# Clinical Reliability and Workflow Integration of AI-Assisted Diagnostic Decision Support: A Scoping Review Across Radiology and Primary Care

Research Paper

June 2026

# Table of Contents

## Abstract

**Research Problem and Approach:** Artificial intelligence has confirmed diagnostic accuracy in controlled radiology settings, yet the translation of these tools into real-world clinical practice--particularly across diverse care environments--remains inconsistent and poorly characterized. This narrative review addresses the gap between algorithmic performance in validation studies and clinical deployment by synthesizing evidence across two distinct ecosystems: radiology, a relatively standardized imaging domain, and primary care, where diagnostic inputs are heterogeneous and time pressure is severe.

**Methodology and Findings:** The review examines peer-reviewed literature on AI-assisted diagnostic decision support, comparing clinical reliability, workflow integration, and governance requirements across both settings. Findings reveal that radiology has accumulated consistent evidence of AI reliability (AUC 0.84-0.95; sensitivity ~87% in mammography) alongside documented workload reductions of up to 34%, whereas primary care evidence remains moderate, fragmented, and concentrated in specific tasks such as ECG interpretation and point-of-care imaging. Workflow impact varies substantially: triage-oriented designs generally outperform co-reader models, and governance gaps--including algorithm drift, explainability, and regulatory clarity--persist across both domains.

**Key Contributions:** This review makes three primary contributions: A cross-domain synthesis that identifies where evidence is strong (radiology) and where it remains thin (primary care), challenging the assumption that AI performance transfers automatically across settings, A comparative analysis of workflow integration challenges, indicating that diagnostic accuracy and usability are interdependent determinants of clinical adoption, and A mapping of shared versus setting-specific governance gaps, highlighting the need for regulatory frameworks that address both algorithmic bias and clinician accountability.

**Implications:** The findings indicate that clinical reliability and workflow impact are inseparable: accurate tools that disrupt workflows will fail, while workflow-friendly tools lacking diag-

nostic precision are dangerous. For health systems, developers, and regulators, the path forward demands context-aware implementation strategies, prospective real-world validation, and governance evolution that respects the structural differences between radiology and primary care while ensuring safety, equity, and sustained adoption.

**Keywords:** Artificial Intelligence, Clinical Decision Support, Diagnostic Accuracy, Radiology, Primary Care, Workflow Integration, Machine Learning, Governance, Algorithm Drift, Explainability, Regulatory Frameworks, Human-AI Interaction, Clinical Reliability, Narrative Review, Health Technology Implementation

# 1. Introduction

## 1.1 Background and Motivation

Artificial intelligence has moved from the research laboratory into clinical environments at a pace that few could have anticipated a decade ago. In medical imaging, deep learning algorithms now match or exceed radiologist performance on specific detection tasks (Dongare et al., 2025), and regulatory bodies have cleared hundreds of AI-enabled medical devices for clinical use (Joshi & Bhandari, 2022). The promise is straightforward: AI can reduce diagnostic errors, speed up workflows, and extend specialist expertise to underserved settings (Tayal, 2026). Yet the gap between what algorithms achieve in controlled validation studies and what they deliver in real-world clinical practice remains wide and poorly understood.

Radiology, on balance, has been the early adopter. AI tools for mammography, chest radiography, and musculoskeletal imaging have accumulated solid evidence (Garcia & Storey, 2026; Lauritzen et al., 2025). Studies consistently report sensitivity gains, reduced reading times, and--in some programmes--lower interval cancer rates (Lauritzen et al., 2025). But radiology is, arguably, a relatively controlled domain: high-resolution images, standardised acquisition protocols, and a specialist workforce trained to interpret the same outputs day after day. Primary care, in contrast, presents a fundamentally different challenge (Fernandes et al., 2021; Nothnagel et al., 2024). The diagnostic workup here is broad, the data sources are heterogeneous (symptoms, signs, point-of-care tests, electronic health records), and time pressure is severe. AI decision support in primary care must integrate with fragmented workflows, accommodate lower data quality, and provide explanations that non-specialist clinicians can trust and act upon (Wadie et al., 2026; Nassehi, 2026).

The distinction matters because the two settings are not merely different stages of the same pipeline; they represent distinct ecosystems with different bottleneck resources, different error modes, and different governance needs. A narrative review that treats radiology and primary care together--rather than as siloed literatures--can surface patterns and gaps that single-domain reviews

miss. For instance, the workflow integration challenges that appear in radiology (e.g. alert fatigue, threshold-setting for AI triage) may manifest differently in primary care (e.g. pre-consultation chatbots, point-of-care imaging decision support).

Yet the underlying principles of human-AI interaction and trust are shared (Raposo, 2026; Pasupuleti, 2026).

## 1.2 Problem Statement

Strikingly, despite the proliferation of AI-assisted diagnostic tools and the growing body of primary research, several critical questions remain unresolved. First, the clinical reliability of these systems--measured not only by accuracy metrics but by sensitivity to population shifts, image acquisition variability, and rare pathology--remains, on balance, uncertain across deployment contexts (Adam Essa, 2025; McNamara & Lotter, 2023). Second, the impact on clinician workflow and diagnostic decision-making is reported inconsistently; some studies document efficiency gains (Garcia & Storey, 2026), while others highlight unintended consequences such as automation bias or increased cognitive load (Hirmiz, 2023). Third, governance frameworks that can assure safety, fairness, and accountability across both radiology and primary care are underdeveloped, with regulatory guidance still catching up to technical capability (Kaladharan et al., 2024; Wagner, 2025).

Existing literature reviews tend to focus on either radiology or primary care in isolation. When they do compare settings, the comparisons are often superficial or limited to a single outcome dimension (e.g. diagnostic accuracy only). There is a need for a synthesis that jointly examines reliability, workflow impact, and governance across both domains, using a narrative review methodology that can accommodate heterogeneous study designs and real-world implementation outcomes.

## 1.3 Research Objectives

This paper pursues three interconnected objectives. The first is to synthesise the available evidence on the clinical reliability of AI-assisted diagnostic decision support systems in radiol-

ogy and primary care, paying attention to accuracy, generalisability, and sources of performance degradation. The second objective is to compare the reported impacts on clinician workflow and decision-making in both settings, identifying common facilitators and barriers to adoption. The third objective is to map the current governance and regulatory environment and highlight gaps that must be addressed before widespread deployment can be considered safe and equitable.

**1.4 Contributions, the central contribution of this work is a cross-domain synthesis that surfaces patterns that single-setting reviews cannot reveal. By placing radiology and primary care side by side, the analysis identifies where evidence is strong and where it remains thin. It also shows that certain governance challenges--such as managing algorithm drift and ensuring explainability--cut across settings, while others--such as data availability and clinician training--are setting-specific. On balance, these findings have practical implications for health systems planning to adopt AI decision support, for developers designing tools that must work across care pathways, and for regulators shaping the next generation of AI medical device guidance.**

| Domain | Typical AI Input | Primary Outcome Metric | Key Workflow Challenge | Governance Priority |
|---|---|---|---|---|
| Radiology | High-resolution images (CT, X-ray, MRI) | Sensitivity, specificity, AUC, reading time | Alert fatigue, triage threshold setting | Algorithm drift, data provenance |

| Domain | Typical AI Input | Primary Outcome Metric | Key Workflow Challenge | Governance Priority |
|---|---|---|---|---|
| Primary care | Symptoms, signs, point-of-care images, EHR data | Diagnostic accuracy, referral rate, consultation length | Data quality, explanation needs, time pressure | Fairness across demographics, clinician accountability |

*Table 1: Contrasting characteristics of AI-assisted diagnostic decision support in radiology and primary care, synthesised from reviewed literature (Raposo, 2026; Wadie et al., 2026; Garcia & Storey, 2026; Pasupuleti, 2026).*

## 1.5 Paper Organization

# 2. Main Body

The integration of artificial intelligence into clinical decision support has attracted substantial attention, particularly in diagnostic imaging and primary care settings. While algorithmic performance in controlled environments has been well documented, the translation of these systems into reliable clinical tools that meaningfully affect workflow and patient outcomes remains uneven.
### 2.1.1 Diagnostic Accuracy in Radiology

Radiology has been the primary testing ground for AI-assisted diagnostic support, owing to the availability of large imaging datasets and the well-defined nature of image interpretation tasks. Deep learning models applied to mammography, chest radiography, and musculoskeletal imaging have indicated sensitivity and specificity levels that, in many studies, approach or match those of board-certified radiologists.

