RESEARCH PAPER

Comparative Evaluation of Diagnostic Accuracy between Perplexity Pro and Physicians using Stepwise Inpatient Case Vignettes: A Retrospective Cross-Sectional Observational Study

Sridhar S[1], Zeno LF,[1] Arun S[1], Suresh K[1*]
[1*]Department of General Medicine,
Sri Venkateshwaraa Medical College Hospital and Research Centre, Ariyur, Puducherry.
Email: sureshsuchu06@gmail.com

ABSTRACT

Background: Artificial intelligence (AI)–based large language models are increasingly being explored as tools to support clinical decision-making and diagnostic reasoning. Recent advances in generative AI have demonstrated promising performance in medical knowledge assessments; however, direct comparisons between AI systems and physicians using realistic inpatient clinical scenarios remain limited. Evaluating AI using sequential clinical vignettes that simulate the stepwise acquisition of clinical information may therefore provide a more clinically relevant assessment of its diagnostic capabilities.

Objective: To compare the diagnostic accuracy of an AI-based system with physicians using stepwise inpatient clinical vignettes derived from real-world hospital cases and to evaluate the level of diagnostic agreement between the AI system and physicians across different stages of clinical information.

Methodology: This retrospective cross-sectional observational study included 31 de-identified inpatient cases from the Department of General Medicine. Each case was converted into three sequential clinical vignettes representing progressive stages of inpatient clinical evaluation:

Stage 1 – Initial Presentation: history and physical examination findings.

Stage 2 – Mid-Admission: basic laboratory investigations, targeted diagnostic tests, and evolution of clinical findings.

Stage 3 – Final Work-up (Pre-discharge): imaging studies and advanced diagnostic investigations prior to discharge.

At each stage, the top five differential diagnoses were independently generated by an AI-based system (Perplexity Pro) and three experienced physicians. Diagnostic accuracy was defined as the presence of the confirmed discharge diagnosis within the top five differential diagnoses generated by each evaluator.

Diagnostic accuracy was calculated as proportions with corresponding frequencies. Differences between the AI system and physicians were assessed using the McNemar test for paired categorical data, while diagnostic agreement was evaluated using Cohen's kappa statistic, with $p < 0.05$ considered statistically significant.

Results: Diagnostic accuracy improved progressively for both the AI system and physicians as additional clinical information became available across the sequential vignette stages. At the final vignette stage (Stage 3 – Final Work-up), the AI system achieved a diagnostic accuracy of 100% (31/31), correctly including the confirmed discharge diagnosis within the top five differential diagnoses for all cases. Physician diagnostic accuracy at this stage was 87.1% (27/31) for Physician 1, 90.3% (28/31) for Physician 2, and 90.3% (28/31) for Physician 3. When physician responses were combined as a pooled consensus, diagnostic accuracy reached 93.5% (29/31). Diagnostic agreement between the AI system and the pooled physician consensus occurred in 93.5% (29/31) of cases.

Conclusion: The AI system demonstrated high diagnostic accuracy when evaluated using stepwise inpatient clinical vignettes that simulate real-world clinical decision-making. These results suggest that AI-based systems are capable of effectively integrating complex clinical information and may support physicians in the diagnostic process, while underscoring the continued value of collaborative clinical expertise in achieving optimal diagnostic outcomes.

Keywords: Artificial Intelligence; Diagnostic Accuracy; Clinical Decision Support; Large Language Models; Case Vignettes; Perplexity Pro

How to cite this article: Sridhar S, Zeno LF, Arun S, Suresh K. Comparative Evaluation of Diagnostic Accuracy between Perplexity Pro and Physicians using Stepwise Inpatient Case Vignettes: A Retrospective Cross-Sectional Observational Study. Int J Drug Deliv Technol. 2026;16(33s):626-639. DOI: 10.25258/ijddt.16.33s.75

INTRODUCTION

Artificial intelligence (AI) is a broad term referring to the application of computational algorithms that can analyze large data sets to classify, predict, or gain useful conclusions.[1] It has emerged as a transformative technological advancement across multiple sectors, with rapidly expanding applications in healthcare, particularly in medical diagnostics and clinical decision-making.[1,2] AI can broadly be categorized into expert systems and machine learning. Expert systems simulate human

decision-making using a structured knowledge base and inference engine; however, their effectiveness is often limited by challenges in knowledge acquisition and adaptability. In contrast, machine learning, which forms the core of modern AI, relies on large datasets and advanced algorithms to continuously improve predictive performance and increasingly aims to match or surpass human expertise in complex decision-making tasks.[3]

In recent years, large language model (LLM)–based artificial intelligence systems have demonstrated remarkable capabilities in processing and synthesizing large volumes of information. Unlike traditional AI systems designed for single-query responses, LLM-based chatbots utilize extensive training datasets to perform sequential reasoning across related tasks. Comparative evaluations of contemporary models have shown that GPT-5 Pro tends to provide deeper analytical responses but with fewer citations, whereas Gemini 2.5 Pro and the Perplexity Pro search engine retrieve broader literature with comparatively limited critical analysis.[4] These evolving capabilities highlight the potential of LLMs to support clinical reasoning and medical decision-making.

The importance of integrating AI into standard medical practice is further highlighted by the increasing digitization and interconnection of healthcare systems. Medical imaging interpretation, early disease identification, prognosis prediction, risk stratification, large-scale patient data analysis, and individualised treatment planning are just a few of the areas of patient care that AI-driven solutions have shown promise in improving.[5] Furthermore, remote monitoring of physiological indicators across vast populations or high-risk groups is becoming possible due to the growing availability of biosensors and compact health monitoring equipment. AI systems that can effectively handle and understand massive datasets will be crucial for the effective use of such extensive health data.[6–8]

Various AI techniques have already shown promising results in clinical medicine. Artificial neural networks (ANNs), for example, have been applied in clinical diagnosis, radiological and histopathological image analysis, intensive care data interpretation, and physiological waveform analysis. These systems have demonstrated utility in diagnosing conditions such as prostate cancer, appendicitis, glaucoma, and other cytological and histological abnormalities. AI-assisted systems like PAPNET have also been developed to aid cervical cancer screening. Furthermore, ANN-based models have been used to analyze medical imaging modalities including X-ray, ultrasound, CT, MRI, and radioisotope scans, as well as physiological signals such as ECG, EEG, EMG, and Doppler waveforms to detect conditions including myocardial infarction, arrhythmias, epilepsy, and sleep disorders.[9–11] Despite these advances, the real-world diagnostic performance of LLM-based AI systems in clinical medicine remains an evolving area of research. In particular, limited studies have directly compared the diagnostic reasoning abilities of LLM platforms with those of practicing physicians using structured clinical case scenarios. Therefore, this study aims to evaluate the diagnostic capability of the LLM-based model Perplexity Pro in interpreting clinical information. Specifically, the study compares the diagnostic accuracy of Perplexity Pro and qualified physicians using identical de-identified stepwise inpatient clinical case vignettes by assessing whether the final confirmed discharge diagnosis appears within the top five differential diagnoses generated at each stage. Additionally, the study evaluates the level of diagnostic agreement between Perplexity Pro and physicians across three sequential vignette stages to assess its potential role as a supportive clinical decision-support tool in general medicine inpatient settings.

