Volume 4, Issue 3, 2026

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

COMPARISON OF ARTIFICIAL INTELLIGENCE BASED DIAGNOSIS WITH CLINICAL DIAGNOSIS IN THE EMERGENCY DEPARTMENT OF A TERTIARY CARE HOSPITAL

[1]Muneera Waris, [2]Uzair Ullah, [3]Ijaz Ahmad, [4]Habib Ullah Jan, [5]Imran Khan, [*6]Kamil Khan

[1]Rehman Medical Institute, Peshawar, Khyber Pakhtunkhwa, Pakistan
[2]Rehman Medical Institute, Peshawar, Khyber Pakhtunkhwa, Pakistan
[3]Rehman Medical Institute, Peshawar, Khyber Pakhtunkhwa, Pakistan
[4]Rehman Medical Institute, Peshawar, Khyber Pakhtunkhwa, Pakistan
[5]Rehman Medical Institute, Peshawar, Khyber Pakhtunkhwa, Pakistan
[*6]Rehman Medical Institute, Peshawar, Khyber Pakhtunkhwa, Pakistan
[5]imranrcahs@gmail.com, [*6]khankamil807@gmail.com

DOI: https://doi.org/10.5281/zenodo.21496032

Keywords:
Artificial Intelligence; Emergency Department; Clinical Diagnosis; Diagnostic Agreement

Article History
Received on 26 Feb, 2026
Accepted on 21 Mar, 2026
Published on 22 Mar, 2026

Abstract

Background: Emergency departments (EDs) require rapid, accurate diagnosis, yet clinical decision-making is frequently constrained by time pressure, high patient volumes, and variable disease presentation. Artificial intelligence (AI), particularly large language models such as ChatGPT, has attracted growing interest as a diagnostic support tool, but evidence from EDs in developing countries remains scarce. Objective: To compare AI-based diagnosis generated by ChatGPT with clinical diagnosis made by physicians in the ED of Rehman Medical Institute (RMI), Peshawar. Methods: This retrospective, comparative cross-sectional study analyzed secondary data from 384 patient records (January 2025–April 2026), selected by non-probability consecutive sampling. De-identified demographic, symptom, vital-sign, and laboratory data (excluding clinical diagnosis) were entered into ChatGPT to generate an independent AI-based diagnosis, which was then compared with the physician's documented diagnosis. Data were analyzed in SPSS v26 using descriptive statistics, cross-tabulation, percentage agreement, and chi-square testing. Results: Exact agreement between AI and physician diagnosis was observed in 87.0% of cases (334/384), an overall diagnostic accuracy of 86.98%. When cases where AI matched with added clinical detail or was clinically close to the physician's diagnosis were included, agreement reached 92.4% (355/384). Agreement was highest for neurological (93.1%), respiratory (92.9%), and renal/genitourinary (84.6–85.6%) presentations, and lowest for infectious disease presentations (82.8%). None of the six organ-system subgroup comparisons reached statistical significance (all $p > 0.05$), with cardiac presentations showing the strongest trend ($p = 0.069$). Conclusion: ChatGPT demonstrated substantial, clinically meaningful agreement with physician diagnosis in a real-world, resource-limited ED setting, supporting its potential as an adjunct decision-support tool. AI cannot substitute for physician judgment and should be integrated alongside, not in place of, clinical expertise.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

1. Introduction

The emergency department (ED) is the first point of contact for patients with acute, potentially life-threatening, or diagnostically uncertain conditions, and its effectiveness depends heavily on prompt, accurate diagnosis (1, 22). Emergency presentations span nearly every organ system, including cardiovascular, respiratory, neurological, gastrointestinal, renal, infectious, and metabolic, each with overlapping symptomatology that complicates rapid decision-making (24). Triage systems such as the Emergency Severity Index help prioritize care by urgency (2, 23), but diagnostic accuracy ultimately depends on clinical judgment, which is vulnerable to cognitive biases such as anchoring and premature closure, and to systemic pressures such as overcrowding and fatigue (3, 4, 21).

Artificial intelligence (AI), meaning computer systems capable of learning, reasoning, and decision-making (7), has expanded rapidly into healthcare, with demonstrated success in radiology, dermatology, and ophthalmology, at times matching specialist-level accuracy (8, 9, 16, 17). Systematic reviews report that AI systems can achieve diagnostic performance comparable to healthcare professionals across specialties (10, 15), and AI is increasingly explored for predictive analytics, triage support, and early detection of time-sensitive conditions such as sepsis (11, 20).

Large language models (LLMs) such as ChatGPT extend this capability to unstructured clinical text, interpreting symptoms and history to propose a differential diagnosis. Early evaluations suggest meaningful promise in clinical reasoning tasks, though performance remains sensitive to input quality (12). However, AI systems remain dependent on data quality and lack the contextual, ethical, and patient-specific reasoning that clinicians bring to the bedside (5, 25); most existing evidence also comes from high-resource settings, leaving a clear gap for EDs in developing countries such as Pakistan, where healthcare infrastructure and disease burden differ substantially (11).

This study therefore aimed to compare AI-based diagnosis generated by ChatGPT with clinical diagnosis made by physicians in the ED of Rehman Medical Institute, Peshawar, addressing the research question of what is the level of agreement between AI-based and clinical diagnosis in this setting.

2. Materials and Methods

2.1 Study Design and Setting

This was a retrospective, comparative cross-sectional study conducted in the ED of Rehman Medical Institute (RMI), a tertiary care hospital in Peshawar, Khyber Pakhtunkhwa, Pakistan. Secondary data were drawn from medical records of patients presenting between January 2025 and April 2026; data extraction and analysis were completed over six months.

2.2 Sample and Sampling

Sample size was calculated using the WHO formula for prevalence studies ($n = Z^2P(1-P)/d^2$), with $P = 0.5$, $Z = 1.96$ (95% confidence level), and $d = 0.05$ margin of error, yielding a required sample of 384, confirmed using the OpenEpi calculator. A non-probability consecutive sampling technique was used, enrolling eligible records sequentially until the target was reached.

Inclusion criteria: records of patients of all ages and genders presenting with acute or chronic conditions, with complete history, vital signs, examination findings, and laboratory investigations.

Exclusion criteria: incomplete records, patients who left against medical advice, and trauma only presentations.

2.3 Data Collection and AI Diagnosis Generation

A structured proforma was used to extract demographics, chief complaint, vital signs, Glasgow Coma Scale, pupillary findings, blood glucose, relevant laboratory investigations, and the organ system involved. This information, excluding the clinical diagnosis and organ-system classification, was entered into ChatGPT to generate an independent, blinded AI-based diagnosis, which was subsequently compared against the physician's documented clinical diagnosis.

2.4 Variables and Statistical Analysis

Data were analyzed using SPSS version 26. Continuous variables were reported as mean ± SD, while categorical variables were presented as frequencies and percentages. Cross-tabulation and percentage agreement quantified overall

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

agreement, and chi-square testing assessed association within each of six organ-system subgroups (cardiac, respiratory, neurological, gastrointestinal, renal/genitourinary, infectious).

2.5 Ethical Considerations

Ethical approval was obtained from the Research Ethics Committee of Rehman College of Allied Health Sciences, with institutional permission from RMI. As the study used de-identified secondary data, no direct patient contact occurred; patient confidentiality was maintained throughout and data were used solely for research purposes.

3. Results

3.1 Demographic and Clinical Characteristics

Young adults (18–44 years) comprised the largest age group (37.2%, n = 143), followed by young-old patients (60–74 years, 26.0%, n = 100) and middle-aged adults (45–59 years, 21.4%, n = 82); patients aged ≥60 years together accounted for 35.6% of the sample, indicating a substantial elderly emergency burden. Female patients were a slight majority (53.9%, n = 207). Duration of symptoms before presentation was fairly evenly distributed: 38.0% presented within 24 hours, 30.5% within 1–3 days, and 31.5% after more than 3 days, suggesting delayed health-seeking behavior in a meaningful proportion of the cohort (Table 1).

Table 1. *Demographic and clinical presentation characteristics (N = 384)*

Characteristic	Category	n	%
Age group	Juvenile (<18 y)	22	5.7
	Young adult (18–44 y)	143	37.2
	Middle-aged (45–59 y)	82	21.4
	Young-old (60–74 y)	100	26.0
	Aged (≥75 y)	37	9.6
Gender	Female	207	53.9
	Male	177	46.1
Duration of symptoms	< 24 hours	146	38.0
	1–3 days	117	30.5
	> 3 days	121	31.5

Vital-sign abnormalities were common across the cohort: only 32.3% of patients were normotensive, 20.8% had Stage 2 hypertension, and 7.3% presented in hypertensive crisis; tachycardia was the predominant heart-rate abnormality (32.6%); 21.6–2.6% of patients showed varying degrees of hypoxemia; and hyperglycemia was documented in 20.1% of cases, consistent with a high local burden of diabetes as a comorbidity. Neurological status was largely preserved (GCS 13–15 in 94.5%) (Table 2). By organ system, gastrointestinal presentations were most common (30.5%, n = 118), followed by renal/genitourinary (20.4%, n = 78), neurological (18.8%, n = 72), cardiac (15.4%, n = 59), respiratory (11.0%, n = 42), and infectious (7.6%, n = 29).

Table 2. *Vital signs and clinical parameters at presentation (selected findings)*

Parameter	Key findings	n	%
Blood pressure	Normotensive	124	32.3
	Stage 2 hypertension (>140/>90 mmHg)	80	20.8
	Hypertensive crisis (>180/120 mmHg)	28	7.3
	Hypotension	30	7.8
Heart rate	Normal	243	63.3
	Tachycardia	125	32.6
	Bradycardia	14	3.6
Respiratory rate	Normal	316	82.3
	Tachypnea	67	17.4
Temperature	Normal	313	81.5
	Low-grade fever (99.1–100.4°F)	49	12.8

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Oxygen saturation	Normal (95–100%)	262	68.2
	Insufficient (91–94%)	83	21.6
	Severe hypoxemia (<80%)	10	2.6
GCS	Mild (13–15)	363	94.5
	Moderate (9–12)	12	3.1
	Severe (≤8)	9	2.3
Blood glucose	Normal	288	75.0
	Hyperglycemia (>140 mg/dL)	77	20.1
	Hypoglycemia (<70 mg/dL)	19	4.9

3.2 Overall Diagnostic Agreement

Exact agreement between AI-generated and physician diagnoses, defined as identical diagnoses, was observed in 334 of 384 cases (87.0%), yielding an overall diagnostic accuracy of 86.98% [(334 ÷ 384) × 100]. When agreement was categorized more granularly, 48.2% of cases showed exact ("Same") agreement, 27.6% were "Same and Detailed" (AI matched the physician's diagnosis while adding clinically relevant detail), and 16.7% were "Close to Clinical Diagnosis" (differing terminology but clinically appropriate); together these three categories yielded 92.4% clinically acceptable agreement (355/384). Complete disagreement occurred in 7.6% of cases (Table 3).

Table 3. *Categorization of diagnostic agreement between AI and physician diagnosis*

Category	n	%
Same (exact agreement)	185	48.2
Same and detailed	106	27.6
Close to clinical diagnosis	64	16.7
Different (disagreement)	29	7.6

3.3 Agreement by Organ System

Diagnostic agreement varied by clinical category, ranging from 79.7% (cardiac) to 93.1% (neurological). Chi-square testing did not reach statistical significance for any of the six organ systems (all $p > 0.05$); the cardiac subgroup showed the closest approach to significance ($\chi^2 = 3.30$, $p = 0.069$), while gastrointestinal, renal, and infectious subgroups had p-values above 0.45, suggesting agreement in these categories was statistically indistinguishable from chance, most plausibly reflecting limited within-subgroup sample sizes rather than a true absence of diagnostic utility (Table 4).

Table 4. *Diagnostic agreement between AI and clinical diagnosis by organ system*

Organ system	n (%) of cohort	Agreement, % (n/N)	χ^2 (df)	p-value
Cardiac	59 (15.4)	79.7 (47/59)	3.30 (1)	0.069
Respiratory	42 (11.0)	92.9 (39/42)	1.44 (1)	0.230
Neurological	72 (18.8)	93.1 (67/72)	3.26 (2)	0.196
Gastrointestinal	118 (30.5)	85.6 (101/118)	0.29 (1)	0.591
Renal / genitourinary	78 (20.4)	84.6 (66/78)	0.48 (1)	0.487
Infectious	29 (7.6)	82.8 (24/29)	0.48 (1)	0.486

4. Discussion

This study found substantial agreement between AI-generated and physician diagnoses in a real-world ED setting, with 87.0% exact agreement and 92.4% clinically acceptable agreement. These findings align with, and in some respects exceed, prior benchmarks: Liu et al. reported pooled AI sensitivity and specificity exceeding 80% across specialties (10), and Shen et al. similarly found AI accuracy comparable to clinicians across multiple scenarios (15). The performance observed here is also consistent with Shieh et al.'s finding that ChatGPT shows promising diagnostic capability in simulated clinical cases (12), extending that evidence to real-world, resource-limited practice.

Agreement was highest in conditions with structured, well-defined symptom patterns, namely neurological (93.1%), respiratory (92.9%), and

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

renal/genitourinary (84.6–85.6%) presentations, consistent with prior evidence that AI performs most reliably where diagnostic criteria are clear-cut, as seen in imaging-based detection of pneumonia and other structured conditions (6, 9). Cardiac presentations also showed meaningful agreement (79.7%), plausibly reflecting the distinctive vital-sign and symptom clusters typical of acute coronary syndrome and hypertensive emergencies. By contrast, infectious disease presentations showed the lowest agreement (82.8%), likely reflecting overlapping, non-specific symptomatology and the need for contextual clinical judgment, a pattern consistent with literature noting that AI performance declines in complex, multi-system scenarios requiring nuanced interpretation (5).

Two patients with trauma-only presentations were inadvertently included despite exclusion criteria; given the large sample size, this is unlikely to have materially affected findings. Notably, in a small number of cases the AI-generated diagnosis was more specific than the documented clinical diagnosis, illustrating both the potential and the interpretive complexity of AI-assisted diagnosis.

The demographic profile, marked by a predominance of young adults and a substantial elderly burden (35.6% aged ≥60), alongside frequent vital-sign derangements such as Stage 2 hypertension (20.8%) and tachycardia (32.6%), reflects the acuity typical of a busy tertiary-care ED in a developing-country setting, and underscores the diagnostic complexity that any AI tool deployed in this context must navigate.

4.1 Strengths and Limitations

Key strengths include the use of real-world (rather than simulated) patient data, direct comparison against physician diagnosis across a broad range of organ systems, and evaluation in a resource-limited developing-country setting, an important and underrepresented context in the existing literature, supported by a reasonably large sample (n = 384).

Limitations include reliance on secondary, retrospectively extracted data, which may carry documentation gaps; generation of AI diagnoses outside a live clinical workflow, which may not fully reflect real-time performance; a single-center design limiting generalizability; and use of ChatGPT, a general-purpose rather than clinically validated diagnostic tool. Future work should consider purpose-built platforms (e.g., Isabel DDx, Ada Health, DXplain), multi-center designs, real-time workflow integration, and more rigorous agreement statistics such as Cohen's Kappa.