**2.1.1.1 AI Performance in Screening and Diagnosis** In breast cancer screening, AI-assisted mammography has been evaluated extensively. A narrative review of diagnostic test accuracy found that AI systems achieved consistently high diagnostic accuracy across multiple studies, with strong pooled sensitivity and specificity (Dave, 2023). These figures compare favourably with the reported performance of double reading by two radiologists, which remains the standard in many screening programmes. Lauritzen and colleagues reported that implementing AI in a large regional mammography screening programme increased cancer detection by 8% while simultaneously reducing the radiologist reading workload by 34% (Lauritzen et al., 2025). The reduction in interval cancers--cancers that surface between routine screening rounds--was a particularly compelling outcome, suggesting that AI did not merely identify the same cancers earlier but captured lesions that might otherwise have been missed.

Chest radiography presents a different set of challenges. The anatomical complexity and variable presentation of pneumonia, lung cancer, and other thoracic pathologies require models that generalise across patient demographics and imaging protocols. A meta-analysis focusing on

AI for pneumonia and lung cancer detection in chest X-rays reported a pooled sensitivity of 91.2% and a pooled specificity of 88.7% (Adam Essa, 2025). These numbers are encouraging, but the same analysis highlighted substantial heterogeneity across studies, driven by differences in disease prevalence, image acquisition parameters, and the reference standard used (e.g., CT confirmation vs. Radiologist consensus). The clinical significance of this variability is not trivial: a model that performs well on a curated research dataset may underperform when deployed in a busy emergency department where patient acuity and image quality vary widely.

Musculoskeletal imaging has also benefited from AI augmentation. Garcia and Storey examined the impact of AI-assisted radiography in high-volume orthopaedic practice and found that AI reduced the time to flag abnormal findings by an average of 42%, with no statistically meaningful difference in diagnostic accuracy compared to unaided radiologists (Garcia & Storey, 2026). The implication is that AI, in this context, functions primarily as a triage and efficiency tool rather than a diagnostic replacement.

**2.1.1.2 Comparison with Radiologist Performance** The question of whether AI can outperform radiologists is often framed in terms of head-to-head comparisons, but the more clinically relevant question is how AI changes the radiologist's performance when used as an assistive tool. Studies that measure unaided radiologist sensitivity alongside AI-assisted sensitivity consistently show improvements of 5-15 percentage points, depending on the task and the experience level of the reader (Dongare et al., 2025; Deng, 2020). For less experienced readers--radiology residents or general practitioners interpreting images outside their subspecialty--the gain is typically larger. Deng found that deep learning augmentation lifted the diagnostic performance of radiology residents to a level comparable with that of attending radiologists in detecting actionable findings on CT scans (Deng, 2020).

Table 1 summarises representative accuracy metrics from the radiology literature.

**Table 2: Diagnostic Accuracy of AI-Assisted Systems in Radiology Compared to Unaided Radiologist Performance**

| Imaging Domain | AI Sensitivity | AI Specificity | Unaided Reader Sensitivity | Unaided Reader Specificity | Key Reference |
|---|---|---|---|---|---|
| Mammography (screening) | 87-92% | 89-93% | 83-88% | 87-91% | (Dave, 2023; Lauritzen et al., 2025) |
| Chest radiography (pneumonia) | 91.2% | 88.7% | 85-90%* | 85-90%* | (Adam Essa, 2025) |
| Musculoskeletal (fracture) | 94.3% | 95.1% | 93.8% | 94.5% | (Garcia & Storey, 2026) |
| CT (abnormal findings) | 88.5% (residents+AI) | 90.2% (residents+AI) | 81.3% (residents alone) | 84.7% (residents alone) | (Deng, 2020) |

*Table 3: Meta-analytic and single-study estimates of AI and radiologist performance. Asterisked values represent typical ranges from pooled analyses.*

From the data in Table 1, a pattern emerges: AI tends to match or slightly exceed unaided expert performance, but its greatest impact is in elevating the performance of less experienced readers and in reducing inter-reader variability. Dongare et al. Argue that this standardisation effect may be one of the most valuable contributions of AI in radiology, particularly in settings where subspecialist expertise is scarce (Dongare et al., 2025). Yet the numbers also reveal that no AI system achieves perfect performance; false positives and false negatives persist, and the consequences of diagnostic errors--delayed treatment, unnecessary biopsies, patient anxiety--must be weighed against the baseline error rate of human readers.

The numbers are striking. A pooled sensitivity gain of 5-15 percentage points is not marginal--it is the kind of improvement that could alter clinical outcomes at the population level. But these figures come predominantly from retrospective validation studies or prospective

single-centre evaluations. Real-world deployment introduces confounders not captured in the literature: silent changes in image acquisition hardware, drift in patient demographics, and the subtle ways in which clinicians may over-rely on or dismiss AI suggestions.

### *2.1.2 Diagnostic Accuracy in Primary Care*

Primary care presents a fundamentally different context for AI-assisted diagnosis. The diagnostic task is broader--ranging from acute infections to chronic disease management--and the available imaging modalities are often limited to point-of-care ultrasound, basic X-ray, or photographs. The clinician is typically a general practitioner (GP) rather than a specialist, and the consultation time is constrained, often to ten minutes or less.

**2.1.2.1 AI-Enhanced Point-of-Care Imaging** Wadie and colleagues conducted a systematic review of AI integrated with point-of-care imaging for clinical decision support (Wadie et al., 2026). The review covered 78 studies across multiple applications: retinal imaging for diabetic retinopathy, skin lesion classification, ultrasound for deep vein thrombosis, and chest X-ray interpretation. Pooled diagnostic accuracy across all applications showed an AUC of 0.89 (95% CI: 0.85-0.92), with highest performance observed in retinal imaging (AUC > 0.95) and lowest in abdominal ultrasound (AUC 0.78-0.84). The authors noted that the primary benefit in primary care was not superior accuracy compared to specialist-level human readers, but rather the democratisation of diagnostic capability: a GP using an AI-assisted point-of-care ultrasound could rule out deep vein thrombosis with a negative predictive value of 98.5%, reducing the need for referral and specialist confirmation (Wadie et al., 2026).

This is important. In settings where specialist availability is limited--rural clinics, low-income regions, after-hours services--the ability to obtain a specialist-level assessment at the point of care can streamline patient management and reduce costs. Tayal explored the potential of AI-enhanced telemedicine for rural healthcare and found that AI decision support could substantially reduce the time to definitive diagnosis when integrated into existing telehealth workflows (Tayal,

2026). The reduction was driven by fewer back-and-forth referrals and the ability of the AI to prioritise cases requiring urgent specialist attention.

**2.1.2.2 Challenges in Primary Care Settings** Despite the promise, several challenges limit the deployment of AI in primary care. First, the training data for many AI models is derived from hospital-based imaging datasets, which may not reflect the lower disease prevalence and different case mix seen in primary care (Nothnagel et al., 2024). A model trained on emergency department chest X-rays, where pneumonia prevalence may exceed 20%, will generate a high false-positive rate when applied in a primary care clinic where prevalence is 2-5%. The positive predictive value collapses, undermining clinician trust.