METHODOLOGY

After obtaining approval from the Scientific Research Committee and Institutional Human Ethics Committee, the study was conducted in the Department of General Medicine at a Tertiary Care Hospital. This study was designed as a retrospective cross-sectional observational study using de-identified inpatient case records. The sample size was calculated based on previously reported diagnostic accuracy of 64.6% for physicians and 40.2% for AI in a similar study by Hirosawa T et al.[12] Using a paired proportion (McNemar) test with $\alpha = 0.05$, power = 80%, and an assumed correlation of 0.5 between paired outcomes, the minimum required sample size was estimated to be 31 cases. Cases were selected using a purposive sampling strategy to ensure inclusion of clinically diverse and diagnostically challenging inpatient cases encountered in the Department of General Medicine. Eligible cases were identified from the inpatient discharge registry and screened based on predefined inclusion and exclusion criteria. Only cases with a clearly documented final confirmed discharge diagnosis and sufficient clinical documentation were included. Cases were intentionally selected to represent a range of common and uncommon diagnostic scenarios encountered in routine inpatient practice. All patient identifiers were removed before vignette development to maintain confidentiality. For each case, the final principal discharge diagnosis was verified from the medical records and used as the gold standard reference for diagnostic accuracy assessment.

Clinical Case Vignettes

Each selected case was converted into three sequential stepwise clinical case vignettes representing different stages of inpatient evaluation: Stage 1 (initial presentation including history and physical examination findings), Stage 2 (mid-admission information including basic laboratory investigations and clinical evolution), and Stage 3 (final pre-discharge workup including advanced investigations and imaging). Thus, a total of 93 vignettes (31 cases × 3 stages) were generated (Figure 1).

Case difficulty was categorized as easy, moderate, or difficult based on physician consensus during vignette preparation. This classification considered clinical complexity, rarity of the condition, and the number of

plausible differential diagnoses.

Additionally, as an exploratory assessment, case difficulty was supported by diagnostic performance patterns across vignette stages, with cases diagnosed correctly at earlier stages or by a higher proportion of evaluators considered relatively easier.

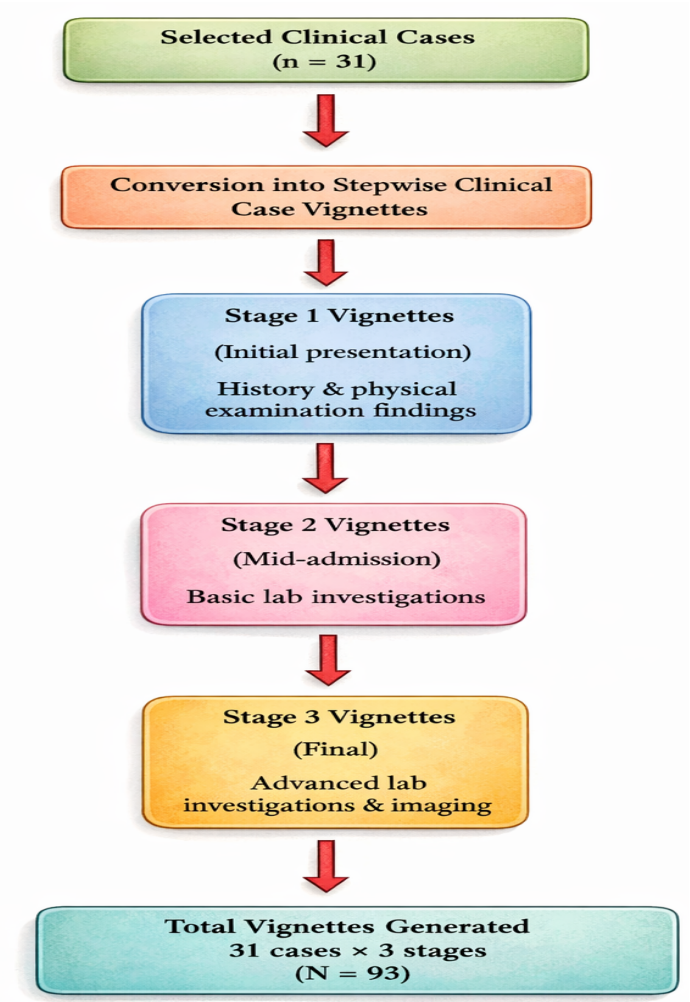

Figure 1: Clinical Case Vignettes

Diagnostic Assessment by Perplexity Pro and Physicians

Diagnostic evaluation was performed independently by Perplexity Pro and three qualified general physicians with a minimum of three years of post-MD clinical experience in general medicine. The participating physicians were not involved in vignette preparation. For each vignette stage, Perplexity Pro was provided with the clinical vignette using a standardized prompt requesting the top five differential diagnoses based only on the information provided.

Standardized Prompt:

"Based only on the information provided below, please list the top 5 differential diagnoses. Provide one diagnosis per line with no explanations. Do not assume unavailable information. [Vignette text here]"

Each vignette was entered into a new chat session to avoid contextual carryover between cases. The same prompt and model version were used for all evaluations. Similarly, the three physicians independently reviewed the same vignettes in stage-wise sessions and recorded their top five differential diagnoses without discussion or knowledge of the AI responses or other physicians' responses. All evaluators were blinded to the final diagnosis and to each other's responses. Also, participating physicians evaluated the vignettes independently and were instructed not to consult external references during the diagnostic process.

Diagnostic Accuracy and Statistical Analysis

Diagnostic accuracy was defined as the presence of the final confirmed discharge diagnosis within the top five differential diagnoses suggested by the evaluator. Diagnostic performance was calculated separately for Perplexity Pro, for each physician individually and for a pooled physician consensus diagnosis formed using a 2-of-3 agreement rule. All collected data were compiled in Microsoft Excel and subsequently analyzed using SPSS

software version 25. Descriptive statistics were used to summarize diagnostic accuracy, while paired comparisons between Perplexity Pro and physician diagnoses were performed using the McNemar test. Inter-rater agreement between Perplexity pro and physicians was assessed using Cohen's kappa statistic, and a p-value <0.05 was considered statistically significant. To evaluate changes in diagnostic accuracy across the three vignette stages, Cochran's Q test was performed using repeated binary outcomes for each case.

Overview of the current research study was given as flowchart in figure 2.

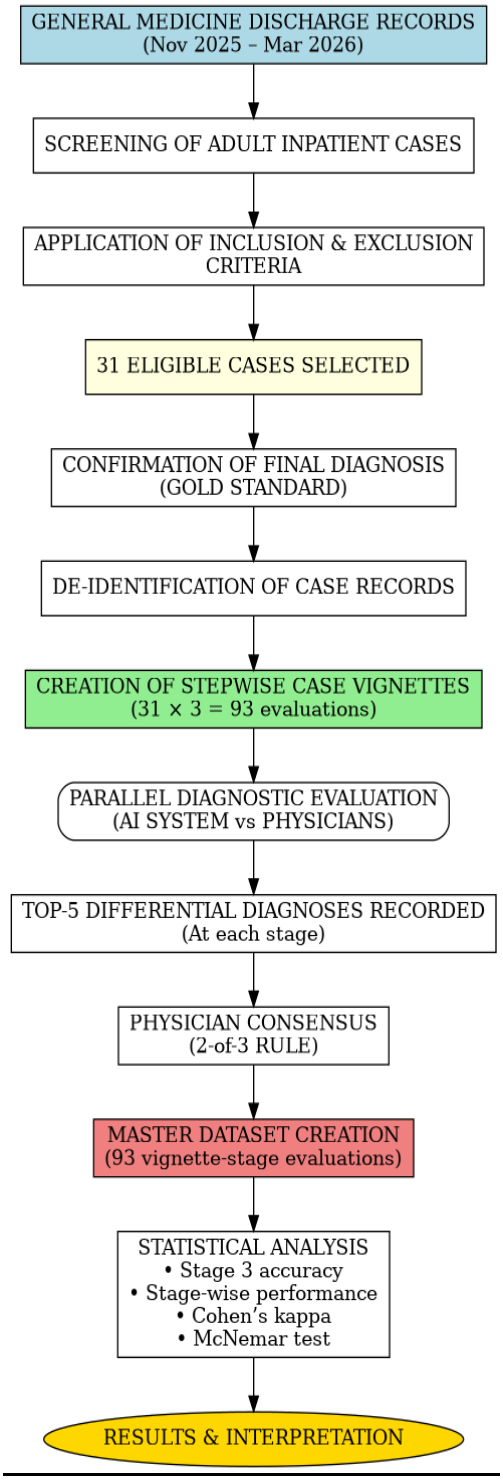

Figure 2: Overview of study design and statistical analysis workflow

RESULTS:

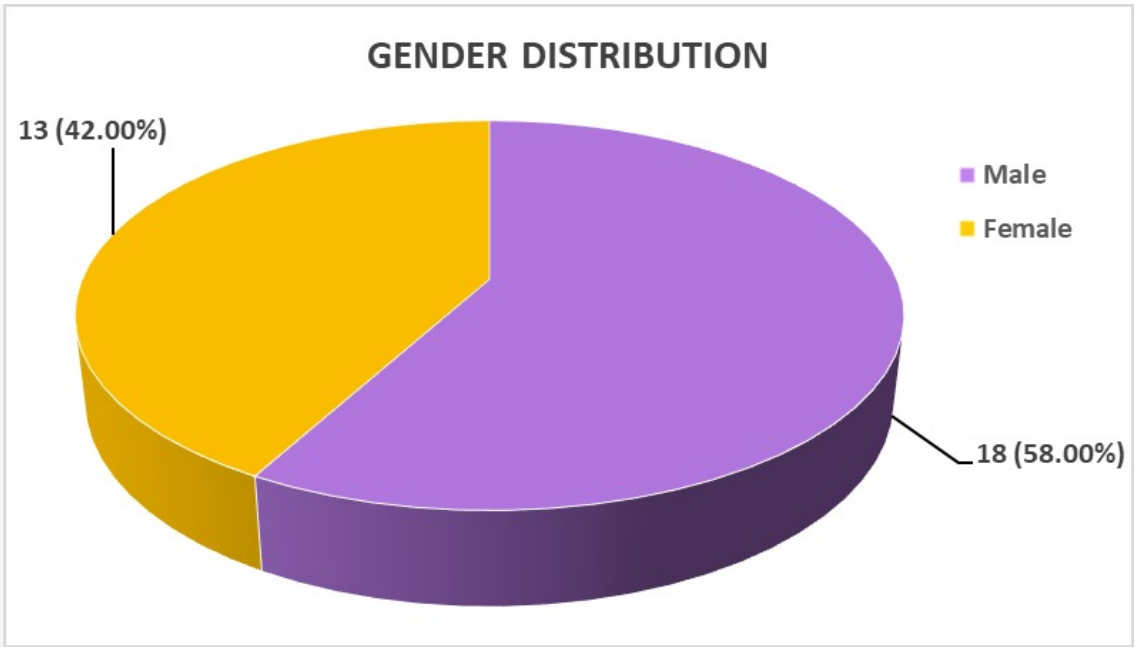

Figure 3: Frequency Distribution of Gender among case vignettes (n = 31)

Figure 3 shows the frequency distribution of gender among case vignettes. Out of the total 31 cases included in the study, 18 cases (58.00%) were males and 13 cases (42.00%) were females.

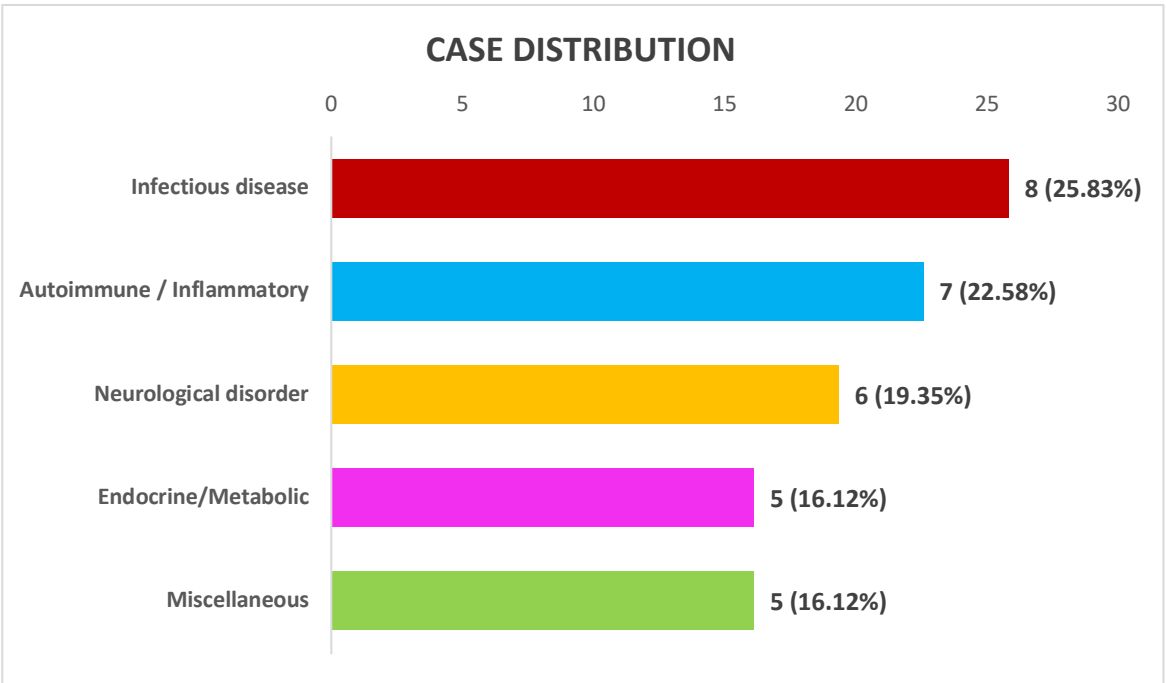

Figure 4: Frequency Distribution of cases categories (n = 31)

The distribution of clinical cases included in the study showed that infectious diseases constituted the largest proportion, accounting for 8 cases (25.83%). This was followed by autoimmune or inflammatory disorders, which comprised 7 cases (22.58%). Neurological disorders accounted for 6 cases (19.35%), while endocrine or metabolic disorders and miscellaneous conditions each represented 5 cases (16.12%) of the total study cases (Figure 4).

Table 1: Overall diagnostic accuracy of AI and physicians (N = 31)

Evaluator	Correct	Incorrect	Accuracy
Perplexity Pro	31/31	-	100.00%
Physician 1	27/31	04/31	87.10%
Physician 2	28/31	03/31	90.30%
Physician 3	28/31	03/31	90.30%
Physician Consensus	29/31	02/31	93.50%

Table 1 shows the frequency distribution of overall accuracy among physician and perplexity pro based on their top five differential diagnosis. Perplexity Pro correctly identified all 31 cases, achieving a diagnostic accuracy of 100% (95% CI: 88.8%–100%). Among the physicians, Physician 1 correctly diagnosed 27 out of 31 cases (87.10%, 95% CI: 70.2%–96.4%), while Physician 2 and Physician 3 each correctly diagnosed 28 out of 31 cases (90.30%, 95% CI: 74.2%–98.0%). When the diagnoses of the three physicians were combined to obtain a physician consensus, a total of 29 out of 31 cases were correctly diagnosed, resulting in an overall accuracy of 93.50% (95% CI: 78.6%–99.2%).

Table 2: Frequency Distribution of stage wise diagnostic accuracy (N = 31)

Stages	Perplexity Pro	Physician 1	Physician 2	Physician 3	Physician Consensus
Stage 1	70.9% (22/31)	64.5% (20/31)	67.7% (21/31)	64.5% (20/31)	71% (22/31)
Stage 2	77.4% (24/31)	77.4% (24/31)	77.4% (24/31)	74.2% (23/31)	83.9% (26/31)
Stage 3	100% (31/31)	87.1% (27/31)	90.3% (28/31)	90.3% (28/31)	93.5% (29/31)

Table 2 shows the stage-wise diagnostic accuracy of Perplexity Pro and the three physicians was assessed across the sequential clinical vignette stages. Perplexity Pro's accuracy in Stage 1 was 70.90% (22/31; 95% CI: 52.5%–85.1%), while the accuracy of Physicians 1, 2, and 3 was 64.5%, 67.7%, and 64.5%, respectively, with a physician consensus accuracy of 71%. As more laboratory and clinical data became available in Stage 2, the diagnostic accuracy improved. Perplexity Pro achieved 77.40% accuracy (24/31; 95% CI: 58.9%–90.4%), while the physicians showed accuracies of 77.40%, 77.40%, and 74.20%, respectively, and the physician consensus rose to 83.90%. Perplexity Pro achieved 100% accuracy (31/31; 95% CI: 88.8%–100%) in Stage 3 when full clinical information, including advanced investigations, was given. Physician 1, Physician 2, and Physician 3 achieved 87.1%, 90.3%, and 90.3%, respectively, and the physician consensus reached 93.5%

Diagnostic accuracy improved progressively across stages for both the AI system and physicians, with the highest accuracy observed at Stage 3. Cochran's Q test demonstrated a statistically significant difference between Stage 1, Stage 2, and Stage 3 ($Q = 14.2$, $p = 0.0008$).

This trend reflects improved diagnostic performance with the availability of additional clinical information.

Table 3: Frequency Distribution of Case-level diagnostic concordance (N = 31)

Category	Cases
Perfect agreement (AI+All physicians correct)	26
AI Failures	0
Physician 1 errors	4
Physician 2 errors	3
Physician 3 errors	3

Table 3 shows the distribution of diagnostic agreement between Perplexity Pro and the physicians is presented in Table 3. Perfect agreement, defined as cases where both the AI system and all three physicians correctly identified the diagnosis, was observed in 26 out of 31 cases. There were no AI failures, indicating that Perplexity Pro successfully included the correct diagnosis within its top differential diagnoses for all cases evaluated. In comparison, Physician 1 made errors in 4 cases, while Physician 2 and Physician 3 each made errors in 3 cases.

Table 4: Frequency Distribution of differential diagnosis Ranking performance (N = 31)

Rank position	Cases
Rank 1	20 (64.5%)
Rank 2	6 (19.4%)
Rank 3	3 (9.7%)
Rank 4	1 (3.2%)
Rank 5	1 (3.2%)

The distribution of the rank position of the correct diagnosis generated by the AI system is shown in table 4. The correct diagnosis was placed at Rank 1 in 20 cases (64.5%), indicating that the AI identified the correct diagnosis as the most likely diagnosis in the majority of cases. In 6 cases (19.4%), the correct diagnosis appeared at Rank 2, while 3 cases (9.7%) had the correct diagnosis at Rank 3. The correct diagnosis appeared at Rank 4 and Rank 5 in 1 case each (3.2%).

Table 5.1: Comparison of discordant case analysis between Perplexity and Physician (N = 31)

Comparison	Discordant pairs	McNemar test (exact p value)
AI v Physician 1	4	0.12
AI v Physician 2	3	0.25
AI v Physician 3	3	0.25
AI v Physician Consensus	2	0.50

(p value <0.05 is considered as statistically significant)

McNemar test analysis showed no statistically significant difference between the AI system and individual physicians (Physician 1: p = 0.12; Physician 2: p = 0.25; Physician 3: p = 0.25). Although the AI system demonstrated a higher number of correct diagnoses, the differences were not statistically significant, likely due to the limited sample size (table 5.1).

Table 5.2: Contingency Tables

Perplexity Pro vs Physician 1		
	Physician Correct	**Physician Incorrect**
AI Correct	27	4
AI Incorrect	0	0
Perplexity Pro vs Physician 2		
	Physician Correct	**Physician Incorrect**
AI Correct	28	3
AI Incorrect	0	0
Perplexity Pro vs Physician 3		
AI Correct	28	3
AI Incorrect	0	0

The contingency tables (table 5.2) demonstrate that all discordant cases occurred when the AI system correctly identified the diagnosis while the physicians did not. No instances were observed where physicians correctly diagnosed cases missed by the AI system.

The diagnostic performance of Perplexity Pro was compared with each physician using the McNemar test to assess differences in paired diagnostic outcomes. The number of discordant pairs between Perplexity Pro and Physician 1 was 4, with an exact p-value of 0.12, indicating no statistically significant difference in diagnostic accuracy. Similarly, the comparison between Perplexity Pro and Physician 2 showed 3 discordant pairs (p = 0.25), and between Perplexity Pro and Physician 3 also showed 3 discordant pairs (p = 0.25), both of which were not statistically significant. When compared with the physician consensus, there were 2 discordant pairs, with a p-value of 0.50, again indicating no statistically significant difference.

These findings suggest comparable diagnostic performance between the AI system and physicians, despite numerically higher accuracy observed with the AI system.

Table 6: Stage-wise diagnostic revision patterns (N = 93)

Transition	AI Improvement	Consensus Improvement
Stage 1 → stage 2	+8%	+13%
Stage 2 → stage 3	+22%	+9.5%
Stage 1 → stage 3	+30%	+22.5%

The improvement in diagnostic accuracy across the sequential clinical stages was evaluated for both Perplexity Pro and the physician consensus. From Stage 1 to Stage 2, the diagnostic accuracy of Perplexity Pro improved by 8%, while the physician consensus improved by 13%. A further improvement was observed from Stage 2 to Stage 3, where Perplexity Pro demonstrated an increase of 22%, whereas the physician consensus improved by 9.5%. Overall, from Stage 1 to Stage 3, the total improvement in diagnostic accuracy was 30% for Perplexity Pro and 22.5% for the physician consensus (Table 6).

Table 7: Agreement Between Physician Consensus and Perplexity Pro Diagnoses Using Cohen's Kappa (N = 93)

Variables	Kappa	P value
Physician and Perplexity Pro	0.81 **(Strong)**	**< 0.05**

(p value <0.05 is considered as statistically significant)

The level of agreement between physician consensus and Perplexity Pro was assessed using the Cohen's kappa statistic. The analysis showed a kappa value of 0.81, which indicates strong agreement between the physician and the AI-generated diagnoses. Agreement was assessed based on whether the final diagnosis appeared within the top five differential diagnoses. This suggests that both the physician and Perplexity Pro produced similar diagnostic conclusions in most of the evaluated cases (Table 7).

Table 8: Diagnostic error type analysis (N = 93)

Error type	AI cases	Physician cases
Incorrect disease category	0	3
Missed rare diagnosis	0	2
Over - diagnosis	0	2
Under-diagnosis	0	1
Total errors	0	8

The types of diagnostic errors made by the AI system and physicians were analyzed. Perplexity Pro did not demonstrate any diagnostic errors across the evaluated cases. In contrast, physicians showed a total of 8 diagnostic errors. Among these, 3 cases involved incorrect disease categorization, 2 cases were due to missed rare diagnoses, and 2 cases represented over-diagnosis, while 1 case was attributed to under-diagnosis (Table 8).

Table 9: Case difficulty stratification - Exploratory Analysis (N = 93)

Case difficulty	Number of cases	AI Accuracy	Physician Accuracy
Easy	12	100%	100%
Moderate	11	100%	91%
Difficult	8	100%	75%

The diagnostic accuracy of Perplexity Pro and physicians was also evaluated based on the difficulty level of the clinical cases and was shown in table 9. Among the 12 easy cases, both Perplexity Pro and physicians achieved 100% diagnostic accuracy. In the moderate difficulty category, which included 11 cases, Perplexity Pro maintained 100% accuracy, while physicians demonstrated an accuracy of 91%. In the difficult case category consisting of 8 cases, Perplexity Pro continued to achieve 100% diagnostic accuracy, whereas physician accuracy decreased to 75%.

Easy cases were typically diagnosed correctly at earlier vignette stages, whereas difficult cases more often required Stage 3 information for accurate diagnosis.

Although the AI system demonstrated uniformly high accuracy across all difficulty levels, physician diagnostic accuracy decreased with increasing case difficulty, supporting the validity of the classification.

Table 10: Diagnostic evaluation (Improvement across stages) (N = 93)

Evaluator	Stage 1 → stage 2	Stage 2 → stage 3	Stage 1 → stage 3
Physician 1	+14%	+8.1%	+22.1%
Physician 2	+9%	+13.3%	+22.3%
Physician 3	+8%	+16.3%	+24.3%
AI system	+8%	+22%	+30%

The improvement in diagnostic accuracy across the sequential clinical stages was analyzed for each evaluator and given in table 10. Physician 1 demonstrated an improvement of 14% from Stage 1 to Stage 2 and 8.1% from Stage 2 to Stage 3, resulting in an overall increase of 22.1% from Stage 1 to Stage 3. Physician 2 showed an improvement of 9% from Stage 1 to Stage 2 and 13.3% from Stage 2 to Stage 3, with a total improvement of 22.3%. Physician 3 demonstrated an improvement of 8% between Stage 1 and Stage 2 and 16.3% between Stage 2 and Stage 3, resulting in an overall improvement of 24.3%. The AI system demonstrated an improvement of 8% from Stage 1 to Stage 2 and a further increase of 22% from Stage 2 to Stage 3, resulting in a total improvement of 30% from Stage 1 to Stage 3.

Representative Case Analysis:

Case A – Stage 1

A male in his late 30s presented with high-grade intermittent fever for one month, associated with chills and rigors. Three days prior to admission, he developed generalized myalgia and continuous holocranial throbbing headache. He reported approximately 4 kg unintentional weight loss over one month. There was no history of cough, dyspnea, chest pain, hemoptysis, vomiting, photophobia, seizures, urinary symptoms, abdominal pain, rash, or joint complaints. He reported frequent travel to rural and hilly regions. No previously diagnosed comorbid illness was documented.

On arrival, he was conscious and oriented. Temperature 100°F, pulse 120/min (regular), blood pressure 120/70 mmHg, respiratory rate 17/min, oxygen saturation 96% on room air. Capillary blood glucose was 409 mg/dL. He appeared dehydrated. No pallor, icterus, cyanosis, clubbing, lymphadenopathy, or edema were noted. Cardiovascular examination revealed normal S1 and S2 without murmurs. Respiratory examination showed normal bilateral air entry without added sounds. Abdomen was soft and non-tender without clinically palpable organomegaly. Neurological examination revealed GCS 15/15 with no focal deficits or meningeal signs. ECG demonstrated sinus tachycardia. Response of Case A-Stage 1 was given in table 11.1.

Table 11.1: Response- Case A – Stage 1

Evaluator	Differential Diagnoses
Perplexity Pro	Malaria; Scrub typhus; Dengue; Enteric fever; Tuberculosis
Physician 1	Scrub typhus; Enteric fever; Tuberculosis; Malaria; Leptospirosis
Physician 2	Malaria; Dengue; Scrub typhus; Enteric fever; Viral febrile illness
Physician 3	Tuberculosis; Scrub typhus; Enteric fever; Malaria; Leptospirosis
Physician Consensus	Scrub typhus; Malaria; Enteric fever; Tuberculosis

Case A – Stage 2

Response of Case A-Stage 1 given in table 11.2. A male in his late 30s presented with 1-month high-grade intermittent fever, chills, rigors, 4 kg weight loss, recent myalgia, and holocranial headache. Frequent rural/hilly travel, no cough/dyspnea/chest pain/ vomiting/ photophobia/ seizures/ urinary/abdominal/rash/joint symptoms, no comorbidities. On exam: conscious, 100°F, HR 120/min, BP 120/70, RR 17, SpO2 96%, glucose 409 mg/dL, dehydrated, normal CV/respiratory/abdomen/neuro exams (GCS 15/15), sinus tachycardia on ECG.

During hospitalization, the patient continued to have fever spikes up to 103°F with persistent tachycardia ranging between 110–130/min. He developed intermittent disorientation to time and place and increasing irritability. Laboratory evaluation showed leukocytosis with neutrophilic predominance, elevated C-reactive protein (24 mg/dL), and markedly elevated procalcitonin (19.6 ng/mL). Serum sodium was reduced. Serum creatinine showed mild elevation. Urinalysis demonstrated glycosuria and proteinuria. Arterial blood gas revealed metabolic acidosis with ketosis. Serial troponin values demonstrated a rising trend. Screening tests for malaria, dengue, and scrub typhus were negative. Respiratory viral screening was negative. Initial blood culture grew skin commensals; repeat cultures showed no growth. Urine culture demonstrated no growth. Chest radiograph revealed heterogeneous opacity in the right mid to lower lung zones. Ultrasonography of the abdomen demonstrated grade I fatty liver changes and splenomegaly (~13 cm). Computed tomography of the brain showed no acute intracranial abnormality. Management included intravenous fluids, insulin infusion, empirical broad-spectrum intravenous antibiotics, and supportive care.

Table 11.2: Response-Case A-Stage 2

Evaluator	Differential Diagnoses
Perplexity Pro	Severe bacterial sepsis; Community-acquired pneumonia; Leptospirosis; Rickettsial infection; Melioidosis
Physician 1	Severe bacterial pneumonia with sepsis; Leptospirosis; Scrub typhus; Melioidosis; Tuberculosis
Physician 2	Sepsis of unknown origin; Pneumonia; Leptospirosis; Melioidosis; Viral infection
Physician 3	Severe pneumonia; Scrub typhus (complicated); Leptospirosis; Melioidosis; Tuberculosis
Physician Consensus	Severe bacterial sepsis/pneumonia; Leptospirosis; Scrub typhus; Emerging suspicion of Melioidosis
Key Difference	AI includes Rickettsial earlier; Physicians favor TB/Scrub typhus

CASE A – Stage 3

A male in his late 30s presented with 1-month high-grade intermittent fever, chills, rigors, 4 kg weight loss, recent myalgia, and throbbing holocranial headache. Frequent rural/hilly travel noted. No cough/dyspnea/chest pain/ vomiting/ photophobia/ seizures/ urinary/ abdominal/rash/joint symptoms. On exam: conscious, 100°F, HR 120/min, BP 120/70, RR 17, SpO2 96%, glucose 409 mg/dL, dehydrated, normal CV/resp/abdomen/neuro exams (GCS 15/15), sinus tachycardia.

Hospital course: Persistent fever (103°F), tachycardia (110-130/min), disorientation, irritability. Labs showed neutrophilic leukocytosis, CRP 24 mg/dL, procalcitonin 19.6 ng/mL, hyponatremia, mild AKI, glycosuria/proteinuria, DKA, rising troponins. Negative malaria/ dengue/ scrub typhus. Chest X-ray: right mid-lower zone opacity. USG abdomen: splenomegaly (13 cm), fatty liver. Normal brain CT. Initial cultures: skin commensals only.

Subsequently, the patient developed recurrent hyperpyrexia with a maximum recorded temperature of 106°F, worsening tachypnea, declining GCS, and hypotension requiring vasopressor support. He underwent elective endotracheal intubation and mechanical ventilation. Laboratory parameters showed persistent leukocytosis, progressive metabolic acidosis, thrombocytopenia trend, and biochemical evidence of multi-organ dysfunction. Contrast-enhanced CT thorax demonstrated consolidation with air bronchograms and surrounding ground-glass opacities in the right upper lobe with associated bronchiectatic and fibrotic changes. Lumbar puncture revealed cerebrospinal fluid with lymphocytic predominance, elevated protein concentration, and elevated glucose levels. Two-dimensional echocardiography demonstrated severe global hypokinesia with markedly reduced left ventricular ejection fraction (15–20%). Repeat blood cultures yielded growth of Gram-negative bacilli demonstrating bipolar staining morphology and non-lactose fermenting characteristics on culture media. Species-level identification and antimicrobial susceptibility testing were completed. Despite escalation of therapy and full supportive management, the clinical course progressed to refractory shock with multi-organ dysfunction. Response of Case A-Stage 1 given in table 11.3 and comparison of performance in three stages given

in table 12.

Table 11.3: Response-Case A-Stage 3

Evaluator	Diagnoses
Perplexity Pro	Melioidosis; Severe gram-negative sepsis; Septic shock with pneumonia; Disseminated infection
Physician 1	Melioidosis; Gram-negative septicemia; Severe pneumonia with MODS
Physician 2	Melioidosis; Septic shock; Disseminated bacterial infection
Physician 3	Melioidosis; Severe sepsis; Pneumonia with multi-organ dysfunction
Physician Consensus	Melioidosis

Table 12: Comparison of Performance in Three Stages

Stage	AI Performance	Physician Performance	Key Observation
Stage 1	Broad tropical infections	Similar but includes TB earlier	No clear diagnosis
Stage 2	Focus on sepsis	Earlier suspicion of Melioidosis	Physicians outperform AI
Stage 3	Converges to diagnosis	Converges to diagnosis	Full agreement

DISCUSSION

Misdiagnosis is still a major problem in clinical practice despite tremendous advancements in medical understanding and diagnostic tools. The potential of artificial intelligence (AI) to enhance diagnostic precision and assist therapeutic decision-making is becoming more and more apparent. Our study is distinct in that it evaluates diagnostic accuracy at successive stages of clinical evaluation, namely history taking, physical examination, and investigation findings, whereas other studies have assessed the overall diagnostic performance of AI systems. Additionally, the study contrasts an AI model's diagnostic performance with that of three doctors throughout these phases, and the results show that the AI system performed more accurately.

The mean age of cases in the vignettes is 48 ± 15 years in our study. Similar finding was noted in Jabbour S et al and Graf M et al studies.[13,14] In the present study, male cases constituted 58.0%, while females accounted for 42.0%, indicating a male predominance. A similar gender distribution was reported by Harada Y et al., who also observed a higher proportion of male cases.[15] In contrast, Ye C et al. reported a female predominance in their study population, highlighting the variability in gender distribution across different study settings and populations.[16] In our study, 25.83% of the cases were related to infectious diseases, followed by 22.58% autoimmune/inflammatory disorders, 19.35% neurological disorders, and 16.12% endocrine or metabolic disorders. The remaining 16.12% of cases were categorized under miscellaneous disorders. A similar broad distribution of disease categories has been reported in previous vignette-based diagnostic studies (Hammoud M et al and Ben-shabbat N et al) evaluating AI-assisted clinical decision support, where cases were drawn from multiple specialties including infectious, neurological, endocrine, and inflammatory conditions to simulate real-world diagnostic scenarios.[17,18]

In the present study, the overall diagnostic accuracy of Perplexity Pro was 100.00%, whereas the diagnostic accuracies of Physician 1, Physician 2, and Physician 3 were 87.10%, 90.30% and 90.30%, respectively. This indicates that Perplexity Pro demonstrated a higher diagnostic accuracy compared with the participating physicians. Similar findings have been reported in previous studies. For instance, cross-sectional studies conducted by Miranda J et al. and Harada Y et al. demonstrated that AI-based virtual assistants achieved higher diagnostic accuracy than physicians across several medical domains, including internal medicine and surgery.[15,19] Furthermore, a meta-analysis by Takita H et al. comparing generative AI with physicians reported that AI systems frequently demonstrate comparable or even superior diagnostic accuracy in clinical decision-making tasks.[20] In the current study, the diagnostic accuracy improved progressively across the three stages for both AI and physicians. Perplexity Pro showed an accuracy of 70.0% in Stage 1, which increased to 78.0% in Stage 2 and reached 100.0% in Stage 3. Similarly, the physicians also demonstrated improved diagnostic accuracy with the addition of further clinical information, with the physician consensus increasing from 71.0% in Stage 1 to 84.0% in Stage 2 and 93.5% in Stage 3. Rodman et al demonstrated that a generative AI (GPT-4) more accurately estimated diagnostic probabilities after test results than clinicians.[21] Additionally, a randomized clinical trial by Goh E et al reported that large language models outperformed physicians in diagnostic reasoning, while studies using AI assistance have shown improved clinician accuracy.[22] A systematic review by Takita H et al also highlighted that AI systems frequently perform at least as well as physicians across a range of diagnostic tasks.[20]

Diagnostic performance varied across case difficulty levels and vignette stages. Easy cases were generally identified correctly at earlier stages of the vignette, whereas difficult cases often required more comprehensive clinical information, particularly at Stage 3, for accurate diagnosis. This trend was reflected in physician performance, with diagnostic accuracy decreasing as case difficulty increased. In contrast, the AI system demonstrated consistently high diagnostic accuracy across all difficulty categories, suggesting an ability to effectively utilize structured clinical information even in complex scenarios. The requirement of advanced-stage data for diagnosing difficult cases further supports the validity of the difficulty classification and aligns with the observed progressive

improvement in diagnostic accuracy across vignette stages.

In the present study, Perplexity Pro made no diagnostic errors, whereas each of the physicians committed a minimum of three errors in their diagnoses. This finding is consistent with previous studies reported by Shan J et al, Rodman A et al and Miranda J et al which similarly demonstrated that AI systems commit fewer diagnostic errors and often achieve higher accuracy compared with physicians.[3,19,21] In the present study, the majority of cases (64.5%) were correctly identified as the top-ranked diagnosis (Rank-1) by the AI system. A smaller proportion of cases were ranked second (19.4%) or third (9.7%), while only 3.2% of cases each were ranked fourth and fifth. This indicates that the AI system most frequently assigned the correct diagnosis as the highest-ranked option, reflecting strong diagnostic prioritization and reliability.

