Since the individual studies vary in task, population, comparator and outcome the effect sizes are not pooled across the studies. The results are presented with individual comparators and denominators. Where a systematic review has pooled estimates, these are reported as reported by the review, and comparisons across quality domains are qualitative, with a focus on the direction, consistency and certainty of effects rather than on a single summary measure.

Table 1 Evidence Grading Used in this Study

Grade	Design	Included references
A	Randomized or cluster-randomized trial (including prespecified analyses)	[4]–[6], [9]–[11]
B	Systematic review, meta-analysis, scoping review or overview of reviews	[1]–[3], [7], [8], [13]–[15]
C	Uncontrolled pre–post or quality-improvement study	[12]

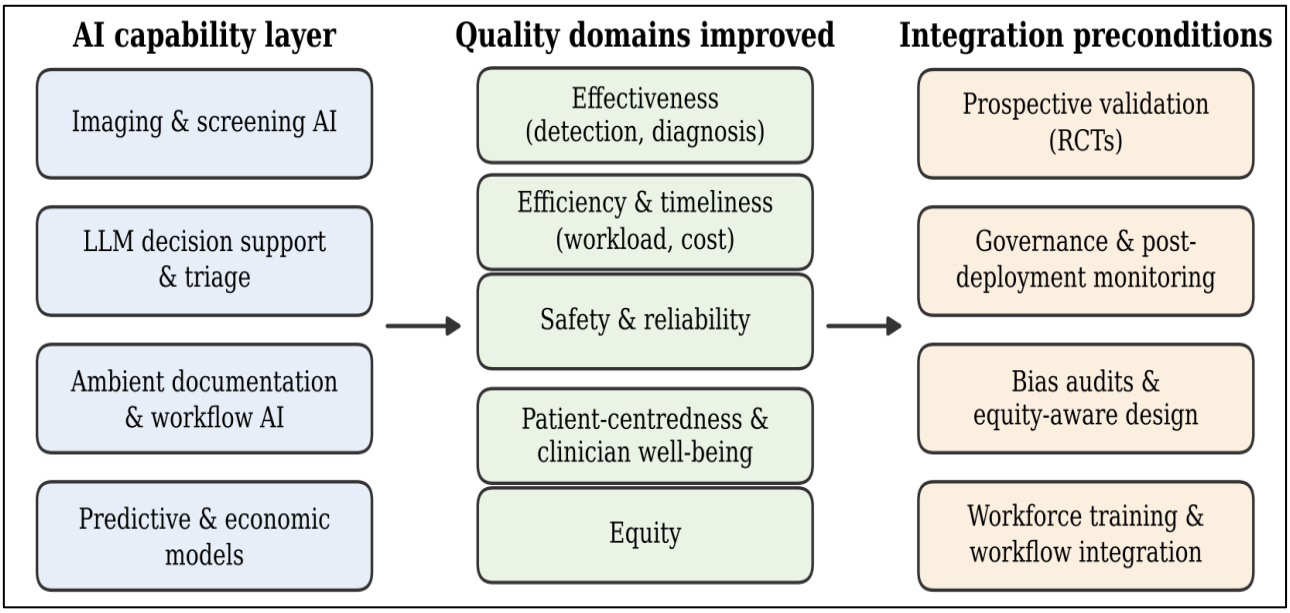


Fig 1 Conceptual Framework: AI Capabilities, the Care-Quality Domains they can Improve, and the Preconditions that Determine Whether Improvement is Realised (Author's Synthesis of [1]–[15]).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1515

III. EVIDENCE ANALYSIS

➢ *Effectiveness: Imaging and Diagnosis*

The strongest human RCT evidence that AI improves effectiveness is from population mammography screening. The MASAI trial randomized 105,934 women in the Swedish national program to AI-supported screening or double reading control. AI-supported reading identified 338 cancers among 53,043 women (6.4 per 1,000) versus 262 among 52,872 in the control arm (5.0 per 1,000) a 29% relative increase-but with half the screen-reading workload and not an excess of false positives [4]. Recalls were only modestly increased in the AI-supported arm (1,110 versus 1,027), consistent with an increase in true detections rather than indiscriminate recall [4]. An earlier safety analysis had already established the feasibility and workload savings of AI-supported reading [5], and the final results, which address the endpoint (interval cancers-the reason screening exists to reduce), showed a lower interval cancer rate with AI-supported screening [6]. A key caveat is that benefit was shown for one particular AI system (Transpara) within a specific reading protocol, in which radiologists saw AI risk and lesion information at the time of reading [4]; it should not be extrapolated to "AI mammography" as a generic term.

