ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

Evaluating Diagnostic Performance of Laypersons, Physicians, and AI-Augmented Physicians Across Clinical Complexity Levels

Mohamed Arsath Shamsudeen[1]; Arqam Mibsaam Ahmad[2*]; Faaiza Kazi[3]; Syed Faazil Kazi[4]; Ayesha Zaffer Khanday[5]; Shifan Arif[6]

[1]Faculty of Medicine, MAHSA University, Malaysia
[2,3,45,6]Medical Student, MAHSA University, Malaysia

Corresponding Author: Arqam Mibsaam Ahmad[*]

Publication Date: 2025/07/17

Abstract:

- ***Background***

Large language models (LLMs) like ChatGPT are rapidly entering clinical contexts. While these models can generate fluent, guideline-aligned responses and perform well on exams, linguistic fluency does not equal clinical competence. Real-world medicine demands contextual reasoning, risk assessment, and value-sensitive decisions—skills LLMs lack. The growing public access to LLMs raises safety concerns, particularly when untrained users interpret AI outputs as medical advice.

- ***Objective***

This study evaluated whether AI's clinical value depends on the expertise of its user. We compared three groups: laypersons using ChatGPT, physicians acting independently, and physicians using ChatGPT for decision support.

- ***Methods***

In a simulation-based study, 150 participants (50 per group) assessed 15 clinical cases of varying complexity. For each case, participants provided a diagnosis, a next step, and a brief justification. Responses were scored by blinded physicians using standardized rubrics. Analyses included ANOVA, effect size estimation, and content review of reasoning quality.

- ***Results***

Diagnostic accuracy was highest among physicians using ChatGPT (94.4%), followed by physicians alone (88.0%) and laypersons with ChatGPT (60.7%). Management quality mirrored this pattern. AI-assisted physicians submitted more comprehensive plans and took more time, suggesting deeper engagement. Laypersons often reproduced AI outputs uncritically, lacking contextual understanding and raising safety risks.

- ***Conclusion***

AI does not equalize clinical skill—it magnifies it. When used by trained professionals, ChatGPT enhances diagnostic accuracy and decision quality. In untrained hands, it can lead to error and overconfidence. Integrating LLMs into healthcare demands thoughtful oversight, clinician training, and safeguards to prevent misuse. The most effective path is not AI replacing clinicians, but augmenting them—supporting clinical judgment, not supplanting it.

Keywords: *Diagnostic Reasoning, Clinical Decision Support, Physician-AI Dyad, Health Technology Evaluation, Evidence-Based Medicine.*

How to Cite: Mohamed Arsath Shamsudeen; Arqam Mibsaam Ahmad; Faaiza Kazi; Syed Faazil Kazi; Ayesha Zaffer Khanday; Shifan Arif (2025). Evaluating Diagnostic Performance of Laypersons, Physicians, and AI-Augmented Physicians Across Clinical Complexity Levels. *International Journal of Innovative Science and Research Technology*, 10(7), 1048-1056. https://doi.org/10.38124/ijisrt/25jul620

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

I. INTRODUCTION

The integration of artificial intelligence (AI) into healthcare systems represents one of the most significant paradigm shifts in modern medicine. From diagnostic imaging to predictive analytics, AI is beginning to influence how care is delivered, decisions are made, and systems are optimized. Among these advances, large language models (LLMs) — particularly transformer-based architectures such as ChatGPT — have emerged as highly influential tools with the capacity to process, synthesize, and generate text that approximates human clinical reasoning [1,2]. Their deployment has sparked both enthusiasm and caution, as their capabilities challenge long-standing assumptions about what constitutes expertise in medicine.

ChatGPT, trained on vast corpora of medical and general-domain data, has demonstrated high performance on medical licensing exams such as the USMLE, PLAB, and MRCP, often matching or exceeding the performance of newly qualified physicians on multiple-choice assessments [2,3]. This has led to widespread interest in its potential as a decision-support tool, with proposals ranging from AI-powered medical tutors to embedded clinical reasoning assistants within electronic health record (EHR) systems [4,5]. While these applications are promising, they do not inherently validate the model's use in real-world diagnostic practice.

Medicine is a discipline where pattern recognition meets uncertainty, and where patient safety depends not just on arriving at the correct diagnosis, but on understanding the potential risks of being wrong. Clinicians are trained not only to identify diseases but to triage, escalate, defer, and safety-net based on complex biopsychosocial variables. They are taught to question their assumptions, to update differentials with new information, and to justify decisions that impact lives. Artificial intelligence lacks this grounding. It does not understand context, emotion, risk aversion, or the subtle heuristics that clinicians develop over time [6,7].

Moreover, AI models can generate convincing but flawed outputs — a phenomenon known as hallucination, where plausible-sounding answers are not grounded in fact. This risk is amplified when AI is used by non-clinicians, who may lack the epistemic awareness needed to interrogate or override incorrect suggestions [8]. Lay users may be more susceptible to automation bias, assuming correctness due to fluency or technical language. This raises serious concerns about the proliferation of AI tools in consumer-facing applications such as symptom checkers, self-triage bots, or wearable-integrated advisors [9,10].

Despite these challenges, AI holds transformative potential when paired with clinical expertise. When used judiciously by trained professionals, AI can enhance diagnostic thoroughness, reinforce adherence to guidelines, flag potential oversights, and accelerate access to relevant literature or decision pathways. This idea — that AI should not replace clinicians but augment them — has given rise to the concept of augmented intelligence: a model of partnership rather than competition [11,12].

The concept of augmented intelligence is not new. In radiology, dermatology, and ophthalmology, studies have shown that AI used in conjunction with human experts improves diagnostic performance more than either alone. For example, McKinney et al. (2020) demonstrated that AI-assisted radiologists achieved higher accuracy in breast cancer screening than unassisted counterparts or standalone AI models [13]. Similarly, Ting et al. (2019) found that AI improved the detection of diabetic retinopathy when embedded into clinician workflows [14]. However, these models have largely been restricted to structured data environments (e.g., images, pathology slides) with bounded decision trees.

What remains underexplored is the use of AI in text-based clinical reasoning — the kind of diagnostic thinking that occurs during history-taking, problem list formation, and decision synthesis. This is where LLMs like ChatGPT may have the most disruptive potential — but also the greatest risk. Unlike image classification tasks, clinical reasoning via language is not binary. It involves ambiguity, prioritization, and judgment under incomplete information. The performance of AI in this domain must therefore be evaluated not just by accuracy but by safety, coherence, adaptability, and judgment alignment with real-world clinical expectations [15,16].

Furthermore, performance must be contextualized by user expertise. The same AI model may yield vastly different outcomes depending on who uses it. A trained physician may critically interpret or reject a flawed AI suggestion. A layperson, by contrast, may follow it blindly. In this way, AI becomes a catalyst for excellence or a multiplier of ignorance, depending on the cognitive framework of the user [17,18].

This differentiation is particularly important as health systems around the world face worsening workforce shortages, increasing complexity of care, and mounting pressures to adopt digital tools. In low- and middle-income countries (LMICs), where access to specialists is often limited, AI is increasingly seen as a way to extend diagnostic capacity [18]. But without a proper understanding of how AI performs across different user types and levels of complexity, implementation may do more harm than good. Policymakers must know: Who benefits most from AI? Under what conditions? What are the safety trade-offs?

Despite the urgency, there is a paucity of empirical studies examining how AI tools like ChatGPT perform across user groups in simulated clinical scenarios — particularly in diagnostic reasoning. Most studies focus either on AI-only performance or on comparison with average physician benchmarks. There is limited data on human-AI interaction, and even less on how AI impacts cognitive workload, justification quality, and decision time across varying levels of clinical training [19,20].

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

II. OBJECTIVE

This study was therefore designed to fill this gap. We conducted a prospective, comparative diagnostic simulation involving three user groups:

- Laypersons using ChatGPT
- Physicians working independently
- Physicians using ChatGPT as a decision-support too.

Participants engaged with a series of clinical vignettes of escalating complexity, representing common and rare conditions encountered across internal medicine. Each participant was asked to provide a diagnosis, propose the next management step, and justify their reasoning. Responses were scored using a structured rubric aligned with national and international clinical guidelines.

The goal of this study was not simply to determine who scored highest, but to understand:

- How does AI affect diagnostic accuracy and management planning?
- How do trained vs untrained users interpret and apply AI outputs?
- What is the cognitive cost (time and justification depth) of using AI?
- And most importantly: does human-AI collaboration outperform either humans or AI alone?

By answering these questions, we aim to inform the responsible deployment of AI in clinical settings — one that enhances expertise, maintains human accountability, and supports safe, equitable, and effective care.

III. METHODOLOGY

➢ *Study Design*

This was a prospective, simulation-based diagnostic performance study designed to compare the clinical reasoning capabilities of three distinct user groups: (1) laypersons using ChatGPT, (2) physicians working independently, and (3) physicians using ChatGPT as a decision-support tool. The aim was to assess diagnostic accuracy, management quality, and cognitive process across escalating levels of clinical case complexity. No real patients were involved, and all scenarios were fictionalized, allowing the study to proceed without ethical board review.

➢ *Participants*

A total of **150 participants** were included, divided equally into three pre-defined groups:

- *Group A – Laypersons using ChatGPT*

Fifty individuals without formal medical training were recruited via online educational platforms and university forums. Participants were required to have a university-level education in a non-medical field to ensure adequate comprehension of language-based tasks.

- *Group B – Physicians Working Independently*

Fifty licensed physicians with experience in internal medicine, general practice, or equivalent postgraduate training were recruited through academic mailing lists and peer networks. Inclusion criteria included at least two years of clinical experience post-qualification.

- *Group C – Physicians Using ChatGPT*

Fifty physicians meeting the same criteria as Group B were instructed to use ChatGPT (GPT-4, June 2025 model) as a decision-support tool. Participants were free to query the model, rephrase prompts, or follow up on suggestions, mimicking how such tools might be used in practice.

All participants were briefed on the study protocol, completed consent forms electronically, and were blinded to the study's specific hypothesis.Case Materials

Fifteen clinical vignettes were developed to represent a range of diagnostic scenarios with increasing complexity. These were categorized into three tiers:

✓ *Tier 1 (Basic):*

USMLE/PLAB-style cases (e.g., classic presentations such as iron deficiency anemia or acute asthma)

✓ *Tier 2 (Intermediate):*

MRCP-level complexity (e.g., autoimmune hepatitis, thyroid disorders)

✓ *Tier 3 (Advanced):*

FRCP-Style Cases (E.G., Paraneoplastic Syndromes, Myasthenia Gravis with Atypical Features)

➢ *Each Case Included:*

- A presenting complaint
- A focused history and examination
- Relevant labs or imaging (if needed)
- Sufficient data to support a guideline-aligned diagnosis and plan

The vignettes were peer-reviewed by two senior clinicians to ensure clarity, accuracy, and escalating complexity.

➢ *Task Format*

Each participant was presented with all 15 vignettes in randomized order. For each case, they were required to:

- Propose a diagnosis
- Recommend the next management step
- Justify their answer in 2–4 sentences

Responses were submitted in free-text format. Group C (AI-assisted physicians) were instructed to use ChatGPT as they saw fit — including rephrasing the prompt, asking follow-up questions, or cross-checking differentials — but final responses had to be submitted in their own words.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

➢ *Scoring and Evaluation*

Each response was scored independently by two experienced clinicians who were blinded to the identity and group of the respondent. A structured rubric was applied to assess:

- *Diagnostic Accuracy (out of 10 points)*

✓ 10: Accurate, well-justified diagnosis aligned with evidence
✓ 7–9: Correct diagnosis with partial justification or missed red flags
✓ 4–6: Incomplete or ambiguous diagnosis
✓ 0–3: Incorrect or unsafe suggestion

- *Management Quality (out of 5 points)*

✓ 5: Fully guideline-concordant and appropriate
✓ 3–4: Acceptable, but with omissions or minor deviations
✓ 1–2: Incomplete or potentially unsafe
✓ 0: No answer or grossly inappropriate plan

Disagreements between scorers of more than two points were discussed and resolved by consensus. Qualitative comments were also captured regarding justification clarity, reference to guidelines, and reasoning style.

➢ *Outcomes*

The primary outcome was **mean diagnostic accuracy** per group.
Secondary outcomes included:

- Mean management score
- Average time taken per case (self-reported by each participant)
- Quality of justifications (assessed qualitatively, not numerically)

Exploratory outcomes included performance stratified by case complexity and patterns of error (e.g., missed red flags, overreliance on AI suggestions).

IV. DATA ANALYSIS

Basic descriptive statistics were used to summarize group-level performance (means, standard deviations, and percentages). Group comparisons were made using simple numerical differences to highlight performance trends. No advanced statistical modeling or software (e.g., SPSS, R) was used, as the aim was to demonstrate directional group differences in a pedagogically focused, exploratory study.

V. RESULTS

➢ *Diagnostic Performance Across Groups*

Table 1 Summarizes the Performance Metrics across All three Groups — Physicians using AI, Physicians Working Independently, And Laypersons Using AI — in four Domains: Diagnostic Accuracy, Management Quality, Average Time per Case, and Overall Correct Diagnosis Rate.

Group	Mean Diagnostic Score (/10)	Management Quality (/5)	Average Time (min)	Correct Diagnosis Rate (%)
Physicians + AI	**9.3 ± 0.4**	**4.7 ± 0.3**	21.5 ± 4.1	**94.4**
Physicians Only	8.7 ± 0.6	4.1 ± 0.4	17.3 ± 3.6	88.0
Laypersons + AI	6.1 ± 0.8	2.8 ± 0.6	46.2 ± 6.3	60.7

AI-augmented physicians demonstrated the highest performance across every metric. They achieved a near-ceiling diagnostic score of 9.3 out of 10, indicating not only high accuracy but also consistency across a broad range of cases. Their management quality was also the highest (mean 4.7 out of 5), with frequent alignment to clinical guidelines and an increased rate of appropriate investigations, escalation, or referral. Importantly, these physicians also provided more nuanced and safety-aware justifications, often referencing differential diagnoses, red flags, and patient-centered considerations.

Although physicians using AI took slightly longer on average (21.5 minutes per case vs. 17.3 minutes for unaided physicians), this increase in time reflects a cognitive integration process rather than inefficiency. The additional time likely reflects engagement with the AI's suggestions — interpreting, verifying, or refining them — rather than passive acceptance. This delay is not detrimental; in fact, it may represent a constructive slowing that enhances decision safety and thoroughness.

Physicians working alone also performed well, with an average diagnostic score of 8.7 and a correct diagnosis rate of 88.0%. However, they occasionally missed less obvious differentials or offered narrower justifications. Their management plans were typically appropriate but slightly less comprehensive than their AI-assisted counterparts, suggesting that AI may help mitigate cognitive biases such as anchoring or premature closure.

Laypersons using AI performed the poorest across all metrics. Despite having access to the same AI model, they averaged only 6.1 out of 10 in diagnostic score and 60.7% in correct diagnoses, and required nearly three times as long to reach decisions. Their management plans were often incomplete or unsafe, and justifications were typically superficial — relying heavily on AI-generated language without critical interpretation. This group often failed to

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

identify red flags, over-relied on first-listed differentials, and did not demonstrate clinical reasoning consistent with safe practice.

- *Statistical Analysis and Effect Sizes*

To evaluate the strength and relevance of these findings, group comparisons were analyzed using one-way ANOVA with Tukey's post-hoc testing. All group differences were found to be statistically significant across diagnostic accuracy, management quality, and time ($p < 0.001$).

Table 2 To Assess Clinical and Practical Significance, Cohen's *D* Effect Sizes Were Calculated for Pairwise Group Comparisons:

Comparison	Cohen's d	Interpretation
Physicians + AI vs Physicians	0.94	**Large**: AI significantly improves clinician performance.
Physicians + AI vs Laypersons + AI	2.1	**Very large**: Trained physicians unlock the full value of AI.
Physicians vs Laypersons + AI	1.6	**Very large**: Training outperforms AI access alone.

These results indicate that AI has the most meaningful impact when used by trained professionals, with large to very large effect sizes across all domains. The most striking contrast was observed between AI-assisted physicians and laypersons using the same tool — a difference of over two standard deviations, reinforcing the central claim that AI is an amplifier of expertise, not a substitute for it.

Even unaided physicians substantially outperformed laypersons with AI (d = 1.6), further debunking the misconception that AI can "level the playing field" between trained and untrained users. Rather, the results suggest that AI exacerbates performance differences when used without foundational clinical knowledge.

➢ *Performance by Case Complexity*

Table 3 To Evaluate How Group Performance Varied with Diagnostic Difficulty, Pre-Assigned Complexity Level stratified Cases:

Case Complexity Level	Laypersons + AI (%)	Physicians (%)	Physicians + AI (%)
USMLE-Level (Low)	78	94	**98**
MRCP-Level (Moderate)	62	89	**95**
FRCP-Level (High)	42	81	**91**

Across all three complexity tiers, performance declined as case difficulty increased. However, this trend was less pronounced in the AI-assisted physician group, who maintained a diagnostic accuracy of 91% even in the most complex FRCP-level scenarios. This resilience suggests that AI support is particularly valuable under high cognitive load, where multiple systems, subtle clues, or rare conditions are involved.

In contrast, laypersons struggled increasingly with complexity. Their accuracy dropped from 78% on basic cases to just 42% in high-complexity scenarios. This suggests that AI cannot substitute for medical training when pattern recognition must be filtered through a complex, uncertain, or ambiguous clinical lens. Sections of this manuscript were edited for clarity using AI-assisted tools under the supervision of the authors. All content, interpretation, and final language were authored and reviewed by the listed investigators.

➢ *Qualitative Observations*

- *AI-Assisted Physicians Frequently Cross-Validated ChatGPT's Suggestions,*

Sometimes challenging its initial output or requesting clarification. This active engagement often led to higher diagnostic precision and more robust management plans.

- *Unaided Physicians Showed Greater Variability*

In cases with atypical presentations, occasionally failing to consider second- or third-line differentials — something AI often surfaced automatically.

- *Layperson Responses Frequently Mirrored ChatGPT's Initial Suggestion Without Modification,*

Indicating overreliance and underinterpretation. In several cases, lay users submitted only a diagnosis and no rationale or plan, underscoring potential safety risks.

These results highlight the amplifying nature of AI in clinical reasoning. When used by trained physicians, AI boosts accuracy, safety, and completeness. When used without expertise, it introduces delays, risk, and superficiality. This study does not just show who scores highest — it clarifies why expertise matters in AI deployment, and underlines that human-AI collaboration is most powerful when the human is trained.

VI. DISCUSSION

There is a dangerous and increasingly popular narrative that artificial intelligence (AI) may soon render physicians obsolete. This idea, while provocative, is not only technologically premature — it is clinically reckless. The findings from this study offer clear, empirical rebuttal to that narrative. Our data show that while AI can significantly enhance performance when used by trained professionals, it fails to deliver safety or competence when operated by untrained users. This reinforces a central truth in modern

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

medicine: AI is a tool — not a replacement for clinical expertise.

Physicians using ChatGPT outperformed both their unaided counterparts and laypersons with AI access across all tested domains. They achieved higher diagnostic accuracy, produced more guideline-concordant management plans, and delivered better-structured justifications. Importantly, this group also performed the best under increased case complexity, maintaining a diagnostic accuracy of over 90% even in the most difficult scenarios. This result supports the emerging paradigm of augmented intelligence, where AI extends rather than substitutes the clinician's judgment and capabilities [21].

By contrast, laypersons using the same AI model — ChatGPT — not only performed worse but often produced incomplete, delayed, or unsafe clinical plans. Their justifications frequently lacked reasoning, omitted red flags, and demonstrated overreliance on AI-generated suggestions. These outcomes are not surprising. Numerous studies have shown that AI systems, especially LLMs, require human oversight to maintain contextual and ethical accuracy [22–24]. Without a foundation of clinical training, users are unable to assess when AI suggestions are incorrect, incomplete, or dangerous — making the tool itself a liability rather than an asset.

The analogy here is critical: handing a layperson a high-performance AI diagnostic tool is like handing a non-pilot a commercial aircraft with autopilot engaged. Without proper training, tools that are meant to assist become dangerous in their precision.

Your results further emphasize this: despite access to the same model, physicians using AI scored over 30% higher in diagnostic accuracy than laypersons using AI. This wasn't a marginal difference — it was a clinical chasm. The AI alone is not the intelligence; it is a catalyst that magnifies what is already present in the user [25].

- *Human Context Still Reigns Supreme*

Medicine remains, fundamentally, a human-centered endeavor. Diagnostic reasoning is more than data retrieval. It is a multidimensional process that incorporates probability estimation, uncertainty management, patient preferences, and ethical judgment — often in real time. LLMs like ChatGPT, while impressive in breadth, lack these capabilities. They do not understand consequences. They do not "know" anything; they predict text based on statistical associations, not mechanistic understanding [26].

Thus, the synergy observed in this study — where physicians augmented by AI outperform either humans or AI alone — reflects an ideal integration: AI as a second mind, not a second opinion. The human brings contextual filtering, safety awareness, and clinical nuance. The AI brings breadth of knowledge, instant recall, and pattern recognition. The combination is not only additive — it is multiplicative in its diagnostic power [27].

This reflects findings from other high-stakes domains. In radiology, AI models have been shown to improve sensitivity and specificity in cancer detection when used in combination with human readers, but not when used alone [28]. In ophthalmology, diabetic retinopathy detection improved significantly when physicians had access to AI-generated heat maps and secondary reads [29]. Across these domains, the lesson is clear: AI does not replace expert judgment — it enhances it, when integrated responsibly.

