Evaluating ChatGPT-4's accuracy in identifying final diagnoses within differential diagnoses compared to those of physicians: an experimental study for diagnostic cases

Takanobu Hirosawa, Yukinori Harada, Kazuya Mizuta, Tetsu Sakamoto, Kazuki Tokumasu, Taro Shimizu

Submitted to: JMIR Formative Research
on: April 08, 2024

Table of Contents

Original Manuscript **5**
Supplementary Files **19**
Figures **20**
Figure 1 ..... **21**
Figure 2 ..... **22**
Figure 3 ..... **23**
Figure 4 ..... **24**
Multimedia Appendixes **25**
Multimedia Appendix 1 ..... **26**
Multimedia Appendix 2 ..... **26**
Multimedia Appendix 3 ..... **26**
CONSORT (or other) checklists **27**
CONSORT (or other) checklist 0 ..... **27**
Related publication(s) - for reviewers eyes onlies **28**
Related publication(s) - for reviewers eyes only 0 ..... **28**

Evaluating ChatGPT-4's accuracy in identifying final diagnoses within differential diagnoses compared to those of physicians: an experimental study for diagnostic cases

Takanobu Hirosawa[1] MD, PhD; Yukinori Harada[1] MD, PhD; Kazuya Mizuta[1] MD; Tetsu Sakamoto[1] MD; Kazuki Tokumasu[2] MD, PhD; Taro Shimizu[1] MD, PhD, MPH, MBA, FACP

[1]Department of Diagnostic and Generalist Medicine Dokkyo Medical University Tochigi JP
[2]Department of General Medicine Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences Okayama JP

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

Abstract

Background: The potential of artificial intelligence (AI) chatbots, particularly the fourth-generation chat generative pretrained transformer (ChatGPT-4), in assisting with medical diagnosis is an emerging research area. However, it is not yet clear how well AI chatbots can evaluate whether the final diagnosis is included in differential-diagnosis lists.

Objective: This study aimed to assess the capability of ChatGPT-4 in identifying the final diagnosis from differential-diagnosis lists, and to compare its performance with that of physicians, for case report series.

Methods: We utilized a database of differential-diagnosis lists from case reports in the American Journal of Case Reports, corresponding to final diagnoses. These lists were generated by three artificial intelligence (AI) systems: ChatGPT-4, Google Bard (currently Google Gemini), and Large Language Models by Meta AI 2 chatbot. The primary outcome was focused on whether ChatGPT-4's evaluations identified the final diagnosis within these lists. None of these AIs received additional medical training or reinforcement. For comparison, two independent physicians also evaluated the lists, with any inconsistencies resolved by another physician.

Results: Three AIs generated a total of 1,176 differential diagnosis lists from 392 case descriptions. ChatGPT-4's evaluations concurred with those of the physicians in 966 out of 1,176 lists (82.1%). The Cohen kappa coefficient was 0.63 (95% confidence interval: 0.56-0.69), indicating a fair to good agreement between ChatGPT-4 and the physicians' evaluations.

Conclusions: ChatGPT-4 demonstrated a fair to good agreement in identifying the final diagnosis from differential-diagnosis lists, comparable to physicians for case report series. Its ability to compare differential-diagnosis lists with final diagnoses suggests its potential in aiding clinical decision-making support through diagnostic feedback. While ChatGPT-4 showed a fair to good agreement for evaluation, its application in real-world scenarios and further validation in diverse clinical environments are essential to fully understand its utility in the diagnostic process. Clinical Trial: Not applicable

(JMIR Preprints 08/04/2024:59267)
DOI: https://doi.org/10.2196/preprints.59267

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

Evaluating ChatGPT-4's accuracy in identifying final diagnoses within differential diagnoses compared to those of physicians: an experimental study for diagnostic cases

Authors: Takanobu Hirosawa[1] MD, PhD, Yukinori Harada[1] MD, PhD, Kazuya Mizuta[1] MD, Tetsu Sakamoto[1] MD, Kazuki Tokumasu[2] MD, PhD, Taro Shimizu[1] MD, PhD, MPH, MBA, FACP

Affiliations:
1. Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Tochigi, Japan
2. Department of General Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, 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: The potential of artificial intelligence (AI) chatbots, particularly the fourth-generation chat generative pretrained transformer (ChatGPT-4), in assisting with medical diagnosis is an emerging research area. However, it is not yet clear how well AI chatbots can evaluate whether the final diagnosis is included in differential-diagnosis lists.

Objective: This study aimed to assess the capability of ChatGPT-4 in identifying the final diagnosis from differential-diagnosis lists, and to compare its performance with that of physicians, for case report series.

Methods: We utilized a database of differential-diagnosis lists from case reports in the *American Journal of Case Reports*, corresponding to final diagnoses. These lists were generated by three artificial intelligence (AI) systems: ChatGPT-4, Google Bard (currently Google Gemini), and Large Language Models by Meta AI 2 chatbot. The primary outcome was focused on whether ChatGPT-4's evaluations identified the final diagnosis within these lists. None of these AIs received additional medical training or reinforcement. For comparison, two independent physicians also evaluated the lists, with any inconsistencies resolved by another physician.

Results: Three AIs generated a total of 1,176 differential diagnosis lists from 392 case descriptions. ChatGPT-4's evaluations concurred with those of the physicians in 966 out of 1,176 lists (82.1%). The Cohen kappa coefficient was 0.63 (95% confidence interval: 0.56-0.69), indicating a fair to good agreement between ChatGPT-4 and the physicians' evaluations.