In the present study, the correct diagnosis was most frequently placed in the first rank (64.5%), with progressively fewer correct diagnoses at lower ranks (rank-2: 19.4%, rank-3: 9.7%, rank-4 and rank-5: 3.2% each), indicating that the AI strongly prioritized the most likely diagnosis. Similar rank-based diagnostic patterns have been reported in previous studies like Hammoud M et al and Hirosawa T et al, where AI-generated differential diagnosis lists tended to place the correct diagnosis higher in the ranking, with decreasing frequency at lower ranks.[12,17]

In the Current research, the comparison of diagnostic concordance between Perplexity Pro and the individual physicians showed a small number of discordant pairs, with 4 discordant cases for Physician 1, 3 cases each for Physicians 2 and 3, and 2 cases for the physician consensus. The differences were not statistically significant ($p > 0.05$ for all comparisons), indicating that the AI system's diagnoses were largely in agreement with those of the physicians and the consensus, despite achieving higher overall accuracy. These findings of largely non-significant diagnostic discordance between AI and physicians are supported by previous work. In a systematic review and meta-analysis, Takita H et al. reported that generative AI models did not differ significantly from physicians in overall diagnostic performance ($p = 0.10$) and showed no significant difference when compared with non-expert clinicians ($p = 0.93$).[20] The sequential evaluation of clinical information led to notable improvements in diagnostic accuracy for both AI and physicians. Perplexity Pro showed a total improvement of 30% from Stage 1 to Stage 3, surpassing the 22.5% improvement observed in the physician consensus, highlighting the AI's ability to effectively integrate additional patient data across stages. Similarly, Fukuzawa F et al supports the concept that diagnostic accuracy increases as more detailed clinical information is provided, similar to your observation of sequential improvement from history to investigations in your study.[23]

In this investigation, we observed a Cohen's Kappa of 0.81, indicating strong agreement between Perplexity Pro and physician diagnoses. Similar high inter-rater agreement metrics have been reported in studies examining AI diagnostic systems, such as AI models for diabetic retinopathy screening with near-perfect kappa values, large language models showing substantial agreement with clinicians in differential diagnosis and ChatGPT evaluations in retinopathy of prematurity cases demonstrating kappa values around 0.80, reinforcing the reliability and consistency of AI-supported diagnosis compared with clinical assessments.[24,25] Perplexity Pro achieved 100% accuracy across all case categories, whereas physician accuracy decreased with increasing complexity, from 100% in easy cases to 91% in moderate and 75% in difficult cases, highlighting the AI system's consistent performance even in challenging scenarios. whereas physician accuracy declined with increasing complexity, aligns with findings from studies evaluating large language models in clinical diagnosis. Dinc M et al. reported that advanced LLMs achieved very high accuracy in common conditions and sustained strong performance as case complexity increased, particularly when more comprehensive clinical information was provided.[25]

Case-Based Insight into Diagnostic Reasoning

The AI system and doctors at the intermediate stage of evaluation showed a significant difference in diagnostic reasoning in a representative case of disseminated melioidosis included in this study. Physicians showed early suspicion of melioidosis at Stage 2, when signs of severe sepsis, lung involvement, and epidemiological exposure became apparent, even though both AI and doctors initially thought of typical tropical infections during the early stage of presentation. The AI system, on the other hand, kept giving priority to more prevalent infectious aetiologies and only reached the final diagnosis at the advanced stage after conclusive microbiological proof.

This observation highlights an important limitation of AI-based diagnostic models in the context of rare or region-specific diseases. AI systems predominantly rely on pattern recognition derived from high-frequency data, which may lead to delayed consideration of less common diagnoses such as melioidosis. In contrast, physicians incorporate contextual clinical reasoning, including geographic exposure, disease severity progression, and atypical clinical patterns, enabling earlier diagnostic suspicion. This reinforces the complementary role of clinical judgment in guiding diagnosis, particularly in complex or uncommon clinical scenarios where epidemiological context plays a critical role. AI versus physician diagnostic accuracy across studies was shown in table 13.

Table 13: AI versus Physician Diagnostic Accuracy across Studies

Study	AI System	Physician	AI Accuracy / Rank Metrics	Physician Accuracy / Rank Metrics	Key Finding
Current Study	Perplexity Pro	Physicians 1–3 & Consensus	100% overall; 64.5% top-rank	87.1–90.3%; 19.4%, 9.7% lower ranks	AI maintained perfect accuracy and prioritized top diagnosis across complexity
Hirosawa et al[12]	GPT-3	Physicians	~93% top-rank accuracy	~80%	AI placed correct diagnosis more often higher in ranked lists
Rodman et al[21]	GPT-4	Internists / Residents	Higher at final stage	Lower than AI	AI improved with added clinical information
Takita et al[20]	Generative AI (meta-analysis)	Physicians	Comparable → slightly higher in some cases	Comparable	AI frequently matched or exceeded physician accuracy
Shen et al[3]	ChatGPT-4	Physicians	Higher top-1/5/10 rates in some versions	Variable	AI model performance comparable or better than physicians
Gräff et al[14]	Ada symptom checker	33 Rheumatologists	Top-1 70%; top-3 59%	Top-1 54%; top-3 42%	AI outperformed physicians in ranked diagnostic accuracy
Dinc M et al[25]	GPT-4	Certified Physicians	Strong diagnostic performance	Clinical diagnoses by doctors	AI showed high capability across multiple specialties

LIMITATIONS OF THE STUDY

The study was conducted using a limited number of clinical cases derived from a single tertiary care teaching hospital, which may restrict the generalizability of the findings to other clinical settings. The use of case vignettes, although useful for standardized comparison, may not fully replicate real-time clinical decision-making where physicians have access to dynamic patient interactions and additional contextual information.

It is also important to note that the cases included in this study were previously evaluated in detail by a dedicated clinical unit, where diagnoses were established after thorough workup and multidisciplinary discussion; the structured case vignettes were subsequently derived from these finalized cases. This approach, while ensuring diagnostic clarity, may have reduced the level of ambiguity typically encountered in routine clinical practice. The seemingly perfect diagnostic accuracy observed for the AI system in this study may, in part, be attributed to the structured nature of vignette-based data, particularly where clear diagnostic clues are presented in later stages. Furthermore, real-world clinical environments are inherently more complex, and such factors may influence AI performance differently. Additionally, the diagnostic responses generated by the AI model may vary depending on prompt structure, and the performance observed in this study may not necessarily represent the behavior of future versions of the model.

RECOMMENDATIONS

The findings of this study suggest that large language model–based AI systems may have potential as supportive tools in clinical diagnostic reasoning when used alongside physician judgment. Future studies with larger sample sizes and multicenter designs are recommended to further validate the diagnostic performance of such systems in diverse clinical environments. Integration of AI-assisted diagnostic tools into clinical workflows should be approached cautiously, emphasizing their role as decision-support systems rather than replacements for physician expertise.

CONCLUSION

The present study evaluated the diagnostic accuracy and agreement between Perplexity Pro and physicians using stepwise inpatient clinical case vignettes. The results provide preliminary insights into the ability of large language model–based AI systems to generate relevant differential diagnoses across different stages of clinical information. Although physicians demonstrated stronger diagnostic reasoning in several scenarios, the AI model showed the capacity to produce clinically meaningful differentials, particularly as additional diagnostic information became available. These findings highlight the potential role of AI-assisted tools as complementary aids in clinical decision-making while reinforcing the importance of physician oversight in patient care.