The story is more complicated for LLMs. One systematic review and multilevel meta-analysis of 50 studies and 25 LLMs in 2026 reported that LLMs alone were slightly less accurate than clinicians in determining the top-1 diagnosis (relative accuracy 0.89, 95% CI 0.79 -1.00), roughly equivalent for longer differential lists, and similar for triage (1.01, 0.94-1.09) [7]. Importantly, clinicians aided by an LLM were better than unassisted clinicians: relative accuracy was 1.13 (1.00-1.27) for top-1 and 1.42 (1.16-1.73) for top-5 diagnoses (Fig. 2) [7]. Another earlier meta-analysis of 83 studies found a pooled generative-AI diagnostic accuracy of 52.1% with no difference from non-expert physicians but worse than expert physicians [8]. There are clear reasons to be interested: primary-care misdiagnosis has been estimated at approximately 20% and triage errors at 2-31%, leading to harmful delays in care [7]. Taken together, LLMs seem to have more of the quality of cognitive aids that expand the differential rather than functioning as standalone diagnosticians.

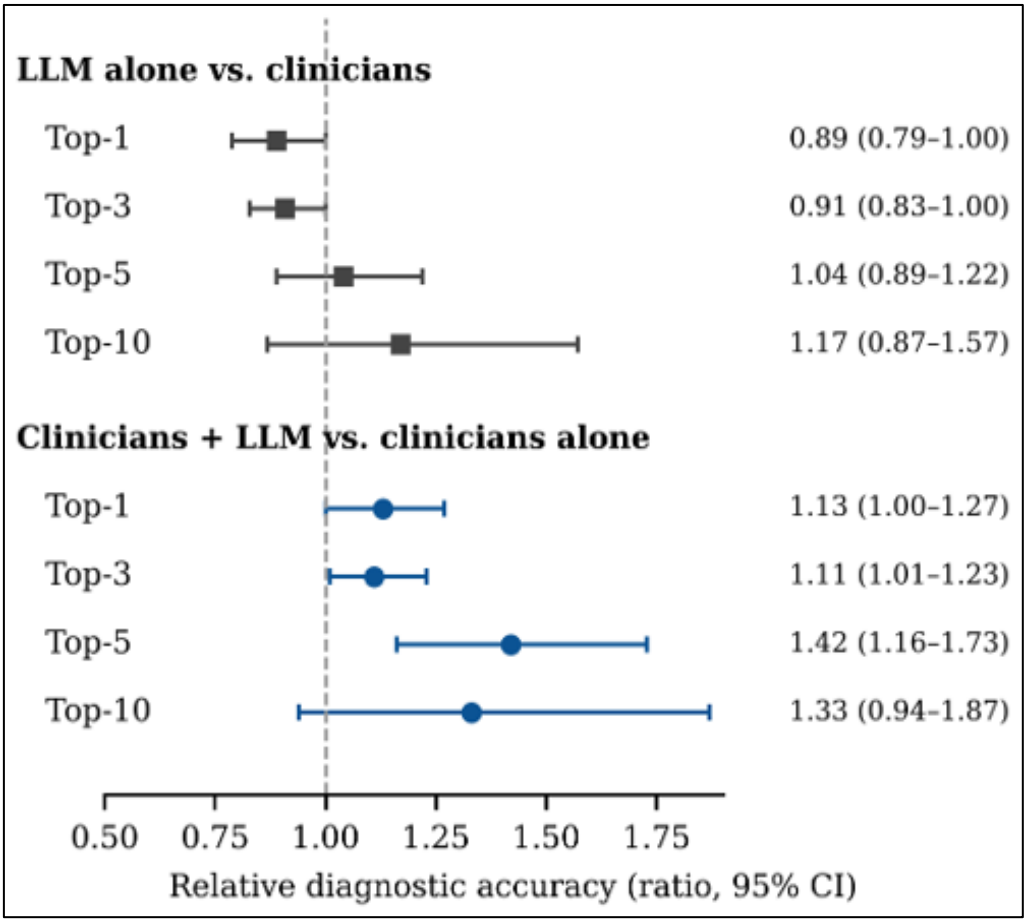


Fig 2 Pooled Relative Diagnostic Accuracy of Large Language Models (LLMs) Versus Clinicians, and of LLM-Assisted Clinicians Versus Clinicians Alone (Values from Chen et al. [7]; Ratios >1 Favor the First-Named Group).

This same review, however, also reveals just how fragile this evidence base is: 3/50 studies were prospective (all, triage), 29 (58%) were at high or unclear risk of bias in participant selection; 40 (80%) had applicability concerns due to LLM inputs not matching routine clinical practice; and 4 (8%) showed publication bias among the top-1 assisted-versus-unassisted comparison [7]. Even the randomized evidence on physician performance is inconsistent: one randomized trial of GPT-4 help measured the effect on performance on patient-care tasks [9], while the prior

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1515

randomized clinical trial found that providing clinicians LLM access not enhancing diagnostic reasoning, despite the model performing well on its own [10]. The recurring warning is that the success comes from the interactions between models and clinicians, not from model performance alone.

➢ *Safety and Reliability in Routine Care*

Safety is the most tangible aspect of translation from benchmark to bedside. A 2026 pragmatic, cluster-randomized trial of the LLM tool embedded in an electronic medical record in 16 primary-care facilities in Kenya ran 103 clinical-officer clusters (52 intervention, 51 control) and examined 9,702 encounters [11]. There was no serious adverse event associated with the intervention and no safety signal; while LLM assistance was not associated with lower treatment failure within 14 days, any effect was likely small [11]. This negative-to-neutral result is informative: it demonstrates safe implementation of a decision-support tool in a resource-limited setting, yet acknowledges that improved documentation and reasoning may not always lead to improved short-term health outcomes.

This is also echoed in the wider systematic evidence: much of the benchmark data for LLM evaluation consists of hand-selected cases rather than live data from patient-care [3], [7], meaning that accuracy estimates may not translate well into clinical contexts with noise, incomplete information, and multi-turn conversational interaction. The benchmarks therefore require prospective validation in the target environment, ongoing tracking of performance, and accountability for final clinical decisions [7].

➢ *Efficiency, Timeliness and Clinician Well-Being*

Efficiency benefits are the most widely reported. In MASAI, the use of ambient AI to support the scribe ship workflow eliminated about 44% of the screen-reading workload [4], [5], freeing up critical radiology time for additional readings or screening a greater number of women. An additional paper demonstrated the benefit of AI to reduce the documentation burden (an acknowledged factor leading to clinician burnout) in a prospective multicenter quality-improvement study of 263 physicians and advanced-practice clinicians across six US health systems, where the use of AI scribes over 30 days resulted in a reduction in the prevalence of burnout from 51.9% to 38.8% (odds ratio 0.26, 95% CI 0.13-0.54), and reduced post-hours documentation by 0.90 hours [12].

Should be viewed with caution: the scribe trial was uncontrolled, self-selected, with outcomes reported by those who participated and several writers had conflicts of interest with the vendor, which also collected pooled survey data [12]. It suggests a large well-being impact is possible but does not quantify the magnitude, durability or impact on note accuracy or patient outcomes. Efficiency is a quality domain only if time gained is used for care rather than loaded on more demand [15].