Conclusions: ChatGPT-4 demonstrated a fair to good agreement in identifying the final diagnosis from differential-diagnosis lists, comparable to physicians for case report series. Its ability to compare differential-diagnosis lists with final diagnoses suggests its potential in aiding clinical decision-making support through diagnostic feedback. While ChatGPT-4 showed a fair to good agreement for evaluation, its application in real-world scenarios and further validation in diverse clinical environments are essential to fully understand its utility in the diagnostic process.

Trial Registration: Not applicable

Keywords: Decision Support System; Diagnostic Errors; Diagnostic Excellence; Diagnosis, Large Language Model, Natural Language Processing

Introduction

Diagnostic Error and Feedback

A well-developed diagnostic process is fundamental to medicine. Diagnostic errors [1], which include missed, incorrect, or delayed diagnoses [2], result in severe misdiagnosis-related harm, affecting up to 795,000 patients annually in the United States [3]. These errors often stem from a failure to correctly identify an underlying condition [4, 5]. Enhancing the diagnostic process is crucial, with diagnostic feedback playing a key role [6]. The feedback enables physicians to assess their diagnostic accuracy and adjust their subsequent clinical decisions accordingly [7]. Common diagnostic feedback methods include self-reflection [8, 9], peer review [1], and clinical decision support systems (CDSSs), which aim to enhance decision-making at the point of care [10]. Unlike the retrospective nature of self and peer review processes, feedback from CDSSs is provided in real-time [11], offering immediate support and guidance during the diagnostic process. This timely feedback is particularly advantageous in fast-paced clinical settings where timely decision-making is critical.

Clinical Decision Support Systems and Artificial Intelligence

CDSSs are categorized into two main types: knowledge-based and non-knowledge-based systems [10]. Knowledge-based CDSSs rely on established medical knowledge, including clinical guidelines, expert protocols, and information on drug interactions. In contrast, non-knowledge-based systems, particularly those utilizing artificial intelligence (AI), leverage advanced algorithms, machine learning, and statistical pattern recognition. Unlike their rule-based counterparts, these systems adapt over time, continuously refining their insights and recommendations. The rapid integration of AI into CDSSs highlights the growing importance of advanced technologies in healthcare [12]. In recent years, generative AI through large language models (LLMs) has been reshaping healthcare, offering improvements in diagnostic accuracy, treatment planning, and patient care [13, 14]. AI systems, emulating human cognition, continuously learn from new data [15]. They assist healthcare professionals by analyzing complex patient data, thereby enhancing clinical decision-making and patient outcomes [10].

Growing Importance of Generative Artificial Intelligence

In this context of rapidly integrating AI into CDSSs, generative AIs have marked a new era in digital health. LLMs are advanced AI algorithms trained on extensive textual data, enabling them to process and generate human-like text, thereby providing valuable insights in medical diagnostics. Several generative AI tools are now available to the public, including Bard (currently Gemini) by Google [16, 17], LLM Meta AI 2 (LLaMA2) by Meta AI [18], and the Chat Generative Pre-trained Transformer (ChatGPT) developed by OpenAI [19]. These AI tools, which utilize LLMs, have successfully passed national medical licensing exams without specific training or reinforcement [20], demonstrating their potential in medical diagnostics. Among these, ChatGPT stands out as one of the most extensively researched generative AI applications in healthcare [21]. Specifically, in diagnostics, a recent study has shown that these generative AI systems, particularly ChatGPT-4, demonstrate excellent diagnostic capability when answering clinical vignettes questions [22]. Additionally, other studies, including our own, have assessed AI systems' performance in one aspect of the diagnostic process, generating differential-diagnosis lists [23-25]. While broader studies compare a variety of state-of-the-art models, our analysis focuses on the distinct capabilities and impacts of these specific tools within medical diagnostics.

Generative Artificial Intelligence Systems in the Diagnostic Process

The diagnostic process involves collecting clinical information, forming a differential diagnosis, and refining it through continuous feedback [26]. This feedback consists of patient outcomes, test results, and final diagnoses [27, 28]. Similar to traditional CDSSs, generative AI systems can enhance this feedback loop [29]. However, a gap previously existed in the systematic comparison of differential diagnoses with final diagnoses through a feedback loop [27]. Given this background, it remains less explored how effectively these AI systems integrate their feedback into clinical workflow. To address this gap, exploring how generative AI systems provide feedback by comparing final diagnoses with differential-diagnosis lists represents a straightforward and viable first step. In this study, we utilized differential-diagnosis lists to assess the diagnostic accuracy. This approach was chosen to mimic a key aspect of the clinical decision-making process, where physicians often narrow down a broad list of potential diagnoses to determine the most likely one. This method reflects a critical use case for AI in healthcare, potentially speeding up and refining diagnostic accuracy. In our previous short communication, we reported that the fourth generation ChatGPT (ChatGPT-4) showed very good agreement with physicians in evaluating the lists for a limited number of case reports published from our General Internal Medicine (GIM) department [30]. Building on this research, our current study focused on assessing the capability of ChatGPT-4 in identifying the final diagnosis from differential-diagnosis lists for comprehensive case report series, compared to those of physicians. Furthermore, this research aimed to demonstrate the role of generative AI, particularly ChatGPT-4, in enhancing the diagnostic learning cycle through effective feedback mechanisms.

Methods

Overview

We conducted an experimental study utilizing ChatGPT-4 and the differential-diagnosis lists generated by three AI systems inputting into case descriptions. The research was conducted at the Department of Generalist and Diagnostic Medicine (GIM), Dokkyo Medical University, Tochigi, Japan. Our research methodology encompassed preparing a dataset for differential-diagnosis lists and the corresponding final diagnoses, assessing these lists using ChatGPT-4, and having physicians evaluate the lists. Figure 1 illustrates the current study flow.

Ethical Considerations

Since we used a database extracted from published case reports, obtaining ethical approval was not applicable.

Database of Differential-diagnosