BMC Rheumatology

RESEARCH

Open Access

Performance of the Large Language Models in African rheumatology: a diagnostic test accuracy study of ChatGPT-4, Gemini, Copilot, and Claude artificial intelligence

Yannick Laurent Tchenadoyo Bayala[1*], Wendlassida Joelle Stéphanie Zabsonré/Tiendrebeogo[1], Dieu-Donné Ouedraogo[1], Fulgence Kaboré[1], Charles Sougué[2], Aristide Relwendé Yameogo[3], Wendlassida Martin Nacanabo[4], Ismael Ayouba Tinni[1], Aboubakar Ouedraogo[1] and Yamyellé Enselme Zongo[1]

Abstract

Background Artificial intelligence (AI) tools, particularly Large Language Models (LLMs), are revolutionizing medical practice, including rheumatology. However, their diagnostic capabilities remain underexplored in the African context. To assess the diagnostic accuracy of ChatGPT-4, Gemini, Copilot, and Claude AI in rheumatology within an African population.

Methods This was a cross-sectional analytical study with retrospective data collection, conducted at the Rheumatology Department of Bogodogo University Hospital Center (Burkina Faso) from January 1 to June 30, 2024. Standardized clinical and paraclinical data from 103 patients were submitted to the four AI models. The diagnoses proposed by the AIs were compared to expert-confirmed diagnoses established by a panel of senior rheumatologists. Diagnostic accuracy, sensitivity, specificity, and predictive values were calculated for each AI model.

Results Among the patients enrolled in the study period, infectious diseases constituted the most common diagnostic category, representing 47.57% ($n=49$). ChatGPT-4 achieved the highest diagnostic accuracy (86.41%), followed by Claude AI (85.44%), Copilot (75.73%), and Gemini (71.84%). The inter-model agreement was moderate, with Cohen's kappa coefficients ranging from 0.43 to 0.59. ChatGPT-4 and Claude AI demonstrated high sensitivity (> 90%) for most conditions but had lower performance for neoplastic diseases (sensitivity < 67%). Patients under 50 years old had a significantly higher probability of receiving a correct diagnosis with Copilot (OR = 3.36; 95% CI [1.16–9.71]; $p=0.025$).

Conclusion LLMs, particularly ChatGPT-4 and Claude AI, show high diagnostic capabilities in rheumatology, despite some limitations in specific disease categories.

Clinical trial number Not applicable.

Keywords Artificial intelligence, Large Language Models, Rheumatology, Diagnostic accuracy, Africa

*Correspondence:
Yannick Laurent Tchenadoyo Bayala
bayalayannick7991@gmail.com

Full list of author information is available at the end of the article

Introduction

Rheumatology is a specialized field of internal medicine dedicated to studying, diagnosing, and managing musculoskeletal and autoimmune disorders [1]. Before the advent of artificial intelligence (AI), disease diagnosis in rheumatology relied primarily on clinical expertise, imaging techniques, and laboratory tests, whose limitations in accuracy, reproducibility, and time efficiency have fostered the development and integration of AI-based diagnostic tools [1]. Since November 2022, significant advances in AI technologies have led to the emergence of innovative platforms in rheumatology [2].

AI tools, particularly Large Language Models (LLMs), have played an increasingly significant role in rheumatology diagnosis [3]. LLMs are advanced natural language processing (NLP) systems designed to interpret and produce human-like language [3]. In contrast to conventional supervised deep learning approaches, LLMs leverage self-supervised learning methods to extract patterns and knowledge from extensive unlabelled datasets [4]. Subsequently, their performance on specific tasks is optimized through fine-tuning using smaller, annotated datasets tailored to the intended application [4]. Recent LLMs integrate advanced algorithms and machine learning methods that have the potential to offer a wide range of applications in rheumatology [4]. These technologies can contribute to various aspects of practice, including disease diagnosis, therapeutic decision support, prediction of adverse drug events, and the development of personalized treatment strategies. By improving data analysis and supporting clinical reasoning, these LLMs hold promise for optimizing the accuracy and efficiency of musculoskeletal and autoimmune diseases decision-making. The strength of these LLMs lies in their ability to process large volumes of medical data, enabling them to detect patterns that may not be immediately apparent to clinicians [5].

ChatGPT-4 remains the most widely used LLM to assess the diagnostic performance of AI in rheumatology, as in other medical specialties, due to its advanced reasoning capabilities, superior accuracy, and broad validation in healthcare applications [4]. A study conducted in 2023 using multiple-choice trivia items, LLMs were evaluated on their ability to assist clinicians in establishing differential diagnoses of rheumatic diseases based on clinical vignettes derived from scenarios [6]. GPT-4 correctly answered 47 (81%) of questions, whereas Claude 1.3 answered 42 (72%) [6]. In a separate study by Enes et al., conducted using board-level rheumatology questions, ChatGPT-4 demonstrated a high diagnostic performance with an accuracy of 86.9%, significantly outperforming Gemini (60.2%) [7]. Beyond ChatGPT-4, other LLMs are increasingly used in clinical settings and deserve to be compared for their potential role in medical decision-making in rheumatology [4]. Copilot, integrated into various Microsoft products and directly embedded in new hardware, is gaining importance in patient interactions; however, its medical capabilities remain poorly studied [4]. Google Gemini, launched in 2023, aims to enhance human-computer interaction, particularly in clinical reasoning [4]. Finally, Claude, developed by Anthropic, has shown promising performance in some diagnostic tasks but still lacks extensive validation in healthcare compared to GPT-4 [4].

Although AI has increasingly been integrated into rheumatology for disease diagnosis, significant geographic disparities persist in its implementation [8]. In particular, the lower penetration of internet access and digital technologies across the African continent may limit the availability and utilization of AI-based diagnostic tools [8]. This digital divide may also contribute to reduced awareness and adoption of AI in clinical practice compared to other world regions, although no knowledge-attitudes-practice survey has yet confirmed this assumption [8]. Furthermore, to the best of our knowledge, no diagnostic test accuracy study has evaluated the performance of LLMs such as ChatGPT-4, Gemini, Copilot, and Claude AI in rheumatology within an African context. Addressing these gaps is essential to understand the applicability and limitations of these models in resource-limited settings. Therefore, this study aimed to assess the diagnostic accuracy, sensitivity, and specificity of ChatGPT-4, Gemini, Copilot, and Claude AI in rheumatology using clinical vignettes derived from African patients. Additionally, we sought to evaluate their ability to generate differential diagnoses and the associated confidence levels, with the ultimate goal of informing their potential integration into rheumatology practice in Africa.

Methods

This study was conducted and reported in accordance with the STARD (Standards for Reporting of Diagnostic Accuracy Studies) guidelines [9].

