Comparative Study to Evaluate the Accuracy of Differential Diagnosis Lists Generated by Gemini Advanced, Gemini, and Bard for a Case Report Series Analysis: An Experimental Study

Takanobu Hirosawa, Yukinori Harada, Kazuki Tokumasu, Takahiro Ito, Tomoharu Suzuki, Taro Shimizu

Submitted to: JMIR Medical Informatics
on: June 07, 2024

Table of Contents

Comparative Study to Evaluate the Accuracy of Differential Diagnosis Lists Generated by Gemini Advanced, Gemini, and Bard for a Case Report Series Analysis: An Experimental Study

Takanobu Hirosawa[1] MD, PhD; Yukinori Harada[1] MD, PhD; Kazuki Tokumasu[2] MD, PhD; Takahiro Ito[3] MD; Tomoharu Suzuki[4] MD; Taro Shimizu[1] MD, PhD, MSc, MPH, MBA, FACP

[1]Department of Diagnostic and Generalist Medicine Dokkyo Medical University Shimotsuga JP
[2]Department of General Medicine Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences Okayama JP
[3]Satsuki home clinic Tochigi JP
[4]Department of Hospital Medicine Urasoe General Hospital Okinawa JP

Corresponding Author:
Takanobu Hirosawa MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
880 Kitakobayashi, Mibu-cho
Shimotsuga
JP

Abstract

Background: Generative artificial intelligence (GAI) systems by Google have recently been updated from Bard to Gemini and Gemini Advanced as of December 2023. Gemini is a basic, free-to-use model after a user's login, while Gemini Advanced operates on a more advanced model requiring a fee-based subscription. These systems have the potential to enhance medical diagnostics. However, the impact of these updates on comprehensive diagnostic accuracy remains unknown.

Objective: This study aims to compare the accuracy of the differential diagnosis lists generated by Gemini Advanced, Gemini, and Bard across comprehensive medical fields using case report series.

Methods: We identified case report series with relevant final diagnoses published from the American Journal Case Reports from January 2022 to March 2023. After excluding non-diagnostic cases and patients under 10 years old, we included the remaining case reports. After refining the case parts as case descriptions, we input the same case descriptions into Gemini Advanced, Gemini, and Bard to generate the top 10 differential diagnosis lists. Two expert physicians independently evaluated whether the final diagnosis was included in the lists and its ranking. Any discrepancies were resolved by another expert physician. The Bonferroni correction was applied to adjust the P values for the number of comparisons among three GAI systems, setting the corrected significance level at P value < .0167.

Results: In total, 392 case reports were included. The inclusion rates of the final diagnosis within the top 10 differential diagnosis lists were 73.0% (286/392) for Gemini Advanced, 76.5% (300/392) for Gemini, and 68.6% (269/392) for Bard. The top diagnoses matched the final diagnoses in 31.6% (124/392) for Gemini Advanced, 42.6% (167/392) for Gemini, and 31.4% (123/392) for Bard. Gemini demonstrated higher diagnostic accuracy than Bard both within the top 10 differential diagnosis lists (P =.0163) and as the top diagnosis (P =.001). Additionally, Gemini Advanced achieved lower accuracy in identifying the most probable diagnosis compared to Gemini, with this result being statistically significant (P =.002).

Conclusions: The results of this study suggest that Gemini outperformed Bard in diagnostic accuracy following the model update. However, Gemini Advanced requires further refinement to optimize its performance for future AI-enhanced diagnostics. These findings should be interpreted cautiously and considered primarily for research purposes, as these GAI systems have not been adjusted for medical diagnostics nor approved for clinical use. Clinical Trial: Not applicable

(JMIR Preprints 07/06/2024:63010)
DOI: https://doi.org/10.2196/preprints.63010

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Comparative Study to Evaluate the Accuracy of Differential Diagnosis Lists Generated by Gemini Advanced, Gemini, and Bard for a Case Report Series Analysis: An Experimental Study

Authors: Takanobu Hirosawa[1] MD, PhD, Yukinori Harada[1] MD, PhD, Kazuki Tokumasu[2] MD, PhD, Takahiro Ito[3] MD, Tomoharu Suzuki[4] MD, and Taro Shimizu[1] MD, PhD, MSc, MPH, MBA, FACP

Affiliations:
1 Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Tochigi, 321-0293, Japan.
2 Department of General Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan.
3 Satsuki home clinic, Tochigi, Japan
4 Department of Hospital Medicine, Urasoe General Hospital, Okinawa, Japan.

Correspondence: Takanobu Hirosawa, MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University, 880 Kitakobayashi, Mibu-cho, Shimotsuga, Tochigi, Japan 321-0293
Tel + 81-282-87-2498
Fax + 81-282-87-2502
Email: hirosawa@dokkyomed.ac.jp

Abstract

Background: Generative artificial intelligence (GAI) systems by Google have recently been updated from Bard to Gemini and Gemini Advanced as of December 2023. Gemini is a basic, free-to-use model after a user's login, while Gemini Advanced operates on a more advanced model requiring a fee-based subscription. These systems have the potential to enhance medical diagnostics. However, the impact of these updates on comprehensive diagnostic accuracy remains unknown.

Objective: This study aims to compare the accuracy of the differential diagnosis lists generated by Gemini Advanced, Gemini, and Bard across comprehensive medical fields using case report series.

Methods: We identified case report series with relevant final diagnoses published from the *American Journal Case Reports* from January 2022 to March 2023. After excluding non-diagnostic cases and patients under 10 years old, we included the remaining case reports. After refining the case parts as case descriptions, we input the same case descriptions into Gemini Advanced, Gemini, and Bard to generate the top 10 differential diagnosis lists. Two expert physicians independently evaluated whether the final diagnosis was included in the lists and its ranking. Any discrepancies were resolved by another expert physician. The Bonferroni correction was applied to adjust the *P* values for the number of comparisons among three GAI systems, setting the corrected significance level at *P* value < .0167.

Results: In total, 392 case reports were included. The inclusion rates of the final diagnosis within the top 10 differential diagnosis lists were 73.0% (286/392) for Gemini Advanced, 76.5% (300/392) for Gemini, and 68.6% (269/392) for Bard. The top diagnoses matched the final diagnoses in 31.6% (124/392) for Gemini Advanced, 42.6% (167/392) for Gemini, and 31.4% (123/392) for Bard. Gemini demonstrated higher diagnostic accuracy than Bard both within the top 10 differential diagnosis lists (*P* =.0163) and as the top diagnosis (*P* =.001). Additionally, Gemini Advanced achieved lower accuracy in identifying the most probable diagnosis compared to Gemini, with this result being statistically significant (*P* =.002).

Conclusions: The results of this study suggest that Gemini outperformed Bard in diagnostic accuracy following the model update. However, Gemini Advanced requires further refinement to optimize its performance for future AI-enhanced diagnostics. These findings should be interpreted cautiously and considered primarily for research purposes, as these GAI systems have not been adjusted for medical diagnostics nor approved for clinical use.

Trial Registration: Not applicable

Keywords: Artificial Intelligence; Clinical Decision Support; Diagnostic Excellence; Generative Artificial Intelligence; Large Language Models; Natural Language Processing

Introduction

Diagnostic team to reduce misdiagnoses

Diagnosis is a crucial step in clinical medicine, where a significant proportion of medical errors and harms are related to diagnostic errors [1]. The formation of a diagnostic team has been proposed as an effective strategy to mitigate the risks associated with misdiagnosis [2, 3]. This team should promote collaboration among medical professionals, patients, their families, and the integration of digital tools to enhance the diagnostic accuracy [4]. Several research, including systematic reviews, have shown that the implementation of clinical decision support systems (CDSSs) in clinical settings has significantly improved diagnostic accuracy, patient care and healthcare process [5-7].

Digital tool for medical diagnosis

Various digital tools, particularly diagnostic CDSSs, have emerged for medical diagnostics. These systems are designed to provide diagnostic suggestions based on clinical data, aiding medical professionals in clinical decision-making [8]. Traditionally, diagnostic CDSSs, such as symptom checkers and differential diagnosis generators, have relied on fixed algorithms and rule-based systems derived from medical databases and expert input [9-11]. Unfortunately, these systems often suffered from poor accuracy and inadequate integration into clinical workflows, limiting their practical utility in real-world medical settings [4]. In this context, artificial intelligence (AI), especially generative artificial intelligence (GAI), has introduced a new category of CDSS [12]. This advancement suggests a future shift in how digital tools can support diagnostic processes.

Generative artificial intelligence in medical diagnosis

GAI systems have shown rapid development and are increasingly influencing various fields, including medicine. This advancement is partly due to the development of machine learning techniques, such as neural networks and natural language processing. GAI represents a shift from rule-based systems to models that can autonomously generate and evaluate new data patterns. Overcoming many limitations faced by traditional CDSSs, GAI systems could significantly enhance future diagnostic processes. Notable examples include ChatGPT developed by Open AI, and Gemini and Gemini Advanced from Google [13]. These systems utilize the advanced large language models (LLMs), which are complex neural networks trained on vast data sets through natural language processing [14]. Recent studies, including one evaluating dermoscopy image descriptions with chatbot responses, have demonstrated promising results in accuracy and diagnostic completeness by ChatGPT [15]. The language model for dialogue applications (LaMDA) developed by Google AI is one such LLM. Their ability to process and generate outputs is particularly promising for future applications in medical diagnostics, where they will analyze complex clinical information and collaborate as part of a diagnostic team.

From Bard to Gemini and Gemini Advanced

Originally, Bard was developed using the LaMDA model primarily for text generation and conversational AI, and later transitioned to the Pathways Language Model (Palm 2). Subsequent developments led to the release of Gemini and Gemini Advanced in December 2023. Gemini Advanced, an upgraded version of Gemini, leverages Ultra 1.0, Google's most advanced model, offered as a fee-based service [16, 17]. These developments reflect the rapid pace at which GAI technology is advancing. Recent updates have transformed Bard into Gemini and Gemini Advanced, enhancing their functionalities and applications in various fields. Previous research, including our

own, has demonstrated that Bard showed promising results in medicine [18-21]. Moreover, a recent study has shown that several GAI systems, including Gemini Advanced, could achieve notable diagnostic accuracy for multiple-choice questions about clinical vignettes [22]. These findings suggest that even without specific training or reinforcement for diagnostics, GAI systems show potential for reliable use in diagnostics. Despite these advancements, the comparative diagnostic accuracy of differential diagnosis lists by these GAI systems across comprehensive medical fields remains to be fully explored. This study aims to fill that gap by evaluating the diagnostic accuracy of the differential diagnosis lists generated by Gemini Advanced, Gemini, and Bard for case report series across various medical disciplines.

Methods

Overview

An experimental study was conducted to assess the diagnostic accuracy of Gemini Advanced, Gemini, and Bard for comprehensive case report series. This study was conducted at the Department of Diagnostic and Generalist Medicine (General Internal Medicine) in Dokkyo Medical University, Japan. This study consisted of preparing case materials, generating differential diagnosis lists, evaluating the lists, and analyzing the diagnostic accuracy. Figure 1 shows the study flow, including the inclusion of case materials and the generation of differential diagnosis lists.

Ethical approval

Given the study's method of utilizing published case reports, approval from an ethical committee was deemed not applicable.

Preparing case materials

We focused on a comprehensive series of case reports from the *American Journal of Case Reports*, covering a broad range of medical fields. The structured format of the journal facilitated easy identification of sections containing the case reports and the final diagnoses. Initially, the inclusion criteria were the case reports published in the *American Journal of Case Reports* from January 1, 2022, to March 1, 2023. A PubMed search identified 557 consecutive case reports. After excluding 130 non-diagnostic case reports and 35 pediatric case reports (patients aged under 10 years), 392 case reports remained. The exclusion criteria were based on prior research for CDSS [23]. We refined the case reports to prepare the case materials, which typically included the initial case report part to the definitive tests for final diagnosis. The relevant final diagnoses were typically described by the authors. We used only textual data exclusively, omitting image data. Specifically, the title, background, final diagnosis, clinical course following diagnosis, discussion, conclusion, figures, tables, and supplemental materials were excluded from the case materials. The main investigator (TH) conducted this process with validation from another investigator (YH). The PubMed search keywords are shown in Multimedia Appendix 1. For example, in a case report titled "Herpes Zoster Following COVID-19 Vaccine Booster," the final diagnosis was herpes zoster [24]. We extracted the case report part from "An 82-year-old.." to "Vesicular breath sounds were heard equally on both lung fields."

Generating differential diagnosis lists

We employed Gemini Advanced, Gemini, and Bard as GAI systems for this research. This was because these systems are popular AI platforms available to the public. These GAI systems were not specifically enhanced for medical diagnosis. Details about the GAI systems utilized in this study are provided in Table 1. To generate the top 10 differential diagnosis lists from GAI systems, the main investigator typically copied and pasted the case materials into the AI systems with the following prompt: "Tell me the top 10 suspected illnesses for the following case: (case materials)." This prompt, developed through preliminary research, aimed to generate the top 10 differential diagnosis lists. The first list produced by the GAI systems was utilized as the differential diagnosis list. The data control setting was adjusted to "Not saving activity," to avoid the influence from the previous conversations. Additionally, before starting a new session, the main investigator refreshed the previous session to prevent any influence from prior conversations.

Table 1. The details of generative artificial intelligence (AI) systems utilized in this study.

Feature	Gemini Advanced	Gemini	Bard
AI model			
	Ultra 1.0	Pro	Pathways Language Model (Palm 2)-based
Availability			
	Fee-based subscription	Free with user login	Discontinued
The setting of apps activity			
	Not saving activity	Not saving activity	Not saving activity
Access date			
	2024/4/3-2024/4/9	2024/3/12-2024/3/28	2023/7/1-2023/8/8
Prompt			
	"Tell me the top 10 suspected illnesses for the following case: (case materials)."	"Tell me the top 10 suspected illnesses for the following case: (case materials)."	"Tell me the top 10 suspected illnesses for the following case: (case materials)."

Evaluating the differential diagnosis lists

Two expert researchers (TI and TS: Tomoharu Suzuki) independently evaluated the differential diagnosis lists from GAI systems. A score "1" was assigned if the differential accurately and specifically identified the final diagnosis or was sufficiently close to the final diagnosis. Conversely, a score "0" was assigned if it diverged significantly from the final diagnosis [25]. When a GAI system could not output the differential diagnosis list, a score "0" was labeled. When the score was "1," the evaluator assessed its ranking within the list. Any discrepancies were resolved by another expert researcher (KT). All evaluators were blinded to which GAI systems produced the differential diagnosis lists.

Analyzing the diagnostic accuracy

In this study, we defined the diagnostic accuracy as the inclusion of the final diagnoses in the differential diagnosis lists.

Outcome

In terms of the outcomes, the primary outcome was the total score for correctly identifying the final diagnosis in the top 10 differential diagnosis lists generated by Gemini Advanced, Gemini, and Bard. The total number of included case reports was used as the denominator. The numerator was the number of case reports that correctly identified the final diagnosis in the top 10 differential diagnosis lists. The secondary outcomes were the total score for correctly identifying the final diagnosis in the top 5 differential diagnosis lists and as the top diagnosis generated from Gemini Advanced, Gemini, and Bard.

Additionally, we evaluated the top 10 rankings of the most frequently named differential diagnoses across generated differential diagnosis lists by a GAI system to find the underlying patterns. We also assessed whether the items in the differential diagnosis lists corresponded to the name of existing diseases.

Moreover, we analyzed how Gemini Advanced, Gemini, and Bard rank the correct diagnosis on average when it appears in the differential diagnosis lists. This metric helps evaluate not only whether the correct diagnosis is included but also its relative priority among other suggested diagnoses. For cases where the correct diagnosis was missing, we assigned a penalty rank; specifically, we used 11 as the penalty rank.

Statistical Analysis

A chi–squared test was used for the categorical or binary variables. The Mann-Whitney U test was applied to analyze the average rankings. For multiple comparisons, the Bonferroni correction was applied [26]. The Bonferroni correction adjusts the *P* value by dividing it by the number of comparisons being made, namely three GAI systems: Gemini Advanced, Gemini, and Bard, setting a more stringent threshold for individual tests to be considered significant. The Bonferroni-corrected significance level was defined as a *P* value of < .0167. Note that all statistical tests were performed using the R (version 4.2.2) for macOS X (The R foundation for Statistical Computing, Vienna, Austria).

Results

Overview of Outputs from Generative Artificial Intelligence Systems

In total, 392 case reports were included. The distribution of diagnoses in these reports highlighted the prevalence of complex medical conditions, with 22.7% (89/392) of cases being diagnosed as malignancy, 22.4% (88/392) as various infections, and 17.1% (67/392) as vascular diseases.

Bard generated differential diagnosis lists for all 392 cases. In contrast, both Gemini Advanced and Gemini encountered challenges with 3 cases, failing to produce outputs. The cases where outputs

were not generated involved complex clinical scenarios that were directly described in the reports as involving overdose, psychological conditions, and vaccination side effects. Representative outputs from Gemini Advanced, Gemini, and Bard are shown in Table 2. The dataset of differential diagnosis generated by AI systems in this study, alongside the final diagnosis, are shown in Multimedia Appendix 2.

Table 2. Representative outputs from Gemini Advanced, Gemini, and Bard.

Final Diagnosis	Gemini Advanced	Gemini	Bard
Herpes zoster			
	1. Shingles (Herpes Zoster) 2. Myocardial ischemia 3. Costochondritis 4. Pericarditis 5. Pleuritis 6. Pulmonary embolism 7. Pneumonia 8. Metastatic malignancy 9. Aortic dissection 10. Esophageal spasm	1. Shingles (Herpes Zoster) 2. Myocardial ischemia (Angina) 3. Costochondritis 4. Musculoskeletal strain 5. Pleurisy 6. Intercostal neuritis 7. Pulmonary embolism 8. Pericarditis 9. Gastritis 10. Esophagitis	1. Herpes zoster (shingles) 2. Myocardial infarction (heart attack) 3. Aortic dissection 4. Pulmonary embolism 5. Pneumonia 6. Thoracic radiculopathy 7. Costochondritis 8. Intercostal neuralgia 9. Costoclavicular syndrome 10. Tietze syndrome
Diffuse alveolar hemorrhage			
	1. Pulmonary alveolar hemorrhage 2. Nocardia pneumonia 3. Pulmonary embolism 4. Acute eosinophilic pneumonia 5. Drug-induced pneumonitis 6. Legionnaires' disease 7. Hypersensitivity pneumonitis	1. Pulmonary Alveolar Hemorrhage 2. Trastuzumab-induced Pneumonitis 3. Aspirin-Exacerbated Respiratory Disease (AERD) 4. Acute Heart Failure 5. Pulmonary Edema 6. Nocardia Pneumonia	1. Acute radiation pneumonitis 2. Acute hypersensitivity pneumonitis 3. Pulmonary edema 4. Pulmonary embolism 5. Aspiration pneumonia 6. Bronchogenic carcinoma 7. Sarcoidosis 8. Lymphoma 9. Wegener's

	8. Idiopathic pulmonary fibrosis 9. Bronchiolitis obliterans organizing pneumonia 10. Acute interstitial pneumonitis	7. Pulmonary Hypertension 8. Acute Interstitial Lung Disease 9. Acute on Chronic Kidney Disease 10. Occult Gastrointestinal Bleed	granulomatosis 10. Goodpasture syndrome
Glomus tumor			
	1. Meralgia paresthetica 2. Soft tissue sarcoma 3. Fibroma 4. Myositis ossificans 5. Neuroma 6. Hemangioma 7. Liposarcoma 8. Lymphoma 9. Abscess 10. Malignant peripheral nerve sheath tumor (MPNST)	1. Meralgia paresthetica 2. Lumbar radiculopathy 3. Femoral neuropathy 4. Piriformis syndrome 5. Muscle strain 6. Myofascial pain syndrome 7. Glomus tumor 8. Neuroma 9. Soft tissue sarcoma (unlikely) 10. Abscess (unlikely)	1. Lipoma 2. Fibromatosis 3. Myositis 4. Neoplasm 5. Infection 6. Inflammation 7. Trauma 8. Hematoma 9. Varicose vein 10. Cellulitis

Diagnostic accuracy

In terms of diagnostic accuracy, the inclusion rates of the final diagnoses within the top 10 differential diagnosis lists were 73.0% (286/392) for Gemini Advanced, 76.5% (300/392) for Gemini, and 68.6% (269/392) for Bard. For the top 5 differential diagnosis, the rates were 60.5% (237/392) for Gemini Advanced, 66.3% (260/392) for Gemini, and 59.9% (235/392) for Bard. The top diagnoses matched the final diagnoses in 31.6% (124/392) for Gemini Advanced, 42.6% (167/392) for Gemini, and 31.4% (123/392) for Bard. Gemini demonstrated higher diagnostic accuracy than Bard both within the top 10 differential diagnosis lists (*P* =.0163) and as the top diagnosis (*P* =.001). Additionally, Gemini Advanced achieved lower accuracy in identifying the most probable diagnosis, compared to Gemini with this result being statistically significant (*P* =.002). Other comparisons were statistically insignificant. Table 3 and Figure 2 show the diagnostic accuracy by Gemini Advanced, Gemini, and Bard.

Table 3. Diagnostic accuracy of Gemini Advanced, Gemini, and Bard.

Variable	Gemini	Gemini	Bard	*P* value[a]

	Advanced					
				Gemini Advanced vs. Gemini	Gemini Advanced vs. Bard	Gemini vs. Bard
Within the top 10, n (%)						
	286/392 (73.0)	300/392 (76.5)	269/392 (68.6)	.29	.21	.0163
Within the top 5, n (%)						
	237/392 (60.5)	260/392 (66.3)	235/392 (59.9)	.10	.94	.08
Top diagnosis, n (%)						
	124/392 (31.6)	167/392 (42.6)	123/392 (31.4)	.002	.99	.001

[a]Chi-squared test. The Bonferroni-corrected significance level at a *P* value < .0167

Most frequently named differential diagnoses

Regarding the top 10 most frequently named differential diagnoses, all rankings included sepsis, pneumonia, pulmonary embolism, lymphoma, and meningitis. Notably, the top three most frequently named differential diagnoses by Gemini Advanced and Gemini were the same. Table 4 shows the top 10 most frequently named differential diagnoses generated by Gemini Advanced, Gemini, and Bard.

Table 4. The top 10 most frequently named differential diagnoses generated by Gemini Advanced, Gemini, and Bard.

The raking in the top 10 most frequently named differentials, (N)	Gemini Advanced	Gemini	Bard
1			
	Sepsis (43)	Sepsis (42)	Sarcoidosis (51)
2			
	Pneumonia (34)	Pneumonia (28)	Sepsis (42)
3			
	Pulmonary embolism (33)	Pulmonary embolism (20)	Pneumonia (41)
4			

	Acute kidney injury (28)	Sarcoidosis (15)	Lymphoma (40)
5			
	Lymphoma (25)	Pericarditis (14)	Pulmonary embolism (39)
6			
	Urinary tract infection (24)	Meningitis (14)	Meningitis (31)
7			
	Heart failure (23)	Lymphoma (13)	Inflammatory bowel disease (29)
8			
	Meningitis (22)	Myocarditis (13)	Tuberculosis (26)
9			
	Myocardial infarction (20)	Acute kidney injury (12)	Encephalitis (25)
10			
	Pericarditis (18)	Systemic Lupus Erythematosus (12)	Myocarditis (24)

Inappropriate diseases names

From all differential diagnosis lists output by generative AIs, we identified inappropriate disease names: 11 items from Gemini Advanced, 9 items from Gemini, and 5 items from Bard. Notably, Gemini Advanced and Gemini both erroneously listed "Wegner's granulomatosis," a misspelling of the previous correct term, "Wegener's granulomatosis," which has now been updated to "Granulomatosis with Polyangiitis" (GPA) [27]. Another error by Gemini Advanced involved "Microcytic colitis," likely a confusion between "microcystic anemia" and "microscopic colitis." Table 5 lists the inappropriate disease names generated by Gemini Advanced, Gemini, and Bard.

Table 5. Inappropriate disease name generated by Gemini Advanced, Gemini, and Bard.

Correct disease name, [cell number in Multimedia appendix 2]	Inappropriate disease name by Gemini Advanced	Inappropriate disease name by Gemini	Inappropriate disease name by Bard
Drug reaction (Dasatinib)			
	Drug reation (Dasatinib) [O15]		
Nonketotic hyperglycemic hyperosmolar coma			

(NKHHC)			
		Nonketotic hyperglycemia hyperosmolar coma (NKHHC) [X29]	
Lipedema			
		Lipoderma [U34]	
Small bowel angiodysplasia			
	Small bowel angiodisplasia [M40]		
Granulomatosis with polyangiitis (GPA)			
	Granulomatous with polyangiitis (GPA) (L68), Wegner's granulomatosis [N111]	Wegner's granulomatosis [U186]	
Costochondritis			
	Costochondritis a [P86]		
Maxillary Sinus Carcinoma			
		Maxillary Sinus Cycinoma [L106]	
Constrictive pericarditis			
	Conrictive pericarditis [L110]		
Scleroderma-related Interstitial Lung Disease (SSc-ILD)			
		Scleroderma-related Interstitial Lung Disease (SLILD) [S117]	
Pericoronitis			
			Pericoronatitis [AA133]
Osteitis			
			Osteoitis [AC133]
Microscopic colitis			