Table 2 Evidence Matrix: AI Applications Mapped to Care-Quality Domains

Quality domain	AI application	Key finding (as reported by the cited authors)	Grade	Ref.
Effectiveness	AI-supported mammography screening	29% higher cancer detection (6.4 vs 5.0 per 1,000; n = 105,934); no increase in false positives; lower interval-cancer rate in final analysis	A	[4], [6]
Effectiveness	LLM-assisted diagnosis	Assisted clinicians vs. alone: relative top-1 accuracy 1.13 (1.00–1.27); LLM alone 0.89 (0.79–1.00)	B	[7], [8]
Safety	LLM decision support, primary care (Kenya)	No related serious adverse events; no reduction in 14-day treatment failure; benefit, if any, modest (9,702 encounters)	A	[11]
Efficiency	AI-triaged mammography reading	About 44% lower screen-reading workload	A	[4], [5]
Well-being	Ambient AI scribe	Burnout 51.9% → 38.8% (OR 0.26); after-hours documentation −0.90 h; uncontrolled, vendor-linked	C	[12]
Economic value	Multiple clinical AI applications	Generally favorable ICERs and savings across 19 evaluations; indirect and infrastructure costs often omitted	B	[13]
Equity	Fairness of clinical AI	Scarce fairness research; narrow attributes; limited clinician-in-the-loop designs	B	[14]
Adoption	Clinician perspectives	49 facilitating and 43 hindering factors in 72 studies; 62 studies hypotheticals	B	[15]

➢ *Economic Benefit*

A 2025 systematic review of 19 economic evaluation (2020-2024) in field of cancer, cardiology, ophthalmology, dentistry, infection, intensive care finds that clinical AI interventions in general lead to greater diagnostic accuracy, quality-adjusted life-year benefits, and potential cost-savings compared to traditional care [13]. For example, review describes machine-learning targeted atrial-fibrillation screening being incrementally cost-effective at £4,847-£5,544 per quality-adjusted life-year (QALY) (well below NHS threshold of £20,000), 14-19.5% per-patient reduction in screening costs of diabetic retinopathy, and AI-assisted colonoscopy in Japan saving US$149.2 million annually and US$85.2 million in US [13]. Quality of reporting according to CHEERS 2022 (score 21-27/28).

The same review is clear about costs. Not all applications cost less-AI glaucoma screening delayed one third of blindness but added long-term cost-and 7 of 19 evaluations used static rather than dynamic models that could not factor in AI learning, only four assessed indirect costs, and infrastructure, integration and ongoing maintenance costs

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1515

were often under-estimated [13]. Cost-effectiveness is therefore contingent on test cost, accuracy and workflow integration.

➢ *Equity and Fairness*

Equity can be exacerbated by integration too. A 2025 scoping review and evidence-gap analysis of fairness in clinical AI, for instance, demonstrated few fairness studies across many medical specialties, limited scope of attributes examined, focus on group-level fairness measures to equalize model accuracy, and limited engagement of the clinician in the loop [14]. An economic review similarly showed that most economic outcomes were not disaggregated by demographic group, so an average that appeared fair may be paying off for one group more than others [13].

➢ *Adoption and Integration By Clinicians*

Technology not trusted or understood by clinicians will not improve quality. A 2025 systematic review of 72 studies and 15,325 healthcare professionals identified 49 enablers and 43 barriers at the individual, interpersonal, institutional, community and policy levels [15]. Enablers included accepted diagnostic quality, workflow efficiency, user-friendliness, education and training, evidence from randomized trials, and clear legal frameworks; barriers included poor AI understanding, lack of education programs, fears of displacement, and blind deference [15]. Almost all studies measured hypothetical or case scenarios of AI, not operational systems – as with the clinical evidence, local evidence too is in its infancy, with the Saudi review recording just twelve studies [2].

➢ *Regulations and Governance*

So governance is crucial to the sustainability of quality improvements. Over 1,000 AI-enabled medical devices have been approved by the US Food and Drug Administration [7], but LLMs challenge existing approval routes because they are generative and overall-purpose. WHO, the US Food and Drug Administration and European Union, as in the 2026 meta-review, agree on lifecycle governance, transparency, adaptability to changing model power, and fairness [7]. The review recommends tiered, risk-based governance, adversarial testing, audit logs of AI recommendations, and that final control rests with an adequately skilled clinician. It also warns that automation bias and skill decline are themselves quality issues [7]. These parallel what clinicians want: clear legal frameworks for data protection and liability [15].