5. Conclusion

ChatGPT demonstrated substantial, clinically meaningful agreement with physician diagnosis in the ED of a tertiary-care hospital in Pakistan, with 87.0% exact agreement (86.98% diagnostic accuracy) and 92.4% clinically acceptable agreement. Agreement was strongest for conditions with well-defined clinical profiles and weakest for infectious and multi-system presentations. These findings support the potential of AI as a diagnostic adjunct in emergency medicine, particularly in resource-limited settings, but AI cannot replace the contextual judgment, patient interaction, and ethical reasoning intrinsic to clinical practice. AI should therefore be integrated as a supportive tool alongside, not a substitute for, physician expertise, with the shared goal of improving diagnostic accuracy and efficiency in emergency care.

REFERENCES

1. Carlson LC, Reynolds TA, Wallis LA, Calvello Hynes EJ. Reconceptualizing the role of emergency care in the context of global healthcare delivery. Health Policy and Planning. 2019;34(1):78–82.
2. Hinson JS, Martinez DA, Cabral S, George K, Whalen M, Hansoti B, et al. Triage performance in emergency medicine: a systematic review. Annals of Emergency Medicine. 2019;74(1):140–52.
3. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of Internal Medicine. 2005;165(13):1493–9.
4. Newman-Toker DE, Peterson SM, Badihian S, Hassoon A, Nassery N, Parizadeh D, et al. Diagnostic errors in the emergency department: a systematic review. 2023.
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019;25(1):44–56.
6. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347–58.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

7. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology. 2017;2(4).
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
9. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Medicine. 2018;15(11):e1002686.
10. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health. 2019;1(6):e271–e97.
11. Di Fazio N, Zanza C, Longhitano Y, Voza A, Balagna R, Mosca S, et al. Artificial intelligence for early diagnosis in emergency department. Journal of Anesthesia, Analgesia and Critical Care. 2026;6(1):7.
12. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Scientific Reports. 2024;14(1):9330.
13. Abdellatif NAM, El-Rawy AS, Abdellattif AR, Al-Shatouri MA. Assessment of artificial intelligence-aided x-ray in diagnosis of bone fractures in emergency setting. Egyptian Journal of Radiology and Nuclear Medicine. 2025;56(1):153.
14. Akhlaghi H, Freeman S, Vari C, McKenna B, Braitberg G, Karro J, et al. Machine learning in clinical practice: evaluation of an artificial intelligence tool after implementation. Emergency Medicine Australasia. 2024;36(1):118–24.
15. Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR Medical Informatics. 2019;7(3):e10010.
16. Ting DS. Artificial intelligence and deep learning in ophthalmology.
17. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–2410.
18. Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. New England Journal of Medicine. 2016;375(13):1216.
19. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digital Medicine. 2020;3(1):41.
20. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digital Medicine. 2020;3(1):17.
21. Pines JM, Hilton JA, Weber EJ, Alkemade AJ, Al Shabanah H, Anderson PD, et al. International perspectives on emergency department crowding. Academic Emergency Medicine. 2011;18(12):1358–70.
22. Shanahan T, Risko N, Razzak J, Bhutta Z. Aligning emergency care with global health priorities. International Journal of Emergency Medicine. 2018;11(1):52.
23. Green NA, Durani Y, Brecher D, DePiero A, Loiselle J, Attia M. Emergency Severity Index version 4: a valid and reliable tool in pediatric emergency department triage. Pediatric Emergency Care. 2012;28(8):753–7.
24. Gok MA. Approach to abdominal trauma in geriatric patient group. Approach to the Geriatric Patients. 2022:145.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

25. Waithira J, Chweya R, Cyprian R. Adversarial debiasing for bias mitigation in healthcare AI systems: a literature review. Open Access Library Journal. 2025;12(5):1–13.