	Microcytic colitis [L152]		
Pneumocystis jirovecii			
			Pneumocystis jerovecii [AG156]
Leukoencephalopathy			
		Leukoencephalomyopathy [X195]	
Strumal carcinoid			
		Struma carcinoid [W208]	
Restrictive ventilatory impairment			
	Restricted ventilatory impairment [J360]		
Moebius syndrome			
		Mobius syndrome [Z369]	
Endometriosis			
			Endometrios [AE385]
Cryptococcus neoformans			
			Chryptococcus neoformans [AJ389]
Unknown			
	Ytzinger hernia [M197]	Y-type appendicitis [V197]	
Unknown			
	(There was partly Arabic language) [L292]		
Unknown (Transaminase elevation is also not disease name)			
	Transaminitis elevation (N354)		

Average ranking

In terms of average ranking, the scores were 5.25 for Gemini Advanced, 4.54 for Gemini, and 5.33 for Bard. The differences in average rankings were not statistically significant between Gemini Advanced and Gemini (P=.99), between Gemini Advanced and Bard (P=.17), and between Gemini and Bard (P=.99).

Discussion

Principal Results

In the following, we discuss our principal findings. Our findings indicate that Gemini demonstrated superior diagnostic accuracy compared to Bard, not only within the top 10 differential diagnosis lists but also in identifying the most likely diagnosis. Specifically, Gemini's diagnostic accuracy for the top 10 lists was 76.5% (300/392), compared to Bard's 68.6% (269/392), with a statistically significant difference (*P* =.0163). Moreover, as the top diagnosis, Gemini's diagnostic accuracy was 42.6% (167/392) versus Bard's 31.4% (123/392), also significant (*P* =.001). This enhancement in Gemini's diagnostic performance may be attributed to its advanced algorithmic framework, which likely incorporates more nuanced medical data and learns from recent case inputs, leading to more refined diagnostic predictions.

However, the performance of Gemini did not statistically outperform in the top 5 differential diagnosis lists. This outcome may suggest that while Gemini's algorithm is effective in a broader exploratory context, its precision may falter when constrained to a narrower list of top diagnoses. This indicates that a balance between breadth of exploration and depth of focus is crucial for optimizing diagnostic accuracy in such AI systems.

Conversely, our analysis showed that Gemini Advanced did not perform as well as expected when compared to Gemini. Despite expectations that the advanced model would provide enhanced diagnostic capabilities, it achieved lower accuracy in identifying the most probable diagnosis with 31.6% (124/392) compared to Gemini's 42.6% (167/392), with this result being statistically significant (*P* =.002). This outcome suggests that the additional features or complexity added in Gemini Advanced may not necessarily translate into improved diagnostic performance. These findings underscore the need for further refinement and optimization of Gemini Advanced to harness its potential for future AI-enhanced diagnostics.

Additionally, our analysis identified issues with inappropriate disease naming in the outputs from generative AI systems, with Gemini Advanced and Gemini producing outdated or misspelled terms for vasculitis, instead of using the updated name. These inaccuracies highlight the challenges in ensuring up-to-date and precise medical terminology in AI outputs, which is crucial for maintaining trust and reliability in AI-assisted diagnostics. Furthermore, these misspellings are often found in published medical articles, suggesting that generative AIs may have learned these errors from these sources. The fact that both Gemini Advanced and Gemini exhibited the same mistakes indicates potential similarities in their underlying models or training data.

Regarding average rankings, there were no statistically significant differences among generative AI

systems. This indicates a level of parity in how each model ranks diagnoses when they include the correct diagnosis, suggesting that while there are differences in overall accuracy, the ranking mechanisms of each model are relatively similar.

Given the current performance metrics, our analysis supports prioritizing the adjustment and enhancement of Gemini for future applications in medical diagnostics, rather than Gemini Advanced. Despite the theoretically superior capabilities of Gemini Advanced [17], Gemini's framework appears more aligned with practical diagnostic needs and shows greater promise in real-world applications. However, it is essential to verify this trend across a variety of sources to ensure that these findings are not specific to the data sets used in this study. Further investigations involving diverse clinical environments and different types of medical data are crucial to confirm the consistency and reliability of Gemini's superior performance.

Finally, the comparative analysis of the differentials by Gemini Advanced, Gemini, and Bard revealed consistent inclusion of sepsis, pneumonia, pulmonary embolism, lymphoma, and meningitis among their top 10 differentials. This underscores not only a shared prioritization of these conditions but also the systems' effectiveness in recognizing critical and prevalent diseases. The consistent identification of sepsis, particularly its second-place ranking by Bard, underscores the potential of these AI systems to enhance diagnostic accuracy and reduce errors in the identification of life-threatening conditions [28]. Importantly, the top 3 differentials by Gemini Advanced and Gemini – sepsis, pneumonia, and pulmonary embolism – are among the most harmful diseases where reducing diagnostic errors is crucial [1]. This suggests a potential for GAI systems to alert medical professionals about the inclusion of these important diseases during diagnosis. Such an understanding could facilitate more effective employment of these GAI systems in future diagnostics processes.

Strengths

This study had several strengths. First, the strengths of this study lie in its direct comparison of three cutting-edge AI systems and its demonstration of the dynamic improvements in their diagnostic accuracy. Unlike some CDSSs like symptom checkers, whose performance has plateaued [29], these GAI systems evaluated in this research show considerable enhancements with each iteration. Second, we evaluated the diagnostic accuracy of GAI systems utilizing a series of the case reports. These case reports often describe rare diseases and atypical presentations, as opposed to common diseases and typical presentations [30]. This showcases the system's diagnostic capabilities under challenging conditions. Third, the comprehensive range of medical conditions covered by the differential diagnosis lists generated by the AI systems represents a significant strength of this study. This extensive coverage demonstrates the systems' capacity to handle a broad spectrum of medical knowledge and its applicability to various clinical scenarios.

Limitations

Several limitations should be discussed. First, the use of case report series might not fully reflect real-world clinical scenarios. This limitation arises because case reports typically focus on novel or rare aspects of diseases rather than typical presentations and common diseases [31]. Second, the exclusive use of a single case reports journal could introduce selection bias. Third, there was no well-established method for evaluating AI diagnostics. In our study, we employed binary evaluation

methods. In contrast, other research on CDSSs employed several rating methods [32, 33] and the ranking averages in the differential diagnosis lists [34]. Fourth, we used only text-data; excluding image data could influence the diagnostic performance. These factors limit the generalizability of the current findings.