➢ *Resource-Limited and Regional Settings*

Most evidence on clinical AI derived from high-income systems and so is often not generalizable. Two, related, findings are promising. First, the Kenyan trial demonstrated that rigorous, pragmatic evaluation of an LLM-based tool is possible in a network of primary-care clinics [11], and the economic evidence includes evidence from Singapore and rural China [13] where AI-based screening was cost-effective or cost-saving. Second, barriers in low- and middle-income systems include insufficient data infrastructure, inability to afford funds and local technical expertise, and the cost-effectiveness is sensitive to unit costs and the area's type of financing [13]. The regional evidence synthesized in the Saudi review [2] shows that locally-commissioned evidence is still scarce, supporting the case for investing in local pilots with prospective evaluation.

➢ *Certainty of the Evidence*

Seven of the fifteen sources, by design, are a randomized trial or prespecified trial analysis (Grade A), eight are a synthesis (Grade B), and one is an uncontrolled trial (Grade C). The preponderance of randomized evidence focused on imaging screening [4]–[6], and on LLM decision support [9]–[11], with none of the randomized studies within this set reporting economic impact, equity, or clinician well-being as the primary outcome. This disequilibrium means that the sector with the strongest evidence (effectiveness) is not the sector that typically informs health policy (cost and equity), and that associations for the latter sectors rely primarily on syntheses of disparate studies [13], [14].

IV. COMPARATIVE CASE ANALYSIS

Three examples demonstrate how the strength of inference depends on study design (see Table 3 and Fig. 3).

➢ *Case 1: AI-Supported Mammography*

Using a narrow clear test (reading screening mammograms), an objective reference test (cancers [4] and, ultimately, interval cancers), individual randomization within a national screening program and a clear reading protocol [4]-[6] combined to give MASAI several of the advantages of a randomized trial. Inference is strengthened: the comparator was the best practice of the time, the outcome measures can be validated, and the intervention was described clearly and in sufficient detail to be repeatable. The disadvantages of MASAI are that the benefit applies specifically to the individual device and the protocol tested [4], higher detection alone is not necessarily equivalent to lives saved, and application to other populations or other devices will need local validation.

➢ *Case 2. How would an LLM Provide Decision Support for Primary Care?*

By contrast, our Kenyan trial targeted a generic decision-support tool in an everyday consultation setting, with cluster randomization of practitioners, a distal outcome (2-week treatment failure) with many other influences than the clinician's decision, and clinicians' independence of decision [11]. The neutral outcome should not be interpreted as indicating that the tool had no value: it was safety-tested and our trial also assessed practice, documentation, patient satisfaction and safety outcomes [11]. It does show that process effects do not necessarily translate into distal outcomes. Reasons-moderate effect size, outcome definitions, differential use of advice-remains hypothetical, not proven, explanations.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1515

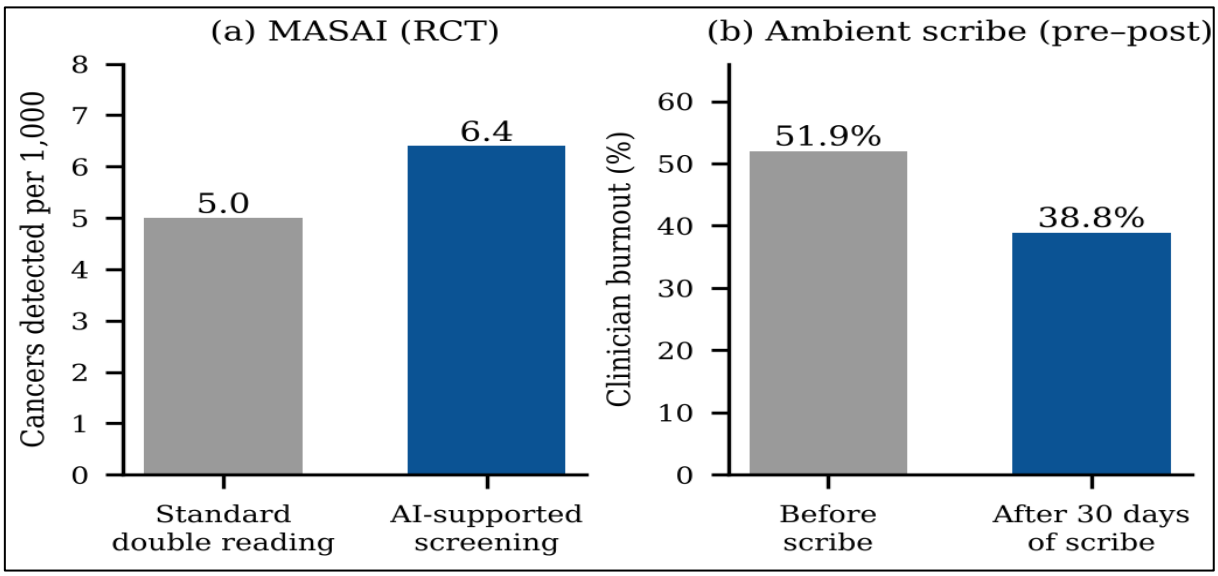


Fig 3 Headline Effects of Two Applications: (a) Cancers Detected Per 1,000 Women in the MASAI Randomized Trial [4]; (b) Clinician Burnout before and after 30 Days of an Ambient AI Scribe in an Uncontrolled Study [12]. The Designs Differ in Strength and the Panels should Not be Compared Directly.

➢ *Case 3: Ambient AI Documentation*

Scribe study asked a different question–does a workflow tool lessen documentation burden–and used a pre-post design without a control [12]. The large reduction in a self-report outcome measured over 30 days is impressive, but novelty effects, regression to the mean, and self-selection of enthusiastic volunteers are among factors that cannot be eliminated, and note accuracy and patient outcomes were not analyzed [12].

➢ *How the Cases Differ*

There are three factors that distinguish the strongest evidence from the weakest. Task specificity: the more specific the task, the easier it is to establish a benefit [4]. The verifiability of outcome: cancer detection is verifiable by design, burnout is self-reported, and treatment failure is multi-factorial [11], [12]. And the closeness of outcome to the intervention: it's easier to report effects on process and workload measures than downstream patient outcomes. Together, these factors account for why we should not compare headline numbers from the authoring or delivery of a variety of different AI solutions, and why claims of "AI improving quality" need to be made with a study design.

Table 3 Design Features of Three Illustrative Applications

Feature	AI mammography	LLM support (Kenya)	Ambient scribe
Task	Screen reading	Consultation decision support	Note drafting
Design	Randomised (1:1)	Cluster-randomised	Pre–post, no control
Comparator	Standard double reading	Same EMR without LLM	Baseline (self)
Setting	Swedish national program	16 Kenyan clinics	6 US health systems
Headline result	29% more cancers detected	No reduction in treatment failure	Burnout 51.9% → 38.8%

Table 4 Five-Pillar Integration Checklist

Pillar	Practical action	Supporting evidence
Validation	Prospective or pragmatic trial in the target setting before scale-up	[3], [7], [11]
Workflow	Define when and how AI output is shown; keep clinician autonomy	[4], [11]
Governance	Auditable logs, post-deployment monitoring, accountable clinician	[7]
Equity	Subgroup reporting and bias audits; representative data	[13], [14]
Workforce	Training, role clarity and AI literacy programs	[12], [15]

V. DISCUSSION

Synthesis. Table 2 summarizes the evidence matrix. The evidence consistently points towards a qualified though robust inference: augmenting AI can raise the bar on care quality; if it does so, but it does so for a narrow task, produces a finite and obvious output and is part of a well-scoped working process (screening mammography as the archetype [4]-[6]), the benefits are the easiest to demonstrate. If AI is a more general purpose and advisory (LLM decision support), its benefits hinge on human-AI interaction and are more substantiated for diagnostic reasoning in pooled analyses [7], [8] than for patient effects in a pragmatic trial [11]. Efficiency and well-being enhancements are plausible and large but comparatively less conclusive [12].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1515

Why integration, not adoption, is the unit of analysis. The same model facilitated or hindered MASAI across contexts. The benefit of MASAI was found when AI information was made available to radiologists at reading [4]; the Kenyan trial found clinical officers maintained full autonomy while having a modest positive effect [11]; and assisted versus unassisted gains were large for broad differentials [7]. Quality improvement is therefore a property of the socio-technical system as an entity, a point reinforced by the multi-level factors found in clinician studies [15].

Evidence gaps. There are 3 [3], [7]: (1) few prospective, outcomes-based trials for LLMs; (2) lack of economic and equity analyses; and (3) lack of evidence from lower-resource settings, but the Kenyan trial [11] and regional reviews [2] suggest such evidence is both feasible and urgently needed.

➢ *Five-Pillar Framework for Integration*

The case implies five pillars (Fig. 1, right).

- Prospective validation: have randomized or pragmatic trials with patient-important outcomes before scale-up [3], [7], [11].
- Workflow design: define the timing and presentation of AI output to clinicians, as MASAI has [4].
- Governance and monitoring: hold final oversight of recommendations with clinicians and audit AI use after deployment [7].
- Equity-aware design: evaluate subgroup out-of-sample performance and audit for bias [13], [14].
- Workforce readiness: train and address source of fear and resistance early [15]. Table IV factors each of these to provide guidance and evidence.

➢ *Stakeholder Considerations*

Clinicians and care teams the evidence supports viewing AI as a diagnostic aid rather than a diagnosis. Assisted clinicians scored better in combined studies than unaided clinicians, whereas LLMs alone did not reach clinicians for the final diagnosis [7]. Instead, teams should be educated to query AI suggestions, to maintain awareness of automation bias, and to keep up their diagnostic skills [7], [15].

Health-system leaders. Leaders should pilot new tools with a prospective evaluation plan, quantify local outcomes rather than quoting manufacturer or external validation accuracy alone [7], [11], and cost-model all expenses, including the costs of infrastructure, integration, and ongoing support [13]. Time saved by AI- such as reading time for screening [4] or documentation time [12]-should be explicitly returned to patient-facing care, otherwise there will be no quality gain.

Regulators and payers. Risk-based oversight should extend post-approval, with each version and its performance on subpopulations made transparent [7], [14]. Reimbursement and procurement should be based on dynamic, distributional economic models rather than stationary models that can inflate expectations [13].

Researchers & developers. Instructions for use should include model name, version, access mode, prompts, inference settings, real or prospective data, clinicians as a comparator & equity-relevant subgroups [1], [7], [14]. There is little value in synthesizing evidence unless the quality of reporting allows policy-makers to have confidence in the findings.

➢ *Research Agenda*

Four further priorities emerge from these gaps. First, pragmatic RCTs of LLM- and ambient-based tools with patient-relevant outcomes such as diagnostic error, treatment failure, re-hospitalization and safety events, rather than accuracy or satisfaction. Second, longer-term evaluations of well-being interventions with control groups and objective outcomes such as documents' quality. Third, dynamic and distributional economic evaluations that incorporate full costs of integration. Fourth, fairness studies that go beyond group performance metrics to explore the unique features of real-time clinical workflows with the clinician in the loop.

VI. LIMITATIONS

This is a focused analysis and synthesis of sources rather than a formal registered systematic review, so relevant studies may have been missed and bias in selection cannot be ruled out. The grading of evidence is a pragmatic, design-based, not formal rating of certainty of evidence (e.g. GRADE). A number of these synthesis sources (e.g. [7], [13]) are themselves heterogeneous and based on non-prospective data. The rapidly moving LLM landscape means findings may quickly become out-of-date; some effects (e.g. The ambient-scribe [12]) were from uncontrolled designs with potential vendor influence. Numerical results are those reported by the authors cited and were not reanalyzed.