The present study evaluated the diagnostic accuracy and agreement between Perplexity Pro and physicians using stepwise inpatient clinical case vignettes. The results provide preliminary insights into the ability of large language model–based AI systems to generate relevant differential diagnoses across different stages of clinical information. Although physicians demonstrated stronger

diagnostic reasoning in several scenarios, the AI model showed the capacity to produce clinically meaningful differentials, particularly as additional diagnostic information became available. These findings highlight the potential role of AI-assisted tools as complementary aids in clinical decision-making while reinforcing the importance of physician oversight in patient care.

Beyond the findings of this study, it is important to recognize that clinical decision-making is inherently dynamic and context-dependent. Physicians integrate clinical experience, contextual judgment, and patient-specific factors in a manner that extends beyond pattern recognition. While AI systems are capable of processing large volumes of information and identifying relevant diagnostic possibilities, their role remains supportive rather than substitutive.

The integration of AI into clinical practice should therefore be viewed as an adjunct to physician expertise, with the potential to enhance diagnostic efficiency, improve consistency, and assist in structured clinical reasoning. This is particularly relevant in healthcare settings with limited access to trained physicians, where AI-assisted tools may contribute to preliminary assessment and support clinical workflows, while ensuring that final decisions remain under appropriate medical supervision.

Further research involving larger datasets, real-time clinical validation, and diverse patient populations is necessary to better define the scope, limitations, and optimal integration of AI systems within routine clinical practice.

IMPLICATIONS

This study highlights the potential role of AI-based large language models as adjunctive clinical decision-support tools in inpatient medicine. By assisting clinicians in generating and refining differential diagnoses as additional clinical information becomes available, AI systems may help enhance diagnostic consistency, reduce cognitive bias, and support clinical reasoning in complex cases. The progressive improvement in diagnostic accuracy across stages underscores the importance of integrating AI outputs within the broader context of evolving clinical data. Rather than replacing physician judgment, AI tools may serve as supportive systems that complement clinician expertise and facilitate more structured diagnostic reasoning. In addition, AI-assisted diagnostic systems may be particularly valuable in healthcare settings with limited availability of physicians or specialist expertise, where such tools could support clinical decision-making and assist in the management of complex cases. Future research involving larger multicenter datasets and diverse clinical scenarios is needed to further evaluate the generalizability, safety, and real-world clinical impact of AI-assisted diagnostic support systems.

Acknowledgements: We would like to thank the College Management for their logistic support and the Faculty of Department of General Medicine for their guidance throughout the study.

Funding: Nil

Conflict of interest: None

Ethical approval: Approved by the Institutional Ethics Committee of Pondicherry University.

REFERENCES

1. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. Injury. 2023;54:S69–73.
2. Tsao H, Olazagasti JM, Cordoro KM, Brewer JD, Taylor SC, Bordeaux JS, et al. Early detection of melanoma: Reviewing the ABCDEs. Journal of the American Academy of Dermatology. 2015;72(4):717–23.
3. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. JMIR Med Inform. 2019;7(3):e10010.
4. Muddana C, Wang B, Sun PT, Tang YJ. Comparative evaluation of large language models for biotechnology review writing. Biotechnology Advances. 2026; 88:108814.
5. Pettit RW, Fullem R, Cheng C, Amos CI. Artificial intelligence, machine learning, and deep learning for clinical outcome prediction. Emerg Top Life Sci. 2021;5(6):729–45.
6. Uzun Ozsahin D, Ozgocmen C, Balcioglu O, Ozsahin I, Uzun B. Diagnostic AI and Cardiac Diseases. Diagnostics (Basel). 2022;12(12):2901.
7. Nichols JA, Herbert Chan HW, Baker MAB. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev. 2018;11(1):111–8.
8. Machine learning and artificial intelligence in research and healthcare - Injury [Internet]. [cited 2026 Mar 9]. Available from: https://www.injuryjournal.com/article/S0020-1383(22)00076-6/abstract
9. Ramesh AN, Kambhampati C, Monson JRT, Drew PJ. Artificial intelligence in medicine. Ann R Coll Surg Engl. 2004;86(5):334–8.
10. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and Diagnostic Accuracy of Online Symptom Checkers: Systematic Review. J Med Internet Res. 2023;25: e43803.
11. Omron R, Kotwal S, Garibaldi BT, Newman-Toker DE. The Diagnostic Performance Feedback "Calibration Gap": Why Clinical Experience Alone Is Not Enough to Prevent Serious Diagnostic Errors. AEM Educ Train. 2018;2(4):339–42.
12. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y et al. ChatGPT-Generated Differential Diagnosis Lists for Complex Case–Derived Clinical Vignettes: Vol. 11. 2023;11(e48808).
13. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, et al. Measuring the

Impact of AI in the Diagnosis of Hospitalized Patients. JAMA. 2023;330(23):2275–84.
14. Gräf M, Knitza J, Leipe J, Krusche M, Welcker M, Kuhn S, et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int. 2022;42(12):2167–76.
15. Harada Y, Katsukura S, Kawamura R, Shimizu T. Efficacy of Artificial-Intelligence-Driven Differential-Diagnosis List on the Diagnostic Accuracy of Physicians: An Open-Label Randomized Controlled Study. Int J Environ Res Public Health. 2021;18(4):2086.
16. Ovid [Internet]. [cited 2026 Mar 13]. Doctor Versus Artificial Intelligence: Arthritis & Rheumatology. Available from: https://www.ovid.com/journals/arrh/fulltext/10.1002/art.42737~doctor-versus-artificial-intelligence-patient-and-physician doi:10.1002/art.42737
17. JMIR AI - Evaluating the Diagnostic Performance of Symptom Checkers: Clinical Vignette Study [Internet]. [cited 2026 Mar 13]. Available from: https://ai.jmir.org/2024/1/e46875?utm_source=chatgpt.com
18. Ben-Shabat N, Sloma A, Weizman T, Kiderman D, Amital H. Assessing the Performance of a New Artificial Intelligence–Driven Diagnostic Support Tool Using Medical Board Exam Simulations: Clinical Vignette Study. JMIR Medical Informatics. 2021;9(11):e32507.
19. Miranda J, Pereira-Silva R, Guichard J, Meneses J, Carreira AN, Seixas D. Artificial Intelligence Outperforms Physicians in General Medical Knowledge, Except in the Paediatrics Domain: A Cross-Sectional Study. Bioengineering. 2025;12(6):653.
20. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. npj Digit Med. 2025;8(1):175.
21. Ovid [Internet]. [cited 2026 Mar 13]. Artificial Intelligence vs Clinician Performance: JAMA Network Open. Available from: https://www.ovid.com/journals/janop/fulltext/10.1001/jamanetworkopen.2023.47075~artificial-intelligence-vs-clinician-performance-in
22. Ovid [Internet]. [cited 2026 Mar 13]. Large Language Model Influence on Diagnostic... : JAMA Network Open. Available from: https://www.ovid.com/journals/janop/fulltext/10.1001/jamanetworkopen.2024.40969~large-language-model-influence-on-diagnostic-reasoning-a
23. Fukuzawa F, Yanagita Y, Yokokawa D, Uchida S, Yamashita S, Li Y, et al. Importance of Patient History in Artificial Intelligence–Assisted Medical Diagnosis: Comparison Study. JMIR Med Educ. 2024;10:e52674.
24. Belenje A, Pandya D, Jalali S, Rani PK. Accuracy of Artificial Intelligence Versus Clinicians in Real-Life Case Scenarios of Retinopathy of Prematurity. Cureus. 2025;17(2):e78597.
25. Dinc MT, Bardak AE, Bahar F, Noronha C. Comparative analysis of large language models in clinical diagnosis: performance evaluation across common and complex medical cases. Jamia Open. 2025;8(3):ooaf055.