Concerning the GAI systems, all platforms utilized in this study were not designed for clinical use and have not received approval for medical diagnostics. These systems were not specifically reinforced or enhanced for medical diagnostic purposes. According to a preprint, Med-Gemini, a specialized model in medicine, was developed [35] but is not available to the public. Additionally, we could not include all currently available GAI systems; thus, these findings cannot be generalized to other systems or different clinical scenarios. There was also a risk that these GAI systems may have learned from the published case reports utilized in this study.

Moreover, the utilization of user data to refine models, as seen in Gemini Advanced and Gemini, highlights significant privacy concerns [36]. Future research should address the development of locally deployable LLM solutions tailored specifically for CDSS. Although our dataset is sourced from an open journal, careful consideration must be given to ethical deployment of these models within healthcare settings. Finally, given the rapid pace of GAI technology development, such as the evolution from Bard to Gemini and from ChatGPT-3 to ChatGPT-4 and ChatGPT-4o, our findings may have a limited shelf-life.

Future Direction

Future research will aim to explore the diagnostic accuracy of GAI systems following medical enhancements and adjustments. Once approved for medical use, it will also be essential to investigate the performance of GAI systems across various populations and settings, including remote medical consultations, to ensure their effectiveness in real-world diagnostics. Moreover, assessing the impact of AI-enhanced diagnostics on the decision-making process of medical professionals will be crucial.

Additionally, future studies should focus on integrating GAI systems with existing electronic health record systems to understand how AI can augment data accessibility and analysis. This integration will be essential to evaluate how GAI can improve clinical workflows, reduce the cognitive burden and the time to diagnosis, and enhance patient outcomes.

Lastly, the development of ethical guidelines and governance frameworks for the use of GAI in diagnostics is imperative [37]. As AI technologies become more prevalent in healthcare, it is crucial to establish clear protocols that safeguard patient privacy, ensure data security, and maintain transparency in AI decision-making processes.

Comparison with Prior Work

Our research builds on previous findings. We revealed that the diagnostic accuracy of ChatGPT-4 was 86.7% (340/392) for the final diagnoses included in the top 10 differential diagnosis lists, and 54.6% (214/392) for the top diagnosis [38]. ChatGPT-4's performances were still higher than that of Gemini in the lists (76.5% versus 86.7%) and as a top diagnosis (42.6% versus 54.6%); it was similar to Gemini Advanced in the lists (73.0% versus 86.7%) and as a top diagnosis (31.6% versus 54.6%).

Expanding our findings, another study showed that Isabel Pro, a successful CDSS developed by Isabel Healthcare, Ltd. [39], correctly identified diagnoses in 87.1% (175/201) of cases, compared to 82.1% (165/201) for ChatGPT-4 in a series of clinical cases [34]. These findings are partly attributed to the earlier launch of Isabel Pro and the ChatGPT series, allowing them to receive more user feedback and undergo updates to improve performance.

Additionally, another research focused on multiple choice questions on clinical vignettes revealed that ChatGPT-4 achieved a high accuracy rate of 73.3% for Clinical Challenges from the *Journal of the American Medical Association* (*JAMA*) and 88.7% for Image Challenges from the *New England Journal of Medicine* (*NEJM*). In contrast, Gemini, referred to as Gemini Pro in that study, achieved 63.6% for Clinical Challenges from *JAMA* and 68.7% for Image Challenges from the *NEJM* [22]. While these previous findings and current results revealed certain diagnostic performances of generative AI systems, comparing these results poses significant challenges due to methodological differences. Variations stem from differences in dataset preparation, the types of clinical vignettes used, and the specific challenges or images included, which may influence performance outcomes. Additionally, the evaluation criteria used to assess accuracy might differ significantly, affecting the comparability. For instance, the scoring systems or the definitions of a 'correct' answer could vary, necessitating caution when drawing direct comparisons between these findings and those of the present study.

In contrast to the serial evaluation approach of a symptom checker [29], which demonstrated an accuracy of 44.3% (97/219) in first year and 47.7% (43/90) in third year without significant difference, the performance of generative AI systems presents a different dynamic. Specifically, the serial evaluation of generative AI indicated that Gemini outperformed Bard over a relatively short period. This superiority can be attributed in part to adaptability of generative AI systems to incorporate additional data. However, it is crucial to note that this adaptability does not consistently translate into improved diagnostic accuracy, as evidenced by the current comparison between Gemini Advanced and Bard. This observation highlights the nuanced interplay between technological advancement and clinical efficacy, underscoring the need for continued research and validation in integrating these systems into medical practice effectively.

Conclusions

The results of this study suggest that Gemini outperformed Bard in diagnostic accuracy following the model update. However, Gemini Advanced requires further refinement to optimize its performance for future AI-enhanced diagnostics. These findings should be interpreted cautiously and considered primarily for research purposes, as these GAI systems have not been adjusted for medical diagnostics nor approved for clinical use. The potential and limitations highlighted by this study underscore the need for ongoing development and evaluation of GAI systems within medical diagnostics.

Acknowledgements

Contributors: TH, YH, KT, TI, TS (Tomoharu Suzuki), and TS (Taro Shimizu) contributed to the study concept and design. TH performed the statistical analyses. TH contributed to the drafting of the manuscript. YH, KT, TI, TS (Tomoharu Suzuki), and TS (Taro Shimizu) contributed to the critical revision of the manuscript for relevant intellectual content. All the authors have read and approved

the final version of the manuscript.
This research was funded by JSPS KAKENHI (grant number 22K10421). This study was conducted using resources from the Department of Diagnostics and Generalist Medicine at Dokkyo Medical University.

Conflicts of Interest

None declared.

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
GAI: generative artificial intelligence
JAMA: *Journal of the American Medical Association*
LaMDA: language model for dialogue applications
LLM: large language model
NEJM: *New England Journal of Medicine*
Palm2: Pathways Language Model

Multimedia Appendix 1

The PubMed search keywords.

Multimedia Appendix 2

The dataset of differential diagnosis generated by AI systems in this study, alongside the final diagnosis.

References

1. Newman-Toker DE, Nassery N, Schaffer AC, Yu-Moe CW, Clemens GD, Wang Z, et al. Burden of serious harms from diagnostic error in the USA. BMJ Qual Saf. 2024;33(2):109-20.
2. Singh H, Connor DM, Dhaliwal G. Five strategies for clinicians to advance diagnostic excellence. Bmj. 2022;376:e068044.
3. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. Washington (DC): National Academies Press; 2015.
4. 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.
5. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj. 2005;330(7494):765.
6. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med. 2012;157(1):29-43.
7. Sibbald M, Monteiro S, Sherbino J, LoGiudice A, Friedman C, Norman G. Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel. BMJ Quality & Safety. 2022;31(6):426-33.