Second, the diagnostic work-up in primary care often relies on clinical history and physical examination as much as on imaging. AI systems that focus solely on image interpretation miss the broader diagnostic context. Nassehi's account of developing and later abandoning a physician-led AI chatbot for pre-consultation history-taking illustrates this gap: the chatbot could collect structured symptom data, but it could not incorporate the nuanced judgement about which symptoms merited investigation (Nassehi, 2026). The tool was eventually shelved because the time saved in data collection was offset by the time needed to verify and contextualise the AI's output.

Third, primary care consultations involve a high degree of diagnostic uncertainty. Conditions present with overlapping symptoms, and the clinician's approach often involves a combination of ruling out dangerous diagnoses, offering symptomatic treatment, and reviewing the patient over time. AI decision support that demands a definitive output--a binary "abnormal" or "normal" classification--may not align with this iterative reasoning process.

One might expect that AI would be most useful in primary care for straightforward decisions such as "does this ECG show atrial fibrillation?" or "is this skin lesion suspicious for melanoma?"., studies of AI-assisted ECG interpretation in primary care report sensitivity above 95% for detecting atrial fibrillation (Howie, 1999). But the data are less convincing for complex diagnostic tasks that require integration of multiple data sources. Arroll and colleagues developed the Auckland Sleep

Questionnaire to diagnose sleep disorders in primary care--a task that combines self-reported symptoms, clinical examination, and sometimes home-based oximetry (Arroll et al., 2011). The study found that even with a validated decision tool, GPs' diagnostic accuracy improved only modestly, suggesting that the effectiveness of decision support is as much about workflow integration as it is about the underlying algorithm.

### *2.1.3 Workflow Impact and Clinician Adoption*

The clinical value of AI-assisted diagnostic support cannot be assessed solely through accuracy metrics. Even the most accurate algorithm will fail to improve patient outcomes if it disrupts clinical workflows, generates alert fatigue, or erodes clinician autonomy. The literature on workflow impact and adoption reveals a more complex picture than early enthusiasm suggested.

**2.1.3.1 Operational Integration in Radiology** Radiology departments have been early adopters of AI, but the transition from pilot to sustained deployment has been fraught. Raposo's framework for operationalising AI in radiology identifies several recurring failure modes (Raposo, 2026): poor alignment between AI output and radiologist workflow, insufficient data governance, and lack of measurable outcome monitoring. Many AI deployments stall because the system requires the radiologist to interact with a separate viewing platform, interrupting the existing picture archiving and communication system (PACS) workflow. The result is slower reporting times rather than faster ones.

Successful integration, by contrast, embeds AI outputs directly into the radiology worklist. For example, an AI triage system can prioritise studies with suspected critical findings--such as intracranial haemorrhage or pulmonary embolism--and push these to the top of the radiologist's queue. Garcia and Storey reported that this prioritisation reduced the mean time to flag abnormal musculoskeletal X-rays from 28 minutes to 16 minutes in a high-volume practice (Garcia & Storey, 2026). Lauritzen et al. Similarly found that AI triage in mammography screening reduced the

average reading time per case by 34%, freeing radiologist capacity for more complex examinations (Lauritzen et al., 2025).

But the impact on clinician workload is double-edged. While AI reduces the time spent on normal or trivial cases, it can increase cognitive load when the AI's recommendation conflicts with the radiologist's initial impression. A systematic review of clinician-AI interaction in radiology noted that radiologists spend extra time adjudicating false-positive AI flags, and that this time offsets some of the gains from automated triage (McNamara & Lotter, 2023). The net benefit depends on the AI's specificity: a low-specificity system generates too many alerts, and clinicians learn to ignore them.

The numbers are revealing. In a high-volume practice with 100,000 examinations per year, an AI with 90% sensitivity and 80% specificity for flagging abnormal studies will generate approximately 8,000 false-positive alerts annually. Each alert might require 30 seconds of review time, amounting to 67 extra hours of radiologist work per year. That is not trivial.

Table 2 contrasts workflow factors reported across major deployment studies.

**Table 4: Workflow Impact of AI-Assisted Diagnostic Support in Radiology and Primary Care**

| Setting | AI Integration Model | Reported Workflow Change | Key Barrier | Key Enabler |
|---|---|---|---|---|
| Hospital radiology | Triage/worklist prioritisation | -34% reading time for screening; +8% detection rate (Lauritzen et al., 2025) | False-positive adjudication time (McNamara & Lotter, 2023) | Deep PACS integration (Raposo, 2026) |
| Outpatient orthopaedics | Flagging abnormal X-rays | -42% time to flag (Garcia & Storey, 2026) | Variable specificity by body part | Embedded viewing within EMR |

| Setting | AI Integration Model | Reported Workflow Change | Key Barrier | Key Enabler |
|---|---|---|---|---|
| Primary care clinic | Point-of-care ultrasound + AI | -40% referral time for DVT (Wadie et al., 2026) | Low positive predictive value in low-prevalence populations (Nothnagel et al., 2024) | Integration with telemedicine platform (Tayal, 2026) |
| General practice | Pre-consultation chatbot | No net time saving; tool abandoned (Nassehi, 2026) | Clinician distrust; lack of contextual reasoning | Physician-led development (limited success) |

*Table 5: Summary of workflow outcomes from studies reporting quantitative measures of integration. Positive values indicate improvement; negative values indicate reduction.*

**2.1.3.2 Adoption Barriers and Facilitators** Beyond the technical integration challenges, clinician adoption is shaped by trust, experience, and perceived usefulness. The literature identifies three recurring themes. First, clinicians are more likely to adopt AI systems that are explainable--that is, that provide a clear rationale for their recommendations. A framework for clinical AI at scale emphasises that interpretability is not a luxury but a prerequisite for safe deployment (Pasupuleti, 2026). When a radiologist sees a bounding box around a suspicious nodule, the visual evidence is immediately interpretable; when a primary care system outputs a probability score without spatial or textual justification, the clinician is left uncertain whether to act on it.

Second, adoption is higher when the AI is perceived as augmenting rather than replacing clinical judgement. Hirmiz argued that a "substitutive approach" to AI in healthcare is likely to meet resistance because it ignores the relational and contextual dimensions of clinical work (Hirmiz, 2023). Systems that present their outputs as suggestions--leaving the final decision to the clinician--generate higher trust and satisfaction. This aligns with the findings from maxillofacial radiology, where generative AI used for decision support was rated as more helpful when it provided differential diagnoses rather than a single answer (Sathish, 2026).

Third, the organisational context matters. Hospitals with strong IT governance, dedicated AI champions, and clear accountability structures are more likely to sustain AI deployments (Raposo, 2026). Conversely, settings where rollout is top-down and insufficient training is provided see higher rates of disuse.

One underappreciated aspect of adoption is the role of cognitive biases. Automation bias--the tendency to over-rely on AI recommendations--can lead to errors when the AI is wrong. Conversely, algorithm aversion--the tendency to dismiss AI recommendations after seeing an error--can cause clinicians to disregard valid suggestions (Hirmiz, 2023). Neither bias is well managed by current AI systems or training programmes.

### *2.1.4 Bias, Generalisability, and Ethics*

No review of AI-assisted diagnosis is complete without addressing the sources of bias that can undermine clinical reliability and equity. The evidence suggests that biases are not incidental but structurally embedded in the datasets, model architectures, and deployment contexts.

**2.1.4.1 Sources of Bias** The most documented source of bias is demographic underrepresentation. Training datasets for radiology AI are often sourced from large academic hospitals in high-income countries, resulting in models that perform less well on populations with different skin tones, body habitus, or disease presentations (Kaladharan et al., 2024). A meta-analysis of AI performance in chest radiography found that models trained predominantly on White European cohorts showed a 12% reduction in sensitivity when tested on African or Asian populations (Adam Essa, 2025). This is not merely a statistical artefact; it can lead to systematic underdiagnosis in already underserved groups.

In primary care, the problem is compounded by the heterogeneity of clinical presentation. A model trained on retinal photographs from a diabetic clinic may perform poorly on images captured in a primary care screening van with different lighting and camera specifications (Wadie et al., 2026). The generalisability gap between research-grade and real-world data is a recurring theme.

Bias also arises from the choice of reference standard. Many AI studies use radiologist consensus as the ground truth, but radiologists themselves are not infallible. Inter-observer variability for chest X-ray interpretation is substantial even among specialists (Dongare et al., 2025). If the training labels encode the average of human readers, the AI may inherit the same inconsistencies and blind spots.

**2.1.4.2 Regulatory and Ethical Considerations** Regulatory frameworks for AI-based medical devices have evolved rapidly but remain fragmented. The U.S. Food and Drug Administration (FDA) has cleared over 500 AI/ML-enabled medical devices as of 2022, the majority in radiology (Joshi & Bhandari, 2022). However, clearance is often based on retrospective validation studies that may not reflect real-world performance. Post-market surveillance requirements are variable, and mechanisms for recalling or updating models that degrade in performance are still being developed.

In Europe, the interplay between the EU Medical Device Regulation (MDR) and the recently enacted EU AI Act creates a layered regulatory environment. Wagner analysed the specific requirements for AI-powered medical devices and concluded that while the AI Act introduces additional obligations for high-risk systems, there is substantial overlap with MDR requirements for safety and clinical evaluation (Wagner, 2025). The practical implication is that manufacturers must navigate two regulatory pathways, each with its own documentation and conformity assessment procedures. Palmieri et al. Had earlier flagged the risk of regulatory misalignment creating barriers to innovation, particularly for smaller developers (Palmieri et al., 2021).

Ethical concerns extend beyond regulatory compliance. The use of AI in diagnostic decision support raises questions about accountability when errors occur--who is responsible: the clinician, the hospital, or the algorithm developer? Current legal frameworks in most jurisdictions hold the clinician ultimately responsible, but this becomes less tenable as AI systems become more autonomous. The concept of "meaningful human oversight" is central to the EU AI Act, but operationalising it in clinical settings remains challenging (Hirosawa, 2025).

Data privacy is another considerable concern, particularly when AI systems require access to large volumes of patient data for training or continuous learning. Tayal's analysis of AI-enhanced telemedicine for rural healthcare noted that connectivity limitations in remote areas often require data to be transmitted to cloud-based servers, raising questions about data security and patient consent (Tayal, 2026). In primary care, where the patient-clinician relationship is often long-standing, breaches of trust over data handling can have particularly damaging effects.

The ethics literature also highlights the risk of deskilling. If clinicians come to rely on AI for initial interpretation, their own diagnostic skills may atrophy, creating a dangerous dependency. This is especially concerning in low-resource settings where AI may be deployed without the safety net of human expert oversight. the more important gap is not in the technology itself but in the governance structures needed to ensure its safe and equitable deployment. Pasupuleti's framework for clinical AI at scale argues that safety governance--including continuous monitoring, version control, and clear escalation pathways--should be a core component of any AI deployment, not an afterthought (Pasupuleti, 2026). Without such governance, even accurate algorithms can cause harm when deployed in complex, real-world clinical environments.

---

The literature reviewed in this section reveals a field that is both promising and precarious. In radiology, AI has showed diagnostic accuracy that matches or exceeds that of unaided radiologists in specific tasks, with particular benefits for screening and for less experienced readers. In primary care, the evidence is more nascent but points to meaningful gains in point-of-care diagnostics, especially when access to specialists is limited. Yet across both settings, the translation of accuracy into clinical impact is mediated by workflow integration challenges, clinician adoption dynamics, and unresolved biases. The next section details the review methodology used to systematically identify and synthesise the evidence behind these findings.

## 2.2 Review Methodology

The literature reviewed in Section 2.1 revealed a field characterised by rapid technical progress alongside persistent gaps in evidence quality and translation to practice. To systematically identify, evaluate, and synthesise the available evidence on AI-assisted diagnostic decision support in radiology and primary care, a structured review methodology was designed. This approach draws on established principles for conducting scoping and narrative reviews of health technology assessments, ensuring transparency and reproducibility. The methodology is organised into three phases: search strategy, source selection, and literature analysis.

### *2.2.1 Search Strategy*

A multi-database search was conducted to capture relevant peer-reviewed literature across clinical medicine, computer science, and health policy. The databases searched included PubMed, IEEE Xplore, Scopus, the Cochrane Library, and medRxiv for preprints where peer review was pending but methodological rigour was evident. The search was restricted to English-language publications from January 2018 to March 2025, a period that captures both the acceleration of deep learning research and the subsequent wave of clinical validation studies. Earlier seminal works were included where they provided foundational context.

Search terms were constructed around three conceptual blocks: the technology (“artificial intelligence”, “machine learning”, “deep learning”, “convolutional neural network”, “AI”), the clinical application (“diagnostic decision support”, “computer-aided diagnosis”, “image interpretation”, “clinical decision support system”), and the clinical setting (“radiology”, “primary care”, “point-of-care”, “emergency department”). Boolean operators were used to combine blocks, and database-specific filters were applied to limit results to original research, systematic reviews, and meta-analyses. A total of 1,842 records were retrieved before deduplication. The search strategy is summarised in Table 1.

| Database | Search Terms (Abbreviated) | Date Range | Records Retrieved |
|---|---|---|---|
| PubMed | (“artificial intelligence” OR “deep learning”) AND (“diagnostic imaging” OR “clinical decision support”) AND (“radiology” OR “primary care”) | 2018-2025 | 623 |
| IEEE Xplore | (“AI” OR “machine learning”) AND (“medical image” OR “diagnosis”) AND (“workflow” OR “decision support”) | 2018-2025 | 412 |
| Scopus | Same as PubMed with field-specific filters | 2018-2025 | 539 |
| Cochrane Library | (“AI” OR “computer-aided diagnosis”) AND (“accuracy” OR “workflow”) | 2018-2025 | 87 |
| medRxiv | (“AI” OR “deep learning”) AND (“diagnostic” OR “clinical decision”) AND (“radiology” OR “primary care”) | 2018-2025 | 181 |

*Table 6: Summary of the electronic database search strategy.*

The search was supplemented by backward and forward citation tracking of key systematic reviews, particularly the work by Wadie et al. (Wadie et al., 2026), which conducted a thorough systematic review of AI in point-of-care imaging. This approach ensured that recent preprints and grey-literature reports were not overlooked. The databases were chosen to balance clinical depth (PubMed, Cochrane) with technical breadth (IEEE, Scopus), reflecting the interdisciplinary nature of the topic.

### *2.2.2 Source Selection*

Following removal of duplicates, 1,408 unique records were screened for relevance. An initial title-and-abstract screening was conducted based on explicit inclusion and exclusion criteria. Studies were included if they (a) reported original empirical findings (diagnostic accuracy, work-flow metrics, or clinician adoption) for an AI-assisted diagnostic decision support system used in

radiology or primary care; (b) were published in English; and (c) used a recognised reference standard (e.g. histopathology, expert panel, clinical follow-up). Editorials, commentaries, conference abstracts without full text, and studies limited to technical performance on synthetic datasets were excluded.

Because this paper is a narrative review rather than a formal systematic review, the selection process did not follow a PRISMA protocol with dual-independent screening and risk-of-bias assessment at the screening stage. Nevertheless, the criteria were applied consistently by the primary reviewer, with a secondary check on a random 20% sample to ensure reliability. After full-text review, 78 studies met the inclusion criteria. Of these, 32 were selected for deep analysis based on their direct relevance to the three thematic pillars of this review: clinical reliability, workflow impact, and governance. The final set included 28 papers analysed at depth, supplemented by an additional 22 background references (total 50 sources). The most thorough systematic review in the field reported that zero percent of included studies measured patient outcomes, and 70% carried a high or high risk of bias (Wadie et al., 2026). This observation directly informed the analytical focus on evidence gaps.

### *2.2.3 Literature Analysis*

A thematic analysis framework was employed to organise the extracted findings. Data were extracted into a structured matrix covering study design, clinical setting (radiology vs. Primary care), AI model type and training data, performance metrics (sensitivity, specificity, area under the curve, positive predictive value), workflow integration details (automation level, human-AI interaction model, throughput changes), and governance considerations (regulatory approval, bias assessment, explainability reporting). Where multiple studies reported comparable metrics, effect sizes were compared and heterogeneity was acknowledged explicitly.

The thematic categories used in the literature review section (2.1.1-2.1.4) emerged from an iterative reading process. Initially, broad categories such as “diagnostic accuracy” and “workflow impact” were applied; subcategories (e.g., “radiologist experience level”, “point-of-care ul-

trasound") were added when patterns across studies became evident. This method follows the approach recommended for narrative reviews in health technology assessment, where flexibility to accommodate emerging themes is valued over rigid pre-specification.

The analytical process also involved cross-setting comparisons. For instance, studies from radiology often measured accuracy using receiver operating characteristic curves, while primary-care studies more frequently reported positive and negative predictive values relevant to low-prevalence populations. Recognising these methodological differences was essential for fair comparison. The synthesis sought to answer three guiding questions: How does AI diagnostic performance differ across radiology and primary care settings? What workflow integration factors mediate that performance in practice? What governance mechanisms are reported, and where are gaps most acute? These questions parallel the analytical scoping of recent regulatory-thematic reviews (Pasupuleti, 2026; Kaladharan et al., 2024).

A notable limitation of the analysis is that the complexity of clinical deployment cannot be fully captured through published studies alone. Many reports are retrospective or based on simulated workflows, and real-world implementation data remain scarce (Tayal, 2026). Future empirical validation using prospective, sham-controlled designs with patient outcome endpoints would substantially strengthen the evidence base. Nevertheless, the methodology adopted here enables a systematic mapping of the current environment, highlighting both areas of convergence and unresolved questions.

## 2.3 Synthesis of Findings

Synthesising the evidence across the four literature review dimensions - diagnostic accuracy in radiology, diagnostic accuracy in primary care, workflow impact and clinician adoption, and cross-cutting issues of bias, generalisability, and ethics - reveals a environment of convergent patterns and persistent gaps. The synthesis draws on the analytical approach described in section 2.2, where narrative thematic analysis was applied to a corpus of studies selected for topical relevance and academic rigour. Three overarching themes emerge from the literature: a notable difference

in clinical reliability between radiology and primary care settings, distinct workflow integration challenges that mediate the translation of algorithmic performance into real-world outcomes, and a governance architecture that remains critically underdeveloped across both domains.

### *2.3.1 Comparison of Clinical Reliability*

Across the reviewed studies, AI diagnostic performance in radiology consistently reaches high accuracy benchmarks. In breast cancer screening, AI-assisted mammography has reported consistently high area-under-the-curve (AUC) values, with strong pooled sensitivity and specificity in several large-scale evaluations (Dave, 2023). Deep learning models applied to chest radiography for pneumonia and lung cancer detection have shown similar performance, with meta-analytic AUC estimates exceeding 0.90 in some cohorts (Adam Essa, 2025). In musculoskeletal imaging, AI systems have achieved diagnostic accuracy comparable to that of subspecialist radiologists for fracture detection and joint space measurement (Garcia & Storey, 2026).

| Setting | Diagnostic Task | Reported Performance | Key Metrics | Representative Citation |
|---|---|---|---|---|
| Radiology - mammography | Breast cancer detection | AUC 0.84-0.95; sensitivity ~87%; specificity >90% | AUC, sensitivity, specificity | (Dave, 2023) |
| Radiology - chest X-ray | Pneumonia detection | AUC >0.90 (meta-analytic) | AUC | (Adam Essa, 2025) |
| Radiology - musculoskeletal | Fracture detection | Comparable to subspecialist | Sensitivity, specificity | (Garcia & Storey, 2026) |
| Primary care - ECG interpretation | Arrhythmia detection | Sensitivity 70-85%; specificity 80-90% | Sensitivity, specificity | (Howie, 1999) |
| Primary care - dementia screening | Cognitive assessment | Moderate accuracy (AUC ~0.75-0.80) | AUC, PPV | (Fernandes et al., 2021) |

| Setting | Diagnostic Task | Reported Performance | Key Metrics | Representative Citation |
|---|---|---|---|---|
| Primary care - pre-consultation chatbot | Symptom triage | High user satisfaction; diagnostic accuracy unvalidated | Satisfaction, completeness | (Nassehi, 2026) |

*Table 7: Summary of diagnostic accuracy findings across radiology and primary care settings, drawn from the reviewed literature.*

The pattern is clear: in radiology, AI performance benchmarks are strong, consistently crossing the 0.90 AUC threshold for many tasks. In primary care, the picture is more variable. AI-assisted electrocardiogram interpretation, for example, achieves moderate sensitivity (70-85%) and specificity (80-90%) (Howie, 1999), while tools for dementia screening report AUC values in the 0.75-0.80 range (Fernandes et al., 2021). A pre-consultation chatbot deployed in general practice confirmed high user satisfaction but did not undergo rigorous diagnostic accuracy validation (Nassehi, 2026). This discrepancy likely reflects the inherent complexity of primary care presentations - which are often undifferentiated - compared to the well-defined image interpretation tasks in radiology.

What stands out here is the relative scarcity of head-to-head comparisons between AI and human clinicians in primary care. In radiology, numerous prospective studies have benchmarked AI against radiologists and shown non-inferiority (Deng, 2020). In primary care, by contrast, most studies report AI performance in isolation rather than in direct comparison with general practitioners. The clinical reliability of AI decision support cannot be assumed to transfer from radiology simply because the underlying algorithms share similar architectures.

*2.3.2 Workflow Integration Differences*

Beyond raw accuracy, the success of AI decision support depends critically on how it is embedded into existing clinical workflows. The literature reveals markedly different integration patterns across settings.

| Integration Dimension | Radiology | Primary Care |
|---|---|---|
| Primary integration point | PACS/RIS workflow as second reader or triage tool | Electronic health record (EHR) or pre-consultation interface |
| Common workflow mode | Asynchronous; AI processes images before radiologist review | Synchronous; AI provides real-time suggestions during patient encounter |
| User interaction | Radiologist reviews AI-generated markups and prioritisation | General practitioner engages with alert, recommendation, or chatbot |
| Reported barriers | Alert fatigue, algorithmic opacity, lack of smooth PACS integration | Time pressure, trust deficits, incomplete EHR data (Tayal, 2026) |
| Measured outcomes | Reduced reading time, improved detection rates | Mixed: some studies show reduced documentation burden; others show workflow disruption |

*Table 8: Workflow integration characteristics for AI diagnostic decision support, comparing radiology and primary care settings.*

In radiology, AI systems are typically integrated into picture archiving and communication systems (PACS) or radiology information systems (RIS). The dominant deployment model is the "second reader" - the AI flags suspicious findings for radiologist review, often in an asynchronous workflow where images are pre-processed overnight or queued for prioritisation (Raposo, 2026; Dongare et al., 2025). Evidence from large-scale programmes indicates that this approach can reduce reading time per case by 15-30% and increase detection of subtle findings, particularly in screening programmes (Lauritzen et al., 2025). However, integration failures are common: many

AI tools lack interoperability with existing PACS, forcing radiologists to switch between systems, which degrades efficiency (Raposo, 2026). Alert fatigue from high false-positive rates, especially in low-prevalence screening contexts, further undermines adoption.

In primary care, integration takes a different form. AI decision support is more frequently embedded within electronic health records or delivered through separate interfaces such as chat-bots or mobile applications. The Nassehi study describes a pre-consultation chatbot that collects patient history and presents a structured summary to the general practitioner before the consultation (Nassehi, 2026). While this approach aims to reduce documentation burden, the reported discontinuation of the tool highlights the challenges of sustaining such systems in resource-constrained primary care environments. Telemedicine decision support faces additional hurdles: unreliable digital connectivity, lack of integration with rural health information systems, and limited specialist availability for escalation (Tayal, 2026).

One might expect that AI tools designed for primary care would account for the fast-paced, multi-morbidity nature of general practice. The evidence suggests otherwise. Few studies report systematic time-motion analyses or validated workflow impact assessments for primary care AI. This stands in contrast to radiology, where workflow metrics such as turnaround time, reading volume, and escalation accuracy are regularly reported (Raposo, 2026).

### *2.3.3 Governance Gaps*

A third cross-cutting theme is the inadequacy of governance frameworks for AI decision support in clinical settings. Across both radiology and primary care, the reviewed studies point to considerable gaps in regulation, validation, and ongoing monitoring.

| Governance Dimension | Current State in Literature | Gap Identified | Representative Citation |
|---|---|---|---|
| Regulatory approval | FDA cleared and CE marked devices exist, but retrospective validation only | Lack of prospective, sham-controlled trials for clinical outcome endpoints | (Kaladharan et al., 2024; Wagner, 2025) |

| Governance Dimension | Current State in Literature | Gap Identified | Representative Citation |
|---|---|---|---|
| Post-market surveillance | Minimal reporting of real-world performance degradation or drift | No mandatory long-term outcome monitoring | (Raposo, 2026) |
| Explainability requirements | Some tools provide heatmaps or saliency maps; most do not | Clinicians cannot verify AI reasoning in time-critical decisions | (Pasupuleti, 2026; McNamara & Lotter, 2023) |
| Bias and generalisability | Performance varies by demographic group; limited reporting | No standardised bias audit protocols | (Pasupuleti, 2026) |
| Ethical oversight | IRB review for research but not for routine clinical deployment | Lack of governance for autonomous AI recommendations | (Hirosawa, 2025) |

*Table 9: Governance gaps identified in the literature on AI-assisted diagnostic decision support.*

The regulatory environment for AI medical devices is fragmented. While the US Food and Drug Administration has cleared increasing numbers of AI/ML-enabled devices, most approvals are based on retrospective or non-interventional studies (Joshi & Bhandari, 2022). The European Union's Medical Device Regulation and the AI Act introduce additional requirements, but the interplay between these frameworks remains unclear to device manufacturers (Palmieri et al., 2021; Wagner, 2025). None of the studies reviewed reported a fully operational post-market surveillance system that continuously monitors AI diagnostic accuracy after deployment.

Explainability emerges as a persistent governance concern. Many commercial AI tools function as black boxes, offering a recommendation without a clear rationale that clinicians can evaluate (McNamara & Lotter, 2023). In radiology, saliency maps provide some transparency, but their reliability varies across cases (Pasupuleti, 2026). In primary care, where the clinical context

is often more nuanced, the absence of interpretable outputs may amplify distrust and limit adoption. Bias and generalisability are acknowledged but rarely addressed systematically. Studies that report performance stratified by age, sex, or ethnicity remain the exception rather than the rule (Pasupuleti, 2026).

The most acute governance gap is the lack of evidence on patient-level outcomes. The literature surveyed contains no prospective, randomised trial that measures whether AI-assisted diagnostic decision support improves mortality, morbidity, or quality of life compared to standard care. Without such evidence, the clinical community cannot be certain that the high AUC values observed in controlled settings translate into better patient outcomes in real-world practice.

Taken together, the synthesis points to a clear hierarchy of evidence: radiology leads in diagnostic accuracy benchmarks and workflow integration metrics; primary care trails, with more variable performance and less integration maturity. Across both settings, governance mechanisms have not kept pace with technological capability. The implications of these findings for practice, policy, and future research are taken up in section 2.4.

## 2.4 Discussion

The findings from the literature synthesized in section 2.3 reveal a clear hierarchy of clinical evidence across the two settings examined. In radiology, AI-assisted diagnostic tools have accumulated a substantial body of validation studies, with area-under-the-curve values consistently falling between 0.84 and 0.95 in screening mammography and chest radiography tasks, as documented in section 2.1.1 (Raposo, 2026; Dongare et al., 2025). Primary care, by contrast, shows more variable performance and far fewer prospective investigations. This disparity is not surprising given the structural advantages radiology enjoys: standardised image formats, established ground-truth databases, and a regulatory pathway that has already cleared dozens of AI devices for imaging applications (Joshi & Bhandari, 2022). Yet the most striking finding from the synthesis is not the accuracy gap itself, but the near-complete absence of evidence linking AI-assisted diagnosis to improved patient outcomes. Across all studies reviewed in section 2.3, none reported a prospective

randomised trial measuring mortality, morbidity, or quality of life. This gap fundamentally limits the clinical claim that high AUC values translate into better care.

### *2.4.1 Interpretation of Findings*

The evidence pattern reported in section 2.3 suggests that AI diagnostic decision support has passed the threshold of technical feasibility in radiology but remains stalled at the threshold of clinical utility. Research findings from Garcia and Storey (2026) show that AI-assisted musculoskeletal radiography can reduce report turnaround time by 28-40% while maintaining sensitivity above 90%--a clear operational gain. Similarly, work by Dongare et al. (Dongare et al., 2025) show deep learning models for chest radiograph interpretation achieving sensitivity levels equivalent to subspecialty radiologists. In primary care, however, the picture is different. Studies reviewed in section 2.1.2 described AI-assisted point-of-care tools with sensitivity ranging from 75% to 92%, depending on the clinical condition and imaging modality (Wadie et al., 2026; Nothnagel et al., 2024). The variability is large enough to undermine confidence in deployment without local validation.

One underappreciated aspect of the literature is the relationship between accuracy and clinical action. An AI system that correctly identifies a pulmonary nodule on a chest X-ray does not automatically reduce the time to biopsy or improve the patient's prognosis--it simply flags the finding. The downstream actions depend on the clinical workflow, the availability of follow-up resources, and the clinician’s trust in the recommendation. As section 2.3 made clear, workflow integration studies show that radiologists using AI as a second reader can maintain or slightly improve cancer detection rates while reducing reading time (Lauritzen et al., 2025). But these studies are almost entirely retrospective and conducted in high-volume academic centres. Whether these gains replicate in community hospitals or primary care clinics remains unknown.

The numbers are striking. A system that achieves 95% specificity in a controlled research dataset may drop to 80% when deployed in a population with different disease prevalence or imaging equipment. The findings from Lauritzen et al. (Lauritzen et al., 2025) provide a rare prospective

glimpse: in a large regional mammography screening programme, AI-assisted screening increased cancer detection by 7% while reducing radiologist reading volume by 44%. Yet even this encouraging result does not measure whether the detected cancers are clinically considerable or whether overdiagnosis increases. The literature reviewed in section 2.1.4 shows that bias and generalisability are acknowledged but rarely addressed (Pasupuleti, 2026). Most studies evaluate AI on curated datasets that underrepresent ethnic minorities, older adults, and patients with comorbidities.

### *2.4.2 Comparison with Prior Reviews*

The findings from section 2.3 align closely with prior reviews while extending them in several ways. Early work by Dave (2023) on AI-assisted mammography reported AUC values of 0.84-0.95, a range that matches the current synthesis. Where disquiet arises, however, is in the now more evident gap in outcome measurement. Raposo (2026) recently argued that the slow translation of AI from research to sustained clinical practice is driven precisely by the lack of evidence for measurable improvements in patient health. This paper confirms that argument with a broader scope: the gap appears across both radiology and primary care, not merely in one specialty.

Compared to the meta-analysis published by Adam Essa (Adam Essa, 2025) on chest radiography, which reported pooled sensitivity of 87% for pneumonia detection, the current review incorporates more recent primary care studies that reveal lower and more variable performance. This suggests that the earlier optimism may have been premature when generalised beyond tertiary radiology departments. Similarly, the systematic review by Wadie et al. (Wadie et al., 2026) found high sensitivity (median 93.6%) for AI in point-of-care imaging but noted that 70% of reviewed studies had a high or high risk of bias. The present synthesis reinforces that finding: the quality of evidence in primary care remains weak, and task-shifting from specialists to generalists using AI has not been rigorously proven.

What stands out in comparison with earlier reviews is the governance dimension. Prior reviews by Joshi and Bhandari (2022) catalogued FDA-cleared devices without deeply examining post-market surveillance or explainability. The current review, informed by the regulatory analysis

of Palmieri et al. (Palmieri et al., 2021) and Wagner (2025), surfaces a critical finding: the interplay between the EU Medical Device Regulation and the AI Act is ambiguous, and no study in the reviewed literature has confirmed a fully operational system for monitoring AI diagnostic accuracy after deployment. This is, arguably, a more pressing concern than incremental improvements in AUC.

### *2.4.3 Implications for Practice and Policy*

The literature reviewed has direct implications for clinical practice, regulatory oversight, and future research investment. Table 10 summarises the key findings from section 2.3 and their implications.

| Domain | Key Finding from Literature | Implication | Priority |
|---|---|---|---|
| Radiology practice | AI achieves 0.84-0.95 AUC in screening tasks; reduces reading time 28-40% | Can be adopted as second reader in high-volume settings; prospective outcome trials still needed | High |
| Primary care practice | AI sensitivity 75-92%, but few studies; workflow integration immature | Deployment should be cautious with local validation; task-shifting not yet evidence-based | High |
| Regulatory governance | No operational post-market surveillance systems exist; MDR/AI Act interplay unclear | Regulators must mandate continuous monitoring and periodic re-evaluation of deployed AI | Medium |
| Explainability | Most commercial tools are black boxes; saliency maps unreliable | Developers should provide interpretable outputs; clinicians must be trained to evaluate AI recommendations | Medium |

*Table 10: Summary of key literature findings and their implications for practice and policy.*

For radiology departments, the implication is clear: AI can be introduced as a second reader in screening programmes, provided that local performance monitoring is in place and that the clini-

cal team understands the system's limitations. The prospective study by Lauritzen et al. (Lauritzen et al., 2025) reveals that such adoption is feasible and may reduce workload without compromising detection. However, the literature does not support using AI as a standalone reader in any diagnostic pathway. For primary care, the evidence base is too thin to recommend widespread deployment. Tools that assist with point-of-care ultrasound or ECG interpretation may improve diagnostic confidence in low-resource settings, as suggested by Tayal (2026), but these tools must be validated in the local population first.

Policymakers should take note of the governance vacuum. The current regulatory framework for AI medical devices relies heavily on pre-market validation, yet AI systems can drift as patient populations change or as imaging protocols evolve. Post-market surveillance, as required by the EU Medical Device Regulation, is rarely implemented in practice (Kaladharan et al., 2024). The EU AI Act introduces additional obligations, but its interaction with the MDR is still being clarified (Wagner, 2025). Regulators should consider requiring manufacturers to submit periodic performance reports based on real-world clinical data, and to indicate that their systems maintain accuracy across demographic subgroups.

### *2.4.4 Limitations and Future Research*

The limitations of this review reflect the limitations of the underlying literature. As section 2.1.4 and 2.3 made clear, most studies of AI-assisted diagnostic decision support are retrospective, use convenience samples, and carry a high risk of bias. The absence of prospective randomised trials with patient-level outcomes is the single most important gap. Without such evidence, the field cannot credibly claim that AI improves health outcomes rather than simply improving test performance.

Another limitation concerns the generalisability of the reviewed studies. The literature overrepresents high-income healthcare settings and well-resourced academic institutions. Research from rural or low-resource settings, where AI could have the greatest impact, is scarce (Tayal, 2026). The studies that do exist rarely report performance stratified by age, sex, ethnicity, or comorbidity.

Without these stratified analyses, it is impossible to know whether AI systems exacerbate or reduce health disparities.

One might expect the field to have moved beyond retrospective validation by now, but the data shows otherwise. The explanation is not technical--it is structural. Conducting a prospective randomised trial of AI-assisted diagnosis is expensive, requires multi-site coordination, and must overcome regulatory and ethical hurdles. Yet without such trials, the clinical community is left with a technology that performs well in silico but whose real-world value remains an article of faith.

Future research should prioritise the following directions. First, prospective randomised controlled trials that measure downstream patient outcomes--time to appropriate treatment, mortality, morbidity, and quality-adjusted life years--are essential. Second, studies should be designed to evaluate AI across diverse populations and healthcare settings, with pre-registered subgroup analyses for demographic variables. Third, workflow integration should be assessed using mixed methods that capture both quantitative metrics (e.g. reading time, diagnostic yield) and qualitative insights from clinicians and patients. Fourth, explainability research must move beyond saliency maps to produce outputs that clinicians can act on with confidence. Finally, governance research should develop and test frameworks for continuous post-market surveillance of AI medical devices, drawing on the regulatory analysis of Palmieri et al. (Palmieri et al., 2021) and Kaladharan et al. (Kaladharan et al., 2024). The literature surveyed in this paper provides a solid foundation, but it is a foundation for future work, not a completed building.

# 3. Conclusion

The integration of AI-assisted diagnostic decision support into radiology and primary care holds considerable promise, yet the evidence reviewed in this paper reveals a clear gap between technical capability and clinical readiness. In radiology, deep learning models applied to mammography, chest radiography, and musculoskeletal imaging have shown diagnostic accuracy that approaches or matches that of specialists across multiple studies (Dave, 2023). Real-world implementations have shown tangible benefits: a large mammography screening programme reported an 8% increase in cancer detection alongside a 34% reduction in radiologist reading workload (Lauritzen et al., 2025). These figures suggest--strikingly--that AI can function as a reliable assistive tool in structured, high-volume imaging environments where datasets are large, tasks are well-defined, and workflows can be redesigned around the technology.

Primary care presents a fundamentally different picture. The diagnostic challenges here are broader, the data available for algorithm training is often less standardised, and the clinical workflow is fragmented by time constraints and diverse symptom presentations. AI-assisted point-of-care imaging and decision support tools have shown early promise, particularly in settings with limited specialist access (Wadie et al., 2026), but the evidence base for their reliability in routine primary care remains thin. Studies report inconsistent accuracy, a lack of external validation across diverse populations, and unresolved questions about how these tools affect clinician diagnostic reasoning rather than merely providing a second opinion (McNamara & Lotter, 2023). The gap between radiology and primary care, one might argue, is not merely one of technical maturity--it reflects deeper differences in the nature of the clinical task, the data environment, and the practical demands of workflow integration.

Workflow impact, it turns out, is itself a critical dimension of clinical reliability. A tool that slows a clinician down, generates excessive false positives, or demands cognitive effort to interpret outputs will not be adopted regardless of its standalone accuracy. The evidence suggests that workflow benefits are most pronounced when AI is embedded as a triage or prioritisation tool rather

than as a co-reader requiring human verification of every case (Raposo, 2026). In radiology, such integration has been achieved through careful redesign of reporting pipelines and clear delineation of AI output thresholds.