Study design

This was a cross-sectional, analytical, and comparative study with retrospective data collection.

Participants

Patients' data were collected retrospectively from hospitalization records over a six-month period (January 1 to June 30, 2024) in the Rheumatology Department of the Bogodogo University Hospital Center, Ouagadougou, a national referral center for rheumatology in Burkina Faso. The department is a specialized unit dedicated to the management of musculoskeletal and autoimmune diseases. The department is staffed by a team of seven

senior rheumatologists, including one full professor, one associate professor, and one rheumatology assistants both considered senior experts in the field. The medical staff also includes four rheumatologists and several rheumatology residents undergoing specialty training. The department does not currently have an electronic health record system; patient data and clinical records are managed using paper-based files. All patients hospitalized for a rheumatologic condition during the study period and who provided informed consent were included. An exhaustive sampling method was used. This approach aimed to include all eligible patients hospitalized during the study period in order to minimize selection bias and ensure the representativeness of the study population. Data were extracted from paper-based files in our center. Patients were excluded if their medical records were deemed unusable, defined as having more than 75% missing data in the data collection form. In other words, a medical record was considered exploitable when at least 75% of the required variables were available. Patients were also excluded in cases of persistent disagreement among senior rheumatologists regarding the final diagnosis. The reference diagnosis (gold standard) corresponded to the diagnosis retained by senior rheumatologists after a collegial discussion.

Tests methods

Index test

These AI models were selected based on their popularity, free access, and widespread use as AI-driven clinical diagnostic tools [10]:

- ChatGPT-4, developed by OpenAI (San Francisco, USA). The study used the version available from December 1 to December 31, 2024, accessible at www.openai.com.
- Copilot, formerly known as Bing AI, developed by Microsoft Corporation (Washington, USA). The version evaluated was available from December 1 to December 31, 2024, at www.microsoft.com.
- Gemini 2.0 Flash, formerly known as Bard, developed by Google LLC (Alphabet Inc.) (California, USA). The version tested was available from January 1 to January 31, 2024, at www.google.com/gemini.
- Claude 3.5 Sonnet, developed by Anthropic (San Francisco, USA). The version evaluated was available from December 1 to December 31, 2024, at www.anthropic.com.

Reference standard

The gold standard diagnosis was determined by a panel of three senior rheumatologists (D-D.O, W.J.S.Z/T, F.K) with at least five years of experience, following consensus during clinical staff meetings. Patients were excluded in cases of persistent disagreement among senior rheumatologists regarding the final diagnosis. Diagnoses were established based on epidemiological, clinical, biological, and radiological criteria.

Data collection and harmonization

Data collection was performed by a team of trained rheumatology residents under the supervision of senior rheumatologists. Before data extraction, the residents received specific training on the use of the standardized data collection form, the definitions of the clinical variables, and the modalities for querying the AI tools used in the study. The extracted data included demographic characteristics (sex, patient origin), medical history, lifestyle factors, physical examination findings, laboratory and imaging results, and the final diagnosis established by the rheumatologist.

Data were standardized prior to submission to AI models, text correction was performed. This involved correcting spelling errors, unifying medical terminology, removing unnecessary or redundant information, and ensuring consistency and readability. The expert rheumatologist diagnoses, as well as radiology and laboratory conclusions, were removed from the records before submission. Radiological images were not directly uploaded to the chatbot systems; instead, detailed written descriptions of the imaging findings were provided as part of the clinical vignettes.

Clinical vignettes were presented in a harmonized, structured format in English, maintaining comprehensive clinical details including history, physical examination findings, and diagnostic test results, with identical prompts utilized across all AI sessions to ensure methodological consistency. The study's rigor was enhanced through a blinded evaluation process where all AI-generated responses were analyzed by a single independent evaluator who possesses both clinical expertise as a rheumatologist and technical knowledge in artificial intelligence, ensuring comprehensive assessment without knowledge of the gold standard diagnoses, employing a validated assessment rubric to minimize cognitive bias during the comparative analysis between AI diagnostic output and reference standards. We conducted two test sessions spaced 10 days apart to minimize response variability.

Evaluation of index test

Each AI model was queried in two phases using identical prompts, with the patient's origin specified: first, based solely on clinical data, followed by a second query after incorporating paraclinical results (Supplementary file 1). To minimize bias, a separate session was initiated for each patient on each AI platform.

- First set of questions (clinical data only):
  - What is the most probable diagnosis?
  - What are the differential diagnoses?
- Second set of questions (after adding paraclinical results):
  - What is the most probable diagnosis?
  - What is your confidence level in percentage?

Definition and rationale for test positivity

The AI-generated diagnoses were standardized using the ICD-10 classification (Supplementary file 2) [11]. AI-generated diagnoses were classified as follows (Supplementary file 2):

- Correct: When they matched exactly with the expert rheumatologist's diagnosis, following ICD-10 criteria.
- Partial: When the AI identified the correct disease category or pathophysiological mechanism but lacked diagnostic specificity (for example, identifying "infectious spondylitis" but not specifying "tuberculous Pott's disease").
- Incorrect: When the diagnosis had no clinical relevance to the rheumatologist's diagnosis.

The quality of differential diagnoses was assessed using Bond et al. ordinal score [12]: (5) The actual diagnosis is included in the differential. (4) A very close suggestion is included. (3) A roughly approximate but useful suggestion is included. (2) A related but unlikely useful suggestion is included. (0) No relevant suggestion.

Analysis

Quantitative variables were summarized as mean ± standard deviation (SD) if normally distributed or median with interquartile range (IQR) if non-normally distributed, after assessing normality using the Shapiro-Wilk test. Qualitative variables were presented as frequencies, percentages, and their respective 95% confidence intervals (CI) calculated using the Wilson method.

The degree of agreement between AI-generated diagnoses was assessed using Cohen's kappa coefficient (κ) with 95% CI [13]. Kappa values were interpreted as: <0.20 (poor), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (good), and 0.81-1.00 (very good agreement). The confidence levels of the AI models were compared using the Friedman test for repeated measures with post-hoc Dunn's multiple comparison tests when significant differences were found. Statistical significance was set at $p < 0.05$.

Diagnostic performance was assessed across etiological groups using sensitivity, specificity, positive predictive value, negative predictive value, and accuracy, all reported with 95% CI. The accuracy of each AI was evaluated using the area under the receiver operating characteristic curve (AUC) with 95% CI, interpreted according to Swets et al. criteria [14]: AUC ≥ 0.80 considered excellent, 0.70–0.79 as good, 0.60–0.69 as fair, and < 0.60 as poor.

To identify factors potentially affecting AI diagnostic performance, we selected variables based on both previous literature for rheumatological conditions and clinical expertise. These factors included patient demographics (age, sex) and etiological groups. Pearson's chi-square test or Fisher's exact test was used in univariate analysis to determine associations between categorical variables and AI diagnostic accuracy. Variables with $p < 0.2$ in univariate analysis were included in a multivariate logistic regression model to assess independent associations with AI diagnostic performance. Odds ratios with 95% CI were calculated, and statistical significance was set at $p < 0.05$.

Results are presented as text supplemented by tables and figures. Data were entered and analyzed using Excel (Microsoft Office Professional Plus 2019, Washington, WA, USA), Epi Info 7.2 (Centers for Disease Control and Prevention, CDC, Atlanta, USA), and GraphPad Prism (version 10.0.2; GraphPad Software Inc., Boston, Massachusetts).

Ethical considerations

This study was conducted in accordance with the principles of the Declaration of Helsinki [15]. Anonymity and data confidentiality were strictly maintained throughout the research process. Informed consent was obtained from all participants prior to enrollment. For participants unable to provide consent due to cognitive impairment or severe illness, consent was obtained from legal guardians or next of kin before any data collection commenced. Special attention was paid to participant inclusivity across demographic categories including age, gender, ethnicity, and socioeconomic status, following recent guidelines on equitable research participation. The study protocol received approval from the Ethics Committee of the Bogodogo University Hospital Center (approval number: N202202-32) before participant recruitment began.

Results

General characteristics of the population

A total of 112 patients were included during the study period, with 9 patients excluded due to incomplete medical records. In total, 103 hospitalization reports were analyzed (Fig. 1). The mean age was 51.9 ± 20.9 years, ranging from 10 to 88 years. The sex ratio was 1.64, with a predominance of male patients. Infectious diseases were the most common diagnoses, accounting for 47.57%

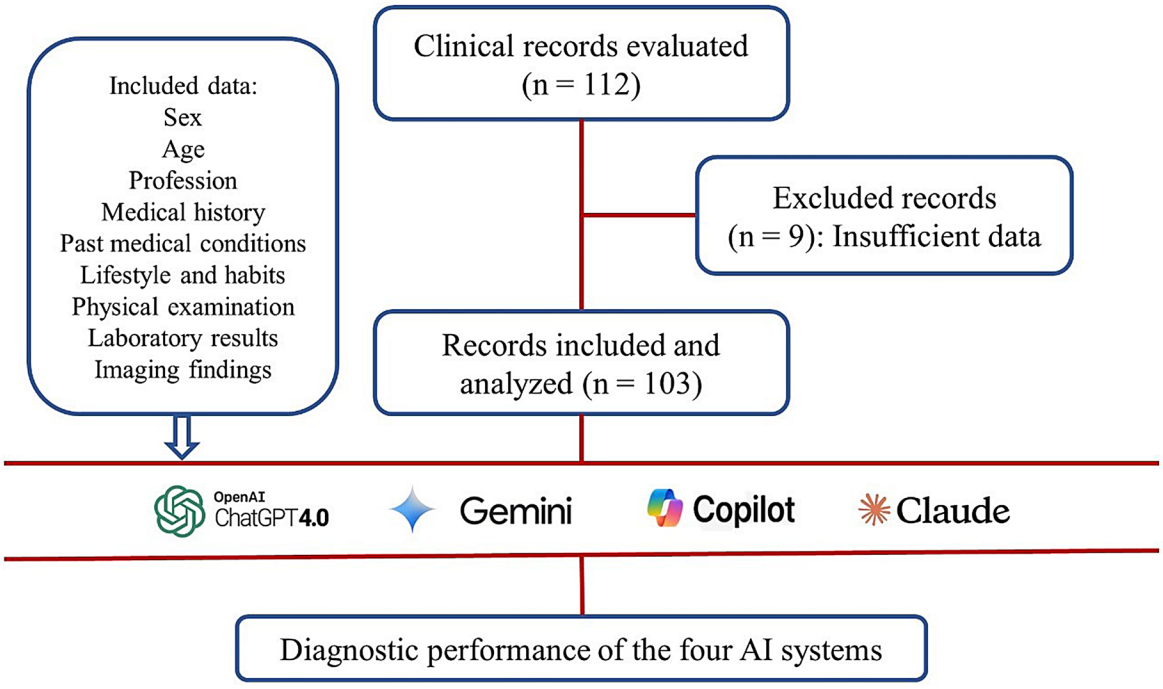

Fig. 1 Flowchart of the study design

($n=49$), followed by chronic inflammatory rheumatic diseases (16.50%, $n=17$). The general characteristics of the study population and the distribution of etiological groups are summarized in Table 1.

Overall performance of AI models

ChatGPT-4 correctly identified the rheumatologist's diagnosis based on clinical data alone in 80 patients (77.66%), compared to 56 patients (54.36%) for Gemini, 58 patients (56.31%) for Copilot, and 79 patients (76.69%) for Claude AI. When combining clinical and paraclinical data, the overall diagnostic accuracy was 86.41% for ChatGPT-4 ($n=89$), 71.84% for Gemini ($n=74$), 75.73% for Copilot ($n=78$), and 85.44% for Claude AI ($n=88$). The concordance between AI models and rheumatologists is illustrated in Fig. 2.

The agreement among AI models was moderate, with Cohen's kappa coefficients ranging from 0.43 to 0.59. The highest concordance was observed between ChatGPT-4 and Copilot ($\kappa=0.59$; 95% CI [0.405–0.786]), followed by Copilot and Claude AI ($\kappa=0.57$; 95% CI [0.377–0.766]) (Table 2).

Using Bond et al. ordinal score, the rheumatologist's diagnosis was included as a differential diagnosis with a score of 5 in 75.72% of cases ($n=78$) for ChatGPT-4, 57.28% ($n=59$) for Gemini, 55.33% ($n=57$) for Copilot, and 75.72% ($n=78$) for Claude AI. The differential diagnosis scoring for each AI model is illustrated in Fig. 3.

Confidence rating

The median global confidence level was 90% (IQR: 85–92.5) for ChatGPT-4, 85% (IQR: 80–90) for Gemini, 90% (IQR: 85–90) for Copilot, and 92.5% (IQR: 90–95) for Claude AI (Fig. 4).

Performance metrics by etiological group

The diagnostic performance of AI models varied significantly across different etiological groups, with notable differences in sensitivity, specificity, and accuracy.

ChatGPT-4 showed the highest sensitivity for infectious diseases (91.83%) and chronic inflammatory rheumatic diseases (94.11%), but with relatively low specificity (18.51% and 15.11%, respectively). Claude AI also demonstrated high sensitivity for infectious diseases (91.83%) and chronic inflammatory rheumatic diseases (94.11%). Degenerative diseases were best diagnosed by ChatGPT-4, with a sensitivity of 86.66%. Gemini exhibited the best balance between sensitivity (80.00%) and specificity (29.54%) for degenerative diseases. Neoplastic diseases remained a challenge for all AI models, with sensitivities not exceeding 67%. The detailed performance metrics for each AI across etiological groups are summarized in Table 3.