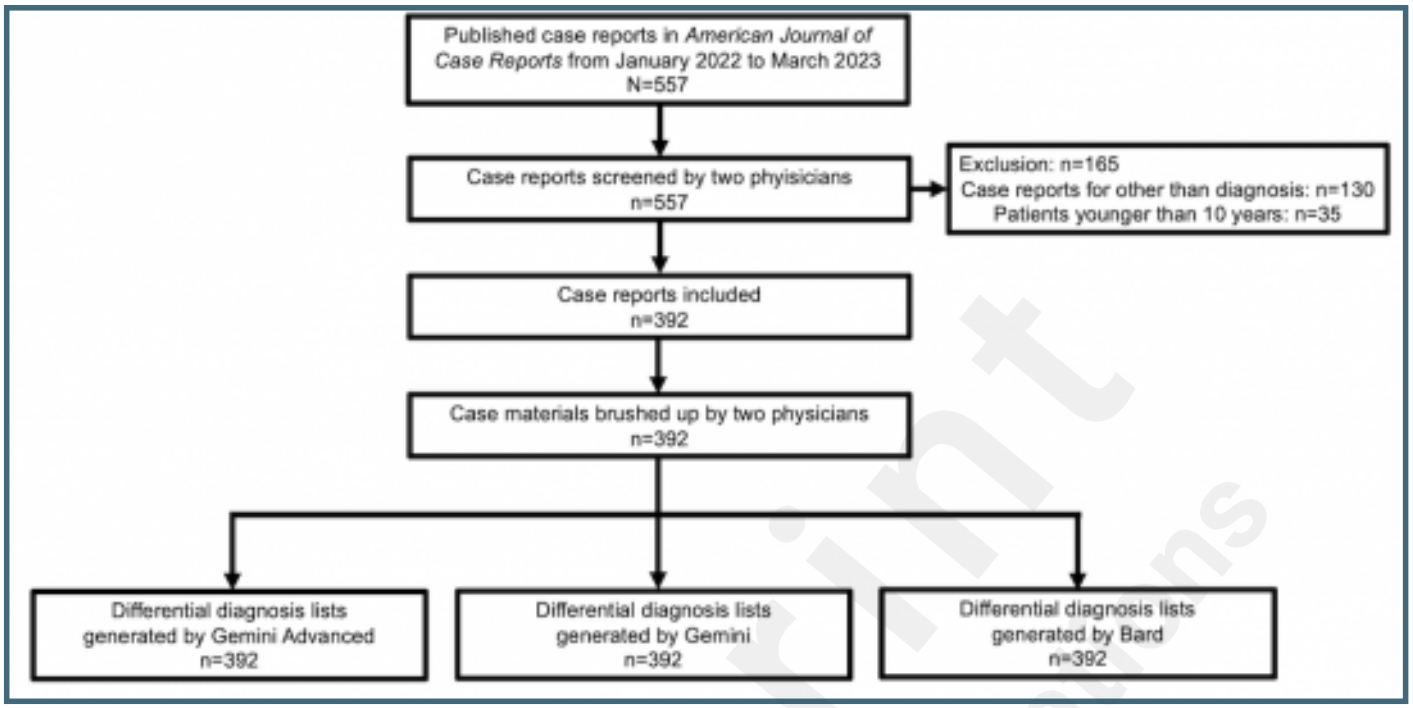
8. van Baalen S, Boon M, Verhoef P. From clinical decision support to clinical reasoning support systems. J Eval Clin Pract. 2021;27(3):520-8.
9. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The Effectiveness of Electronic Differential Diagnoses (DDX) Generators: A Systematic Review and Meta-Analysis. PLoS One. 2016;11(3):e0148991.
10. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage Accuracy of Symptom Checker Apps: 5-Year Follow-up Evaluation. J Med Internet Res. 2022;24(5):e31810.
11. Castaneda C, Nalley K, Mannion C, Bhattacharyya P, Blake P, Pecora A, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. Journal of Clinical Bioinformatics. 2015;5(1):4.
12. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review. JMIR Med Inform. 2024;12:e52073.
13. Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJPC. Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies, and Limitations. IEEE Access. 2024;12:31078-106.
14. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv preprint arXiv:230318223. 2023.
15. Karampinis E, Toli O, Georgopoulou K-E, Kampra E, Spyridonidou C, Roussaki Schulze A-V, et al. Can Artificial Intelligence “Hold” a Dermoscope?—The Evaluation of an Artificial Intelligence Chatbot to Translate the Dermoscopic Language. Diagnostics. 2024;14(11):1165.
16. Sundar Pichai DH. Introducing Gemini: our largest and most capable AI model. Google; 2023.
17. Team G, Anil R, Borgeaud S, Wu Y, Alayrac J-B, Yu J, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.
18. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. medRxiv. 2023:2023.04. 06.23288265.
19. Doshi RH, Amin K, Khosla P, Bajaj S, Chheang S, Forman H. Utilizing Large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, Google Bard, and Microsoft Bing. medRxiv. 2023:2023.06. 04.23290786.
20. Amin KS, Mayes LC, Khosla P, Doshi RH. Assessing the Efficacy of Large Language Models in Health Literacy: A Comprehensive Cross-Sectional Study. Yale J Biol Med. 2024;97(1):17-27.
21. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative Evaluation of Diagnostic Accuracy Between Google Bard and Physicians. Am J Med. 2023;136(11):1119-23 e18.
22. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative Analysis of Multimodal Large Language Model Performance on Clinical Vignette Questions. JAMA. 2024;331(15):1320-1.
23. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med. 2008;23 Suppl 1(Suppl 1):37-40.
24. Shahrudin MS, Mohamed-Yassin MS, Nik Mohd Nasir NM. Herpes Zoster Following COVID-19 Vaccine Booster. Am J Case Rep. 2023;24:e938667.
25. Krupat E, Wormwood J, Schwartzstein RM, Richards JB. Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors. Medical education. 2017;51(11):1127-37.
26. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. New York: john wiley & sons; 2003.
27. Falk RJ, Gross WL, Guillevin L, Hoffman G, Jayne DR, Jennette JC, et al. Granulomatosis with polyangiitis (Wegener's): an alternative name for Wegener's granulomatosis. Ann Rheum Dis.