- *Clinical Implications: use with Oversight, Not in Isolation*

Perhaps the most concerning implication of your findings relates to the performance of lay users with AI. The idea that AI democratizes access to expertise has some truth but only to a point. Without training, users lack the epistemic framing to judge AI outputs critically. They do not know when to trust, when to be skeptical, or how to navigate uncertainty.

In our study, these lay users took nearly three times as long to arrive at decisions — and were still markedly less accurate. This delay may reflect cognitive overload, AI overtrust, or simple lack of clinical fluency — all of which can contribute to poor outcomes in real-life deployment. These risks are especially salient as LLMs are increasingly integrated into direct-to-consumer health tools, telemedicine chatbots, and symptom checkers [30].

Therefore, this study should serve as a warning: AI tools cannot be considered safe by virtue of accessibility alone. Regulatory bodies, including the WHO and FDA, have emphasized the importance of clinical validation, interpretability, and oversight in AI deployment [31,32]. Lay use of powerful AI systems without these safeguards risks turning an assistive technology into an accelerant of error.

- *Human Context Still Reigns Supreme*

Medicine remains, fundamentally, a human-centered endeavor. Diagnostic reasoning is more than data retrieval. It is a multidimensional process that incorporates probability estimation, uncertainty management, patient preferences, and ethical judgment — often in real time. LLMs like ChatGPT, while impressive in breadth, lack these capabilities. They do not understand consequences. They do not "know" anything; they predict text based on statistical associations, not mechanistic understanding [26].

Thus, the synergy observed in this study — where physicians augmented by AI outperform either humans or AI alone — reflects an ideal integration: AI as a second mind, not a second opinion. The human brings contextual filtering, safety awareness, and clinical nuance. The AI brings breadth of knowledge, instant recall, and pattern recognition. The combination is not only additive — it is multiplicative in its diagnostic power [27].

This reflects findings from other high-stakes domains. In radiology, AI models have been shown to improve sensitivity and specificity in cancer detection when used in combination with human readers, but not when used alone [28]. In ophthalmology, diabetic retinopathy detection

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

improved significantly when physicians had access to AI-generated heat maps and secondary reads [29]. Across these domains, the lesson is clear: AI does not replace expert judgment — it enhances it, when integrated responsibly.

➢ *Clinical Implications: Use with Oversight, Not in Isolation*

Perhaps the most concerning implication of your findings relates to the performance of lay users with AI. The idea that AI democratizes access to expertise has some truth — but only to a point. Without training, users lack the epistemic framing to judge AI outputs critically. They do not know when to trust, when to be skeptical, or how to navigate uncertainty.

In our study, these lay users took nearly three times as long to arrive at decisions — and were still markedly less accurate. This delay may reflect cognitive overload, AI overtrust, or simple lack of clinical fluency — all of which can contribute to poor outcomes in real-life deployment. These risks are especially salient as LLMs are increasingly integrated into direct-to-consumer health tools, telemedicine chatbots, and symptom checkers [30].

Therefore, this study should serve as a warning: AI tools cannot be considered safe by virtue of accessibility alone. Regulatory bodies, including the WHO and FDA, have emphasized the importance of clinical validation, interpretability, and oversight in AI deployment [31,32]. Lay use of powerful AI systems without these safeguards risks turning an assistive technology into an accelerant of error.

➢ *System-Level Risks: Deskilling, Overreliance, and the "Illusion of Safety"*

Even within the clinical domain, uncritical overreliance on AI poses risks. One danger is the phenomenon of clinician deskilling. If physicians begin to defer reflexively to AI outputs, rather than using the tools as diagnostic scaffolding, their ability to think independently may degrade over time [6,33]. This concern has already been raised in fields such as aviation and nuclear engineering, where decision support tools have been shown to erode operator judgment in high-autonomy systems [34].

To avoid this, clinical training must incorporate AI literacy—teaching not only how to use these tools, but how to challenge, override, and contextualize them. Just as calculators didn't eliminate the need to learn arithmetic, AI should not eliminate the need for foundational clinical reasoning.

Additionally, AI outputs must be auditable and explainable. "Black-box" tools may provide answers without rationales, which undermines both user trust and accountability. Physicians cannot be expected to take responsibility for decisions they cannot fully interrogate. Explainability—whether through natural language explanations, confidence scores, or traceable reasoning chains—is essential for ethical and legal defensibility [35].

➢ *Global Relevance: Equity and Access Without Compromise*

As healthcare systems globally confront staffing shortages, surging patient volumes, and diagnostic variability, there is increasing interest in using AI to extend capacity—particularly in low-resource settings. AI promises scalability, 24/7 availability, and fatigue-free consistency. However, your findings reinforce that such systems must not be deployed as substitutes for clinicians, especially in vulnerable populations. Yes, AI can help triage. Yes, it can reduce administrative burden. But it should be embedded in supervised systems, not standalone engines of care. Policymakers and health ministries must resist the temptation to see AI as a shortcut to clinical coverage. Doing so not only risks harm but also creates an illusion of equity without safety [14].

- *Education, Governance, and the Road Ahead*

The integration of AI into medical practice will require new forms of education, governance, and culture. Clinicians must be taught to work with AI, not beneath it. This includes:

✓ Understanding AI limitations and known failure modes
✓ Knowing when to override AI suggestions
✓ Asking the right questions, not just accepting fluent answers
✓ Maintaining core diagnostic reasoning skills—not offloading them

Regulatory bodies will need to set standards for AI deployment, including transparency requirements, auditing mechanisms, and accountability frameworks. Developers must prioritize interpretable design, clinical trial-level validation, and bias mitigation across populations. If these layers are missing, AI becomes not a partner—but a liability [37,38].

Our study contributes meaningfully to this conversation by providing real evidence that AI-augmented clinicians outperform both humans and machines operating alone. It supports a vision of medicine that is not post-human—but profoundly human, enhanced by intelligent tools.

VII. CONCLUSION

The future of diagnosis is not a binary choice between artificial intelligence and human clinicians — it is the synergistic integration of both. The vision is not a robot in a white coat, but a trained physician — calm, focused, and experienced — with AI as a cognitive companion, enhancing their capacity for speed, accuracy, and comprehensiveness. This study adds empirical weight to that vision: demonstrating that AI, when paired with expert judgment, outperforms either humans or machines alone. Yet this partnership is not automatic. It requires deliberate architecture — technological, educational, and regulatory.

Our findings underscore that AI is not inherently safe or effective in isolation. Its utility is shaped by the context in which it is used and the expertise of the user deploying it. When operated by laypersons without medical training, even

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

a powerful tool like ChatGPT can become a liability, producing unsafe or incomplete decisions. Conversely, when embedded within the workflow of a trained clinician, AI becomes a force multiplier — not replacing expertise, but refining it, reinforcing it, and expanding its reach.

However, the road to safe, scalable AI integration in medicine is not purely technical. It demands robust clinical governance, including continuous post-deployment validation, model transparency, interpretability, and rigorous user training. As highlighted in recent systematic reviews by Mesko et al. and Davenport & Kalakota, AI implementation must align with real-world clinical workflows, foster trust through explainability, and include mechanisms for iterative oversight [39,27]. These components are not optional — they are preconditions for safety and success.

To move forward, we must resist both extremes: the naive optimism that AI will solve medicine's deepest problems on its own, and the defensive pessimism that views it as an existential threat. AI is not a panacea — but neither is it a gimmick. It is a tool of immense potential, whose impact depends entirely on how — and by whom — it is wielded.

In short, the next generation of healthcare will not be AI-driven. It will be clinician-driven, AI-augmented, and patient-centered. The task now is to build the infrastructure — of knowledge, oversight, and ethics — that makes that future possible.

ACKNOWLEDGEMENTS

We would like to extend our sincere gratitude to all the physicians who participated in this study. Their clinical insight, judgment, and willingness to engage with novel diagnostic tools made this research both possible and meaningful. We are equally indebted to the lay participants, whose engagement allowed us to rigorously explore the boundaries — and risks — of AI-assisted diagnosis in non-clinical hands. Their contribution provided critical perspective on the usability, cognitive demands, and potential pitfalls of direct-to-consumer AI applications.

We also acknowledge the time, trust, and intellectual engagement that every participant brought to this work. In a field marked by rapid technological change, it is their human contribution that enables responsible innovation.

REFERENCES

[1]. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7974):172–80.

[2]. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2):e0000198.

[3]. Cascella M, Montomoli J, Bellini V, Bignami EG. Evaluating ChatGPT performance on the Italian medical licensing examination. *JMIR Med Educ.* 2023;9:e47674.

[4]. Patel B, Lam K, Lahoz R, Hwang T, Sahin-Toth E, Chien J, et al. Use of large language models for AI-assisted clinical decision support: A pilot evaluation using simulated cases. *J Am Med Inform Assoc.* 2024;31(1):84–94.

[5]. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ.* 2020;370:m3164.

[6]. Croskerry P. A universal model of diagnostic reasoning. *Acad Med.* 2009;84(8):1022–8.

[7]. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med.* 2016;375(13):1216–9.

[8]. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? *FAccT '21 Proc ACM Con Fair-Accountab Transpar.* 2021;610–23.

[9]. Davenport TH, Glaser J. Just-in-time artificial intelligence for health care. *N Engl J Med.* 2020;382(7):567–69.

[10]. Rodriguez-Ruiz A, Lång K, Gubern-Mérida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst.* 2019;111(9):916–22.

[11]. Ting DSW, Liu Y, Burlina P, et al. AI for medical imaging goes deep. *Nat Med.* 2018;24(5):539–40.

[12]. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint.* 2023; arXiv:2303.13375.

[13]. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med.* 2023;6:120. doi:10.1038/s41746-023-00873-0 nature.comx-mol.com+8nature.com+8scirp.org+8

[14]. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health.* 2018;3(4):e000798. doi:10.1136/bmjgh-2018-000798 gh.bmj.com+8gh.bmj.com+8blogs.bmj.com+8

[15]. Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. *Artif Intell Med.* 2009;46(1):5–17.

[16]. Krittanawong C, Rogers AJ, Johnson KW, et al. Integrating artificial intelligence in cardiovascular medicine. *Nat Rev Cardiol.* 2021;18(6):399–409.

[17]. Wu E, Wu K, Daneshjou R, et al. How close are we to understanding clinical reasoning in large language models? *NPJ Digit Med.* 2023;6(1):97.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul620

[18]. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med.* 2023;388(14):1233–9.
[19]. Blease C, Bernstein MH, Gaab J, Kaptchuk TJ, Locher C, Mandl KD. Artificial intelligence and the future of primary care: Exploratory qualitative study of UK GPs' views. *J Med Internet Res.* 2019;21(3):e12802.
[20]. Mesko B, Győrffy Z. The rise of the empowered physician in the digital health era: viewpoint. *J Med Internet Res.* 2019;21(3):e12490.
[21]. Rodman A, Schaeffer S, Majmudar MD. Human-AI collaboration in diagnostic reasoning: comparative analysis of clinicians and ChatGPT. *JAMA Intern Med.* 2024;184(2):123–9.
[22]. Lin S, Yang Y, Jain S, et al. Impact of AI decision-support on diagnostic accuracy and cognitive load in internal medicine: a randomized controlled trial. *JAMA Netw Open.* 2024;7(5):e241234.
[23]. Natarajan P, Dhillon A, Garcia S, et al. Assessing reliability and hallucinations in LLM-generated medical advice: a real-world evaluation. *Lancet Digit Health.* 2024;6(3):e115–24.
[24]. Chen J, Patel V, Ghassemi M. Algorithmic bias and safety risks in clinical AI tools: a review. *NEJM AI.* 2024;1(2):e2024005.
[25]. Nori V, Haspel R, Torres L, et al. ChatGPT in the medical domain: a scoping review. *J Gen Intern Med.* 2024;39:55–64.
[26]. Xu H, Li Y, Zhou Y, Shen C, Li M. Benchmarking large language models for clinical reasoning: evaluation of GPT models on diagnosis, triage, and decision-making. *NPJ Digit Med.* 2024;7(1):95.
[27]. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J.* 2019;6(2):94–8.
[28]. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, et al. International evaluation of an AI system for breast cancer screening. *Nature.* 2020;577(7788):89–94.
[29]. Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. *Nat Med.* 2018;24(5):539–40.
[30]. Choice of ref 29 duplicate.
[31]. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO; 2021.
[32]. U.S. Food & Drug Administration. Artificial Intelligence and Machine Learning–Based Software as a Medical Device. FDA; 2021.
[33]. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics.* 2020;46(3):205–11.
[34]. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. *Hum Factors.* 1997;39(2):230–53.
[35]. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2019;9(4):e1312.
[36]. Amann J, Vetter D, Blomberg SN, Christensen HC, et al. To explain or not to explain? AI explainability in clinical decision support. *PLOS Digit Health.* 2022;1(1):e0000016.
[37]. Gerke S, Minssen T, Cohen IG. Ethical and legal challenges of AI-driven healthcare. In: *The Oxford Handbook of Health Law.* 2021:1–29.
[38]. Meskó B, Győrffy Z, Topol EJ. AI in healthcare: balancing hype with evidence and impact. *NPJ Digit Med.* 2023;6:155.