ROC curve analysis

In our study, Gemini achieved the highest AUC of 0.633 (95% CI: 0.533–0.726) for chronic inflammatory rheumatic diseases and 0.548 (95% CI: 0.447–0.646) for degenerative diseases. Claude AI and ChatGPT-4

Table 1 General and etiological characteristics of the study population

Variables	*n*	Percentage (%)
Mean age (years)	51.9 ± 20.9	
Age groups		
< 50 years	53	51.46
≥ 50 years	50	48.54
Sex		
Male	64	62.14
Female	39	37.86
Rheumatological disease categories		
Infectious diseases	49	47.57
Pott's disease	23	22.33
Pyogenic spondylodiscitis	7	6.80
Septic arthritis of peripheral joints	10	9.71
Pyogenic zygapophyseal arthritis	2	1.94
Infectious myositis	3	2.91
Acute rheumatic fever	1	0.97
Septic osteonecrosis of the femoral head	2	1.94
Chronic inflammatory rheumatic diseases	17	16.50
Rheumatoid arthritis	7	6.80
Systemic lupus erythematosus	4	3.88
Systemic sclerosis	1	0.97
Dermatomyositis	1	0.97
Post-streptococcal rheumatism	1	0.97
Ankylosing spondylitis	2	1.94
Sjögren's syndrome	1	0.97
Degenerative diseases	15	14.56
Common low back pain in adults	9	8.74
Avascular necrosis of the femoral head	2	1.94
Acute flare of knee osteoarthritis	2	1.94
Common cervical pain	1	0.97
Microcrystalline diseases	13	12.62
Gout	13	12.62
Neoplastic diseases	9	8.74
Spinal bone metastasis	4	3.88
Benign bone tumor of the spine	1	0.97
Multiple myeloma	3	2.91

Table 2 Cohen's kappa coefficient for agreement between AI model pairs

	Coefficient Kappa de Cohen	CI à 95%
ChatGPT-4 vs. Gemini	0.45	[0.265–0.652]
ChatGPT-4 vs. Copilot	0.59	[0.405–0.786]
ChatGPT-4 vs. Claude AI	0.47	[0.235–0.721]
Gemini vs. Copilot	0.44	[0.254–0.643]
Gemini vs. Claude AI	0.43	[0.241–0.633]
Copilot vs. Claude AI	0.57	[0.377–0.766]

displayed similar performance, with AUC values close to 0.55 for infectious diseases and chronic inflammatory rheumatic diseases. However, AI models demonstrated poor classification ability for neoplastic diseases, with AUC values below 0.40, indicating low discrimination capacity in this category. The ROC curves for AI performance by etiological group are presented in Fig. 5.

Factors affecting the diagnostic performance of AI tools

Multivariate analysis revealed that neoplastic diseases were significantly associated with a lower probability of correct diagnosis by ChatGPT-4 (OR = 0.08; 95% CI [0.01–0.45]; $p = 0.004$) and Copilot (OR = 0.09; 95% CI [0.01–0.54]; $p = 0.007$). Conversely, patients under 50 years old had a significantly higher probability of receiving a correct diagnosis with Copilot (OR = 3.36; 95% CI [1.16–9.71]; $p = 0.025$). The complete logistic regression analysis is summarized in Table 4.

Discussion

Principal finding

In our study, ChatGPT-4 demonstrated the highest overall diagnostic accuracy rate (86.41%) defined as the proportion of correctly identified diagnoses out of total cases, for all-cause rheumatological diseases. It followed by Claude AI (85.44%), Copilot (75.73%), and Gemini (71.84%). Gemini achieved the highest AUC of 0.633 (95% CI: 0.533–0.726) for chronic inflammatory

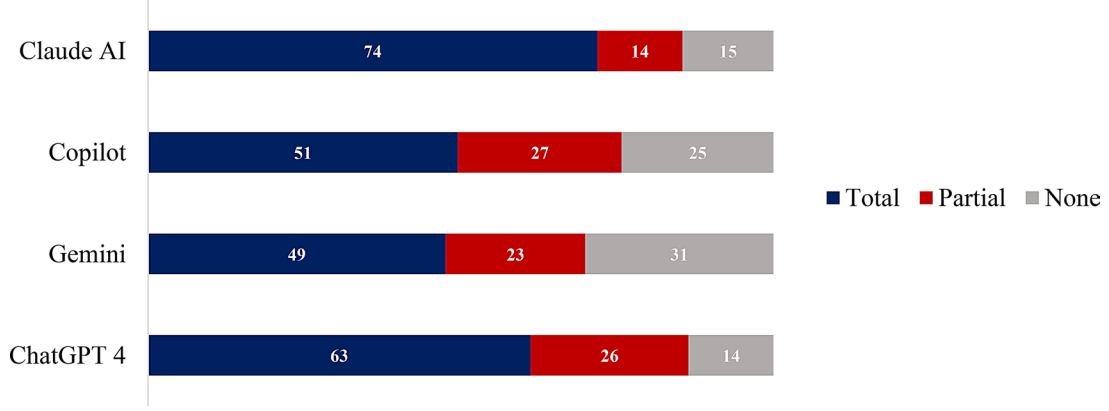

Fig. 2 Stacked bar chart representing the concordance between AI models and the rheumatologist's diagnosis

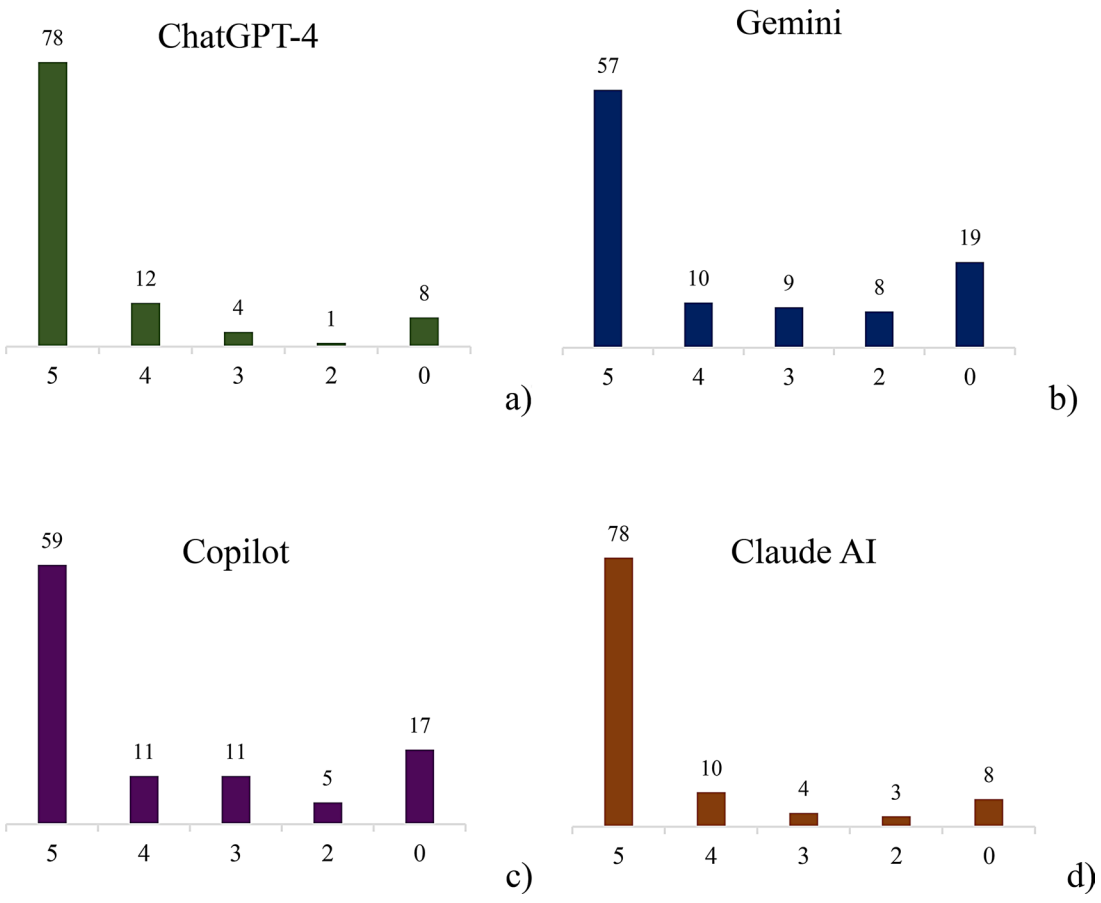

Fig. 3 Histogram of differential diagnosis scores based on Bond et al. ordinal score [6]: (**a**) ChatGPT-4 (**b**) Gemini (**c**) Copilot (**d**) Claude AI. (5) The actual diagnosis is included in the differential. (4) A very close suggestion is included. (3) A roughly approximate but useful suggestion is included. (2) A related but unlikely useful suggestion is included. (0) No relevant suggestion

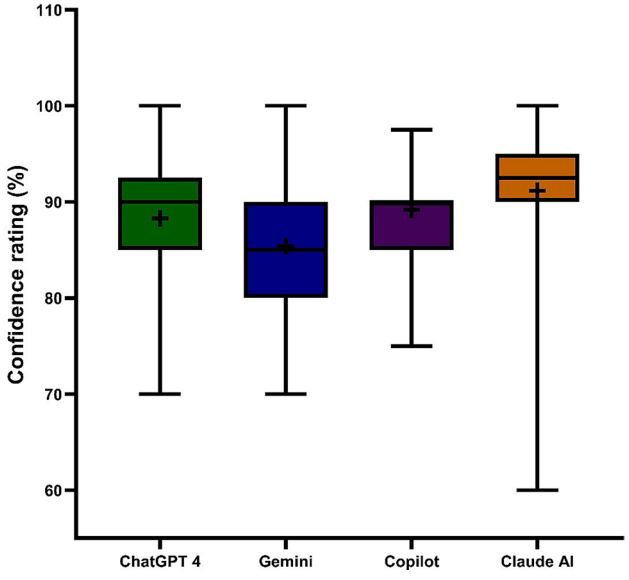

Fig. 4 Boxplot displaying the distribution of AI confidence levels. Black solid line represents the median, boxplot represents the 25th (Q1) and 75th (Q3) percentile. Whiskers range from the minimum to the maximum value. * = $p < 0.0001$

rheumatic diseases and 0.548 (95% CI: 0.447–0.646) for degenerative diseases. Gemini stood out for its best balance between sensitivity and specificity for degenerative diseases, suggesting it may be the most suitable tool for this condition. Conversely, neoplastic musculoskeletal diseases were significantly associated with a lower correct diagnosis rate for ChatGPT-4 (OR = 0.08; 95% CI [0.01–0.45]; $p = 0.004$) and Copilot (OR = 0.09; 95% CI [0.01–0.54]; $p = 0.007$), with AUC values below 0.40 and sensitivities not exceeding 67%.

Interpretation of the principal findings

This difficulty in correctly identifying neoplastic diseases may be due to the rarity and clinical polymorphism of osteoarticular tumors, which can mimic more common conditions such as infections and chronic inflammatory rheumatic diseases [16–20]. Additionally, their diagnosis is often confirmed through histopathological examination, which was unavailable in our study population.

In contrast, patients under 50 years old had a significantly higher probability of receiving a correct diagnosis with Copilot (OR = 3.36; 95% CI [1.16–9.71]; $p = 0.025$).

Table 3 Performance metrics of each AI model based on etiological groups

AI Models	Etiological Groups	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	AUC (95% CI)
ChatGPT-4	Infectious diseases	91.83	18.51	50.56	71.42	53.39	0.552 (0.451–0.650)
	Degenerative diseases	86.66	13.63	14.60	85.71	24.72	0.502 (0.401–0.602)
	Chronic inflammatory rheumatic diseases	94.11	15.11	17.97	92.85	28.15	0.546 (0.445–0.645)
	Microcrystalline diseases	84.61	13.33	12.36	85.71	22.33	0.490 (0.390–0.590)
	Neoplastic diseases	44.44	9.57	4.49	64.28	12.62	0.270 (0.187–0.367)
Gemini	Infectious diseases	69.38	25.92	45.94	48.27	46.60	0.477 (0.377–0.577)
	Degenerative diseases	80.00	29.54	16.21	89.65	36.89	0.548 (0.447–0.646)
	Chronic inflammatory rheumatic diseases	94.11	32.55	21.62	96.55	42.71	0.633 (0.533–0.726)
	Microcrystalline diseases	61.53	26.66	10.81	82.75	31.06	0.441 (0.343-0542)
	Neoplastic diseases	44.44	25.53	5.40	82.75	27.18	0.350 (0.259–0.450)
Copilot	Infectious diseases	81.63	29.63	51.28	64.00	54.36	0.556 (0.455–0.654)
	Degenerative diseases	80.00	25.00	15.38	88.00	33.01	0.525 (0.424–0.624)
	Chronic inflammatory rheumatic diseases	82.35	25.58	17.94	88.00	34.95	0.540 (0.439-0649)
	Microcrystalline diseases	69.23	23.33	11.53	84.00	29.12	0.463 (0.364–0.564)
	Neoplastic diseases	33.33	20.21	3.84	76.00	21.35	0.268 (0.185–0.364)
Claude AI	Infectious diseases	91.83	20.37	51.13	73.33	54.36	0.561 (0.460–0.659)
	Degenerative diseases	73.33	12.50	12.50	73.33	21.35	0.429 (0.332–0.530)
	Chronic inflammatory rheumatic diseases	94.11	16.27	18.18	93.33	29.12	0.552 (0.451–2.568)
	Microcrystalline diseases	76.92	13.33	11.36	80.00	21.35	0.451 (0.353–0.552)
	Neoplastic diseases	66.66	12.76	6.81	80.00	17.47	0.397 (0.302–0.492)

PPV: Positive Predictive Value NPV: Negative Predictive Value

This trend may be attributed to biases in AI training models, which are often based on clinical cases from younger populations, as frequently reported in the literature [21].

Comparison with previous studies

Our findings are generally consistent with existing literature on the diagnostic performance of AI models in rheumatology. Venerito et al. (2023) evaluated three AI models on theoretical rheumatology questions, reporting diagnostic accuracy rates of 81% for GPT-4 and Claude 2, while Claude 1.3 achieved 72% [16]. While their results align with ours, some key differences emerge, particularly in infectious rheumatologic diseases [16]. In our study, diagnostic performance for these conditions ranged from 40 to 55%, whereas in Venerito et al. study, ChatGPT-4's correct response rate dropped to 57%, while Claude 1.3 and Bard did not exceed 14% [16].

Our results also differ from Krusche et al., who compared ChatGPT-4's diagnostic performance to that of rheumatologists. Their study found that ChatGPT-4 correctly identified the final diagnosis in 35% of cases, compared to 39% for rheumatologists [17]. Furthermore, ChatGPT-4 achieved relevant differential diagnoses in 60% of cases, and its accuracy for chronic inflammatory rheumatic diseases was 71%, whereas in our study, it was 94.12%. This discrepancy may be explained by the critical role of immunological tests and imaging in diagnosing chronic inflammatory rheumatic diseases, which were not systematically incorporated into AI prompts in Krusche et al. study [17].

Other studies have confirmed the high diagnostic performance of ChatGPT-4 in rheumatology, particularly in student examination settings. ChatGPT-4o achieved an accuracy of 86.9%, significantly outperforming Gemini (60.2%), with particularly high accuracy in subfields such as osteoarthritis ($p = 0.023$) and rheumatoid arthritis ($p < 0.001$) [18]. Similarly, Madrid-Garcia et al. reported a 93.71% accuracy rate for ChatGPT-4 in rheumatology examinations, suggesting its potential as an educational tool for rheumatology learning [19].

Strengths

Our study offers several methodological strengths. First, it represents the first comparison of four widely accessible AI tools in a clinical rheumatology setting, adhering to established diagnostic test accuracy study guidelines. All data were extracted systematically from paper-based files records using a standardized protocol, ensuring consistency in data collection and minimizing selection bias.

The comprehensive performance assessment using multiple metrics provides a robust evaluation framework. Notably, the AUC analysis revealed Gemini's superior discriminative ability for chronic inflammatory rheumatic diseases (AUC = 0.633; 95% CI: 0.533–0.726) and degenerative conditions (AUC = 0.548; 95% CI: 0.447–0.646), offering insights beyond raw accuracy metrics. Furthermore, our stratified analysis across different etiological groups provides granular performance assessment essential for clinical implementation.

Unlike simulation-based evaluations, our study utilized real clinical cases from routine practice, enhancing

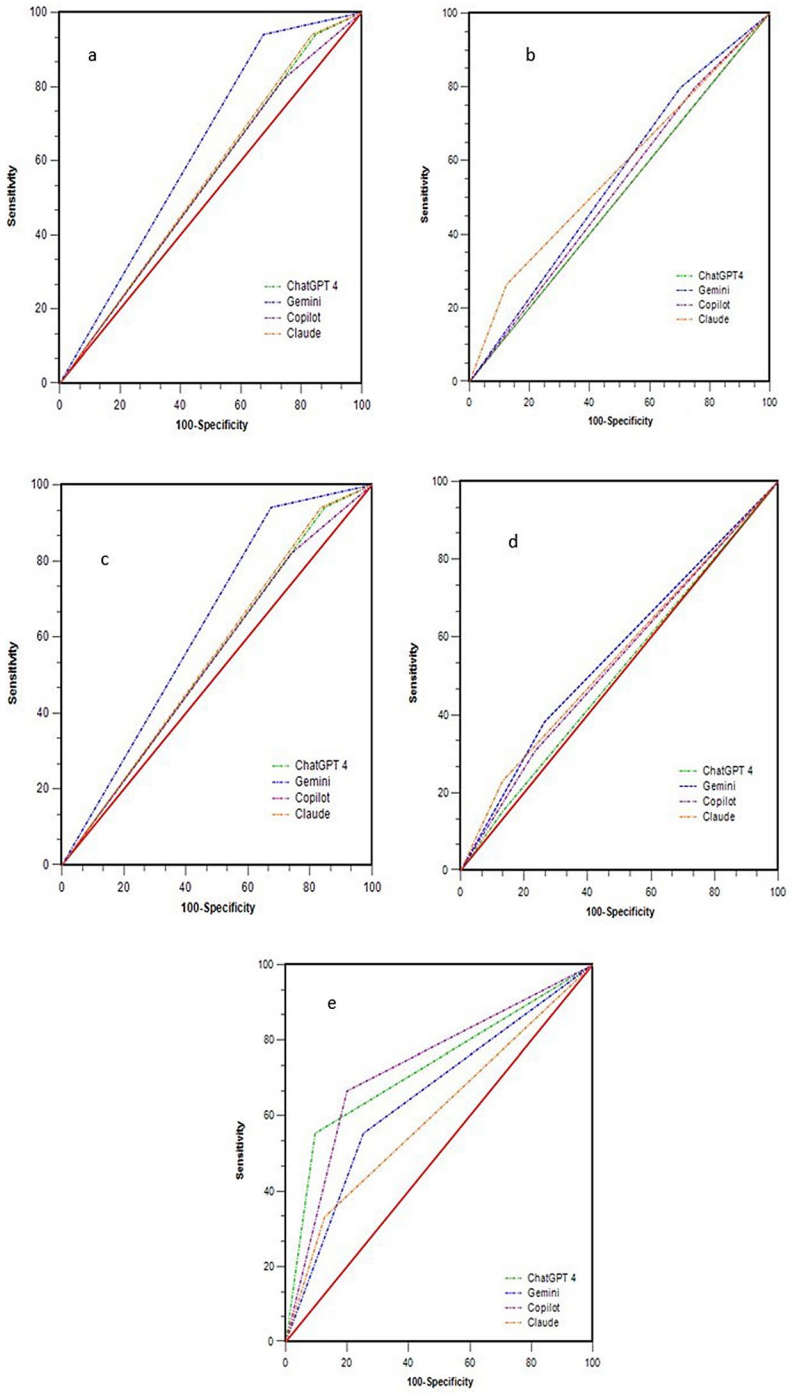

Fig. 5 Receiver Operating Characteristic (ROC) curves for each AI model by etiological group: (**a**) Infectious diseases (**b**) Degenerative diseases (**c**) Chronic inflammatory rheumatic diseases (**d**) Microcrystalline diseases (**e**) Neoplastic diseases

ecological validity and clinical relevance. Additionally, this research represents one of the first diagnostic accuracy studies of AI tools conducted in an African healthcare setting, providing crucial contextual validation necessary for appropriate AI implementation in diverse medical environments, a critical consideration for equitable AI development in medicine across the continent [4].

Limitations

Our study presents several methodological limitations that warrant consideration. This cross-sectional diagnostic test accuracy design, while practical for initial assessment, inherently restricts external validity and application to broader clinical contexts.

Table 4 Univariate analysis and logistic regression of the association between sociodemographic variables, etiological groups, and AI diagnostic performance

Variables		Univariate analysis		Multivariate analysis		
Correct Diagnosis *n* (%)		**OR**	***p* value**	**OR**	**[CI 95%]**	***p* value**
ChatGPT-4						
Age < 50 years	43 (48.31)	0.93	0.454			
Male sex	53 (59.55)	0.40	**0.092**	0.28	[0.06 ; 1.32]	0.108
Infectious diseases	45 (50.56)	2.55	**0.106**	1.49	[0.36 ; 6.05]	0.571
Degenerative diseases	13 (14.61)	1.02	0.644			
Chronic inflammatory rheumatic diseases	16 (17.98)	2.84	0.280			
Microcrystalline diseases	11 (12.36)	0.84	0.558			
Neoplastic diseases	4 (4.49)	0.08	**0.002**	0.08	[0.01; 0.45]	**0.004**
Gemini						
Age < 50 years	38 (51.35)	1.49	**0.187**	1.89	[0.74; 4.81]	0.178
Male sex	42 (56.76)	0.41	**0.056**	0.40	[0.14; 1.14]	0.089
Infectious diseases	34 (45.95)	0.79	0.378			
Degenerative diseases	12 (16.22)	1.67	0.337			
Chronic inflammatory rheumatic diseases	16 (21.62)	7.72	**0.018**	6.13	[0.74; 50.43]	0.091
Microcrystalline diseases	8 (10.81)	0.58	0.281			
Neoplastic diseases	4 (5.41)	0.27	**0.041**	0.27	[0.06; 1.22]	0.089
Copilot						
Age < 50 years	42 (53.85)	2.47	**0.046**	3.36	[1.16; 9.71]	**0.025**
Male sex	46 (58.97)	0.55	**0.176**	0.40	[0.13; 1.20]	0.105
Infectious diseases	40 (51.28)	1.87	**0.135**	1.33	[0.46; 3.80]	0.587
Degenerative diseases	12 (15.38)	1.33	0.480			
Chronic inflammatory rheumatic diseases	14 (17.95)	1.60	0.361			
Microcrystalline diseases	9 (11.54)	0.68	0.389			
Neoplastic diseases	3 (3.85)	0.12	**0.005**	0.09	[0.01; 0.54]	**0.007**
Claude AI						
Age < 50 years	43 (48.86)	1.09	0.549			
Male sex	55 (62.50)	1.11	0.534			
Infectious diseases	45 (51.14)	2.87	**0.068**	3.37	[0.65; 17.50]	0.147
Degenerative diseases	11 (12.50)	0.39	**0.148**	0.82	[0.04; 0.82]	0.826
Chronic inflammatory rheumatic diseases	16 (18.18)	3.11	**0.147**	4.80	[0.43; 52.76]	0.199
Microcrystalline diseases	10 (11.36)	0.51	0.286			
Neoplastic diseases	6 (6.82)	0.29	**0.122**	0.60	[0.09; 3.98]	0.597

A significant limitation concerns our gold standard, the rheumatologist's final diagnosis which, despite being standard clinical practice, lacks the objectivity and reproducibility of universally accepted diagnostic biomarkers. This reference standard may introduce inter-observer variability and potentially limit replication in future studies.

The diagnostic accuracy of AI models demonstrated in this study may vary across different languages, as their performance was only assessed using English clinical vignettes, which limits generalizability to non-English clinical settings including French or Spanish consultations.

Additionally, our study evaluated AI performance using retrospective data from hospitalized patients only, potentially overlooking the heterogeneity of presentations in outpatient settings. The AI models' inability to access longitudinal patient data and disease evolution often crucial diagnostic elements in rheumatology further constrains the clinical applicability of our findings. Moreover, AI diagnostic capabilities are fundamentally bounded by their training datasets, which may harbor inherent biases regarding disease prevalence, demographic representation, and clinical presentations, potentially affecting diagnostic reliability across diverse patient populations.

Implications of our findings for future research, policies, and clinical practice

Our findings have several implications for rheumatological practice in Burkina Faso and in Africa. Given the severe shortage of rheumatologists, AI-assisted diagnostic tools could serve as valuable clinical decision support systems for non-specialist healthcare providers who manage the majority of musculoskeletal conditions. For Africa specifically, the development of context-adapted AI algorithms that account for the local disease profile,

particularly the high prevalence of infectious diseases observed in our study. These tools should be optimized to function with limited paraclinical resources, as many healthcare facilities in Africa operate without advanced imaging or laboratory capabilities.

The implementation strategy should prioritize training programs for general practitioners and nurses at peripheral health centers, where rheumatological expertise is most scarce. Given our finding that ChatGPT-4 and Claude AI demonstrated high sensitivity for infectious diseases and chronic inflammatory rheumatic conditions, these models could be particularly valuable for initial screening and triage in primary care settings throughout Africa [4].

For broader African applications, our results must be contextualized within the digital health landscape described by recent perspectives on rheumatological disease diagnosis in Africa. The diagnostic performance variations we observed across different AI models reinforce the importance of rigorous validation studies in diverse African populations before widespread implementation.

Furthermore, the poor performance of all AI models in detecting neoplastic diseases (AUC values < 0.40) highlights a critical limitation that must be addressed through specialized algorithms and clear clinical guidelines to prevent missed diagnoses of malignancies [21].