2011;70(4):704.
28. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, et al. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. Jama. 2017;318(13):1241-9.
29. Harada Y, Sakamoto T, Sugimoto S, Shimizu T. Longitudinal Changes in Diagnostic Accuracy of a Differential Diagnosis List Developed by an AI-Based Symptom Checker: Retrospective Observational Study. JMIR Form Res. 2024;8:e53985.
30. Riley DS, Barber MS, Kienle GS, Aronson JK, von Schoen-Angerer T, Tugwell P, et al. CARE guidelines for case reports: explanation and elaboration document. Journal of Clinical Epidemiology. 2017;89:218-35.
31. Riley DS, Barber MS, Kienle GS, Aronson JK, von Schoen-Angerer T, Tugwell P, et al. CARE guidelines for case reports: explanation and elaboration document. Journal of Clinical Epidemiology. 2017;89:218-35.
32. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330(1):78-80.
33. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med. 2012;27(2):213-9.
34. Bridges JM. Computerized diagnostic decision support systems – a comparative performance study of Isabel Pro vs. ChatGPT4. Diagnosis. 2024.
35. Saab K, Tu T, Weng W-H, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of Gemini Models in Medicine. arXiv preprint arXiv:240418416. 2024.
36. Gemini Apps Privacy Notice [updated May 29, 2024. Available from: https://support.google.com/gemini/answer/13594961#privacy_notice.
37. Newman-Toker DE, Sharfstein JM. The Role for Policy in AI-Assisted Medical Diagnosis. JAMA Health Forum. 2024;5(4):e241339.
38. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Diagnostic performance of generative artificial intelligences for a series of complex case reports. DIGITAL HEALTH. 2024;10:20552076241265215.
39. Ren LY. Isabel Pro. Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada. 2019;40(2):63-9.

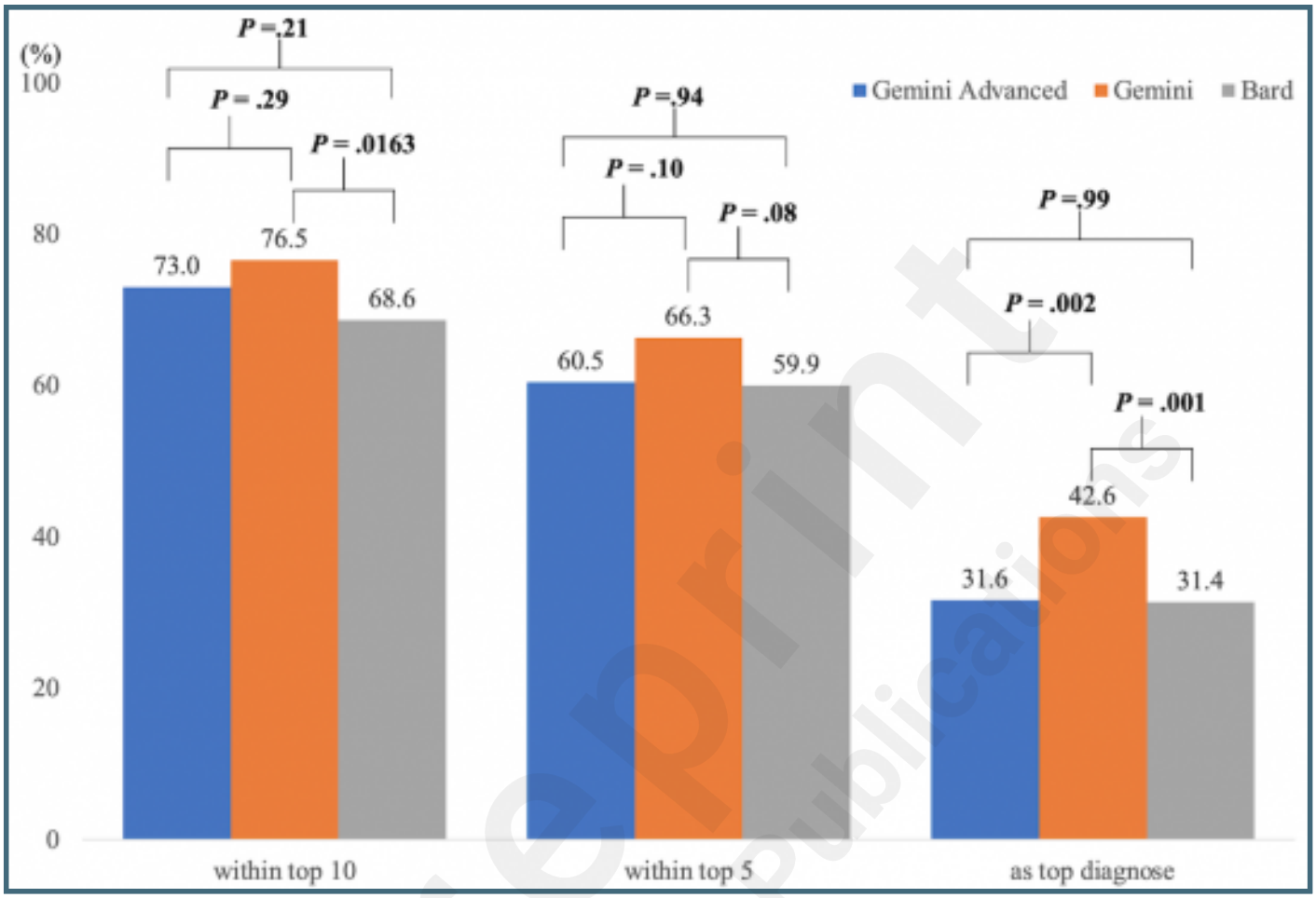
Supplementary Files

Figures

Study flow of Inclusion Case Materials and Generation of Differential Diagnosis Lists.



Diagnostic accuracy of Gemini Advanced, Gemini, and Bard. P values were derived from Chi-squared test. The Bonferroni-corrected significance level at a P value < .0167.



Multimedia Appendixes

The PubMed search keywords.
URL: http://asset.jmir.pub/assets/c25c0cd51bf7fb707c5171347bf35764.docx

The dataset of differential diagnosis generated by AI systems in this study, alongside the final diagnosis.
URL: http://asset.jmir.pub/assets/438d22ccb9eeeb2b5d49e81efd5ba5f6.xlsx

CONSORT (or other) checklists

CONSORT checklist.
URL: http://asset.jmir.pub/assets/4dcba862138f3844da1e25ee5153ae8b.pdf