VII. CONCLUSION

An additional step towards responsible adoption: embedment of AI in a clinical pathway, prospective validation, accountability, audit and optimization, and training clinicians are Identified. Amassing evidence for large language models (LLMs) has proved difficult and shows mixed results: systematic reviews are heterogeneous; bias detection even more so; and impacts on patients, clinicians, and healthcare institutions need further evaluation. Although most efforts focus on correcting prejudice and promoting equity in the use of AI, evidence on the economic impact is scarce. The evidence of AI as a clinical intervention suggests a safe, effective implementation in healthcare. However, further high-quality evidence is needed to be scaled safely and responsibly, involving prospective clinical validation, preparedness of clinicians and health systems, implementation research, and economic evaluation.

ACKNOWLEDGMENT

Researchers wish to express their appreciation to Albayan university, Baghdad-Iraq for all their supports to this research.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1515

REFERENCES

[1]. G. Morone et al., "Artificial intelligence in clinical medicine: a state-of-the-art overview of systematic reviews with methodological recommendations for improved reporting," *Front. Digit. Health*, vol. 7, Art. no. 1550731, 2025, doi: 10.3389/fdgth.2025.1550731.

[2]. E. Alghareeb and N. Aljehani, "AI in health care service quality: systematic review," *JMIR AI*, vol. 4, Art. no. e69209, 2025, doi: 10.2196/69209.

[3]. S. Bedi et al., "Testing and evaluation of health care applications of large language models: a systematic review," *JAMA*, vol. 333, pp. 319–328, 2025, doi: 10.1001/jama.2024.21700.

[4]. V. Hernström et al., "Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomized, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study," *Lancet Digit. Health*, vol. 7, no. 3, pp. e175–e183, Mar. 2025, doi: 10.1016/S2589-7500(24)00267-X.

[5]. K. Lång 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 randomized, controlled, non-inferiority, single-blinded, screening accuracy study," *Lancet Oncol.*, vol. 24, no. 8, pp. 936–944, 2023, doi: 10.1016/S1470-2045(23)00298-X.

[6]. J. Gommers et al., "Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomized, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial," *Lancet*, 2026, doi: 10.1016/S0140-6736(25)02464-X.

[7]. M. Chen et al., "Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage," *npj Digit. Med.*, vol. 9, Art. no. 222, 2026, doi: 10.1038/s41746-026-02409-8.

[8]. H. Takita et al., "A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians," *npj Digit. Med.*, vol. 8, Art. no. 175, 2025, doi: 10.1038/s41746-025-01543-z.

[9]. E. Goh et al., "GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial," *Nat. Med.*, vol. 31, pp. 1233–1238, 2025, doi: 10.1038/s41591-024-03456-y.

[10]. E. Goh et al., "Large language model influence on diagnostic reasoning: a randomized clinical trial," *JAMA Netw. Open*, vol. 7, Art. no. e2440969, 2024, doi: 10.1001/jamanetworkopen.2024.40969.

[11]. A. Agweyu et al., "Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial," *Nat. Med.*, 2026, doi: 10.1038/s41591-026-04503-6.

[12]. K. D. Olson et al., "Use of ambient AI scribes to reduce administrative burden and professional burnout," *JAMA Netw. Open*, vol. 8, no. 10, Art. no. e2534976, 2025, doi: 10.1001/jamanetworkopen.2025.34976.

[13]. R. A. El Arab and O. A. Al Moosa, "Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare," *npj Digit. Med.*, vol. 8, Art. no. 548, 2025, doi: 10.1038/s41746-025-01722-y.

[14]. M. Liu et al., "A scoping review and evidence gap analysis of clinical AI fairness," *npj Digit. Med.*, vol. 8, 2025, doi: 10.1038/s41746-025-01667-2.

[15]. D. Henzler et al., "Healthcare professionals' perspectives on artificial intelligence in patient care: a systematic review of hindering and facilitating factors on different levels," *BMC Health Serv. Res.*, vol. 25, Art. no. 633, 2025, doi: 10.1186/s12913-025-12664-2.