Successful integration of AI into rheumatological care in Africa will require multi-stakeholder collaboration including ministries of health, medical associations, patient advocacy groups, and technology developers. Regulatory frameworks must be established to ensure data protection, ethical use, and equitable access across different socioeconomic groups. Educational initiatives should target both healthcare providers and patients to build trust and facilitate appropriate utilization of these technologies.

In the context of patient data confidentiality and healthcare security concerns, future research should explore smaller locally deployable language models like Llama, which could offer viable alternatives for clinical settings where on-premises deployment would mitigate the privacy risks associated with transmitting sensitive patient information to cloud-based AI systems [21].

Finally, future research should focus on developing AI models specifically trained on African patient data to improve diagnostic accuracy for conditions with unique local presentations and to account for regional genetic, environmental, and socioeconomic factors that influence disease manifestation and progression [4].

Conclusion

This study demonstrated that AI models exhibit promising diagnostic capabilities in rheumatology, with remarkable accuracy for ChatGPT-4 (86.41%) and Claude AI (85.44%), followed by Copilot (75.73%) and Gemini (71.84%). These tools were particularly effective in diagnosing infectious diseases and chronic inflammatory rheumatic conditions, with sensitivities exceeding 90% in some models. However, neoplastic diseases were more challenging to identify for ChatGPT-4 and Copilot, reducing their performance in this domain. In contrast, patients under 50 years old had a higher probability of receiving a correct diagnosis with Copilot. These findings highlight both the potential of AI in rheumatology and its diagnostic limitations, particularly for certain disease groups. Further research is needed to evaluate AI integration into clinical practice, including its impact on patient management timelines, cost-effectiveness, and acceptance by both clinicians and patients.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41927-025-00512-z.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

Not applicable.

Author contributions

Y.L.T.B: Contributing to the conceptualization, the data curation, the methodology, the formal analysis, the supervision, the validation and the writing of the original draft. W.J.S.Z/T: Contributing to the conceptualization, the data curation and the methodology. D-D.O: Contributing to the conceptualization, the data curation the methodology, the supervision, the validation and the writing of the original draft. F.K: Contributing to the conceptualization, the methodology, the validation and the writing of the original draft. C.S: Contributing to the validation and the writing of the original draft. A.R.Y: Contributing to the supervision, the validation and writing – review and editing. I.A.T: Contributing to the supervision, the validation and the writing of the original draft. W.M.N: Contributing to the visualization, the validation and the writing of the original draft. A.O: Contributing to the validation and the writing of the original draft. Y.E.Z: Contributing to the validation and the writing of the original draft.

Funding

No funding.

Data Availability

Data supporting the results of this study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was approved by the local Ethics Committee of the Bogodogo University Hospital center. (N202202-32). Informed consent was obtained from each included patient or their relatives in the study.

Consent for publication

Not applicable

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Rheumatology, Bogodogo University Hospital Center, Sector 51, 14 BP 371, Ouagadougou, Burkina Faso
[2]Department of Internal Medicine, Sourou Sanou University Hospital Center, Bobo-Dioulasso, Burkina Faso
[3]Department of Cardiology, Tengandogo University Hospital Center, Ouagadougou, Burkina Faso
[4]Department of Cardiology, Bogodogo University Hospital Center, Ouagadougou, Burkina Faso

Received: 13 February 2025 / Accepted: 12 May 2025
Published online: 16 May 2025

References

1. Silverman ED. Celebrating the journal of rheumatology's 50th year of publication. J Rheumatol. 2023;50(1):1–2.
2. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and bard against conventional Drug-Drug interactions clinical tools. Drug Healthc Patient Saf. 2023;15:137–47.
3. Chinnadurai S, Mahadevan S, Navaneethakrishnan B, Mamadapur M. Decoding applications of artificial intelligence in rheumatology. Cureus. 2023;15(9):e46164.
4. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y et al. Diagnostic performance comparison between generative AI and physicians: a systematic review and meta-analysis. medRxiv; 2024. https://www.medrxiv.org/content/https://doi.org/10.1101/2024.01.20.24301563v2. Accessed 10 february 2025.
5. Kumari A, Kumari A, Singh A, Singh SK, Juhi A, Dhanvijay AKD, et al. Large Language models in hematology case solving: A comparative study of ChatGPT-3.5, Google bard, and Microsoft Bing. Cureus. 2023;15(8):e43861.
6. Venerito V, Puttaswamy D, Iannone F, Gupta L. Large Language models and rheumatology: a comparative evaluation. Lancet Rheumatol. 2023;5(10):e574–8.
7. Is EE, Menekseoglu AK. Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google gemini and ChatGPT-4o. Clin Rheumatol. 2024;43(11):3507–13.
8. Boateng G, John S, Boateng S, Badu P, Agyeman-Budu P, Kumbol V. Real-World deployment and evaluation of Kwame for science, an AI teaching assistant for science education in West Africa. Artificial intelligence in education. Cham: Springer Nature Switzerland; 2024. pp. 119–33.
9. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799.
10. Kaya Kaçar H, Kaçar ÖF, Avery A. Diet quality and caloric accuracy in AI-Generated diet plans: A comparative study across chatbots. Nutrients. 2025;17(2):206.
11. CIM-10 FR à usage PMSI. 2025| Publication ATIH. https://www.atih.sante.fr/cim-10-fr-usage-pmsi-2025. Accessed 29 Jan 2025.
12. Carini C, Seyhan AA. Tribulations and future opportunities for artificial intelligence in precision medicine. J Transl Med. 2024;22(1):411.
13. Martín Andrés A, Álvarez Hernández M. Hubert's multi-rater kappa revisited. Br J Math Stat Psychol. 2020;73(1):1–22.
14. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988;240(4857):1285–93.
15. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191–4.
16. Venerito V, Bilgin E, Iannone F, Kiraz S. AI am a rheumatologist: a practical primer to large Language models for rheumatologists. Rheumatol Oxf Engl. 2023;62(10):3256–60.
17. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large Language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int. 2024;44(2):303–6.
18. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78–80.
19. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, Pérez-Sancristóbal I, Pato-Cour E, Plasencia-Rodríguez C, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep. 2023;13(1):22129.
20. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large Language models in orthodontics: a comparative study of ChatGPT, Google bard, and Microsoft Bing. Eur J Orthod. 2024;46:cjae017.
21. Nacanabo M, Seghda TAA, Bayala YLT, Millogo G, Thiam A, Yameogo N, et al. Evaluation of the diagnostic capabilities of artificial intelligence in the cardiology department of the Bogodogo University Hospital Center using Chat GPT; 2024.

Publisher's note