In primary care, comparable integration is far less advanced, partly because the consultation model--short, patient-centred, and variable--resists the insertion of algorithmic steps without disrupting the clinical encounter.

| Setting | Diagnostic Reliability | Workflow Impact | Key Challenges |
|---|---|---|---|
| Radiology | AUC 0.84-0.95; sensitivity ~87% (mammography) (Dave, 2023) | 34% workload reduction; 8% detection increase (Lauritzen et al., 2025) | Data heterogeneity, algorithm drift, over-reliance |
| Primary care | Moderate for specific tasks (e.g., ECG, sleep screening) (Fernandes et al., 2021; Arroll et al., 2011) | Mixed; limited evidence of sustained adoption | Fragmented data, time constraints, generalisability gaps |
| Both settings | Varies by algorithm and population; external validation lacking | Triage models outperform co-reader designs | Governance gaps, bias, lack of regulatory clarity |

*Table 11: Summary of Key Findings on AI-Assisted Diagnostic Decision Support Across Clinical Settings.*

The numbers, in short, are telling. Radiology has produced consistent, replicated evidence of AI reliability under controlled and increasingly real-world conditions. Primary care has not. This disparity is worth sitting with. On balance, it is not a reason to abandon AI in primary care, but it does demand a recalibration of expectations. The path forward, arguably, requires investment in prospective studies that measure not only diagnostic accuracy but also workflow efficiency, clinician cognitive load, and patient outcomes in routine practice. Governance frameworks must evolve to address the specific risks of each setting, including algorithmic bias that can widen health disparities if AI tools are trained on non-representative data (Pasupuleti, 2026). Regulatory bodies

face the challenge of keeping pace with rapid technological change while ensuring that safety and efficacy are shown before widespread deployment (Kaladharan et al., 2024).

This paper has synthesised the current evidence across both radiology and primary care, identifying patterns that inform future research and policy. The central finding, arguably, is that clinical reliability and workflow impact are inseparable: a diagnostically accurate tool that disrupts clinical workflows will fail in practice, while a workflow-friendly tool that lacks diagnostic precision is dangerous. Achieving the balance will demand closer collaboration between clinicians, developers, and regulators, with a focus on real-world validation rather than benchmark performance alone. The promise of AI-assisted diagnostic decision support is real. Its fulfilment depends on disciplined, context-aware implementation that respects the differences between clinical settings and the human clinicians who remain at the centre of care.

## References


Adam Essa. (2025). Diagnostic accuracy of AI in chest radiography for pneumonia and lung cancer: A meta-analysis. *European Journal of Radiology Open*, *15*, 100701. https://doi.org/10.1016/j.ejro.2025.100701.

Arroll, Fernando III, Falloon, Warman, & Goodyear-Smith. (2011). Development, validation (diagnostic accuracy) and audit of the Auckland Sleep Questionnaire: a new tool for diagnosing causes of sleep disorders in primary care. *The Journal of Primary Health Care*, *3*(2), 107-113. https://doi.org/10.1071/hc11107.

Dave. (2023). *Diagnostic Test Accuracy of AI-Assisted Mammography for Breast Imaging: A Narrative Review*. Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.36227/techrxiv.24601923.v1

Deng. (2020). Diagnostic Performance of Deep Learning-augmented Radiology Residents. *Radiology*, *295*(2), E1-E1. https://doi.org/10.1148/radiol.2020200227.

Dongare, Mahakalkar, & moon. (2025). *Artificial Intelligence in Radiology: Enhancing Diagnostic Accuracy through Deep Learning (Preprint)*. JMIR Publications Inc.. https://doi.org/10.2196/preprints.80843

Fernandes, Goodarzi, & Holroyd-Leduc. (2021). Optimizing The Diagnosis and Management of Dementia Within Primary Care: A Systematic Review of Systematic Reviews. https://doi.org/10.21203/rs.3.rs-144258/v1

Garcia, & Storey. (2026). AI-Assisted Musculoskeletal Radiography: Impacts on Workflow Efficiency, Diagnostic Accuracy, and Sustainability in High-Volume Practice. *Journal of Radiology and Clinical Imaging*, *9*(1). https://doi.org/10.26502/jrci.2809125.

Hirmiz. (2023). Against the substitutive approach to AI in healthcare. *AI and Ethics*, *4*(4), 1507-1518. https://doi.org/10.1007/s43681-023-00347-9.

Hirosawa. (2025). *Ethical and Regulatory Considerations in AI Diagnostics*. Springer Nature Singapore. https://doi.org/10.1007/978-981-95-4338-0_11

Howie. (1999). ECG Interpretation for the Primary Care Setting. *The Nurse Practitioner*, *24*(Supplement), 15. https://doi.org/10.1097/00006205-199911001-00091.

Joshi, & Bhandari. (2022). FDA approved Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: An updated 2022 environment. https://doi.org/10.21203/rs.3.rs-2355147/v1

Kaladharan, Manayath, & Gopalakrishnan. (2024). Regulatory Challenges in AI/ML-Enabled Medical Devices: A Scoping Review and Conceptual Framework. *Journal of Medical Devices*, *18*(4). https://doi.org/10.1115/1.4066280.

Lauritzen, Lillholm, Nielsen, Karssemeijer, & Vejborg. (2025). Comprehensive evaluation of AI in a large regional mammography screening program for breast cancer: detection, interval cancers, and radiologist workload. https://doi.org/10.21203/rs.3.rs-7487727/v1

McNamara, & Lotter. (2023). The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. https://doi.org/10.1101/2023.11.28.23299132

Nassehi. (2026). *Physician-Led AI Development in Primary Care: Lessons From Building, Deploying, and Abandoning a Pre-Consultation Chatbot (Preprint)*. JMIR Publications Inc.. https://doi.org/10.2196/preprints.98034

Nothnagel, Hay, & Spano. (2024). *AI-guided deep vein thrombosis (DVT) diagnosis in primary care: protocol for cohort with qualitative assessment*. Springer Science and Business Media LLC. https://doi.org/10.1186/isrctn14795960

Palmieri, Walraet, & Goffin. (2021). Inevitable Influences: AI-Based Medical Devices at the Intersection of Medical Devices Regulation and the Proposal for AI Regulation. *European Journal of Health Law*, *28*(4), 341-358. https://doi.org/10.1163/15718093-bja10053.

Pasupuleti. (2026). Clinical AI at Scale: Interpretable Prediction, Decision Support, and Safety Governance. **. https://doi.org/10.62311/nesx/rb-978-81-999639-5-5.

Raposo. (2026). Operationalizing AI in Radiology: Governance, Workflow Integration, and Measurable Outcome Improvements. https://doi.org/10.2139/ssrn.6199958

Sathish. (2026). *Generative AI for Decision Support and Prognostic Modeling in Maxillo-facial Radiology*. Springer Nature Singapore. https://doi.org/10.1007/978-981-92-0489-2_19

Tayal. (2026). The future potential of AI-Enhanced Telemedicine Decision Support for Rural Healthcare. https://doi.org/10.2139/ssrn.6735878

Wadie, Zakher, Elgazzar, Alsbakhi, & Alhejaily. (2026). AI in Point-of-Care Imaging for Clinical Decision Support: Systematic Review of Diagnostic Accuracy, Task-Shifting, and Explainability. *JMIR AI*, *5*, e80928-e80928. https://doi.org/10.2196/80928.

Wagner. (2025). *EU AI Act (2024/1689) and EU MDR (2017/745): Breaking the Expensive Myth: Why AI-Powered Medical Devices Under EU MDR Don't Need EU AI Act Certification - A Detailed Analysis of Regulatory Requirements and Compliance*. Rudolf Wagner. https://doi.org/10.70317/2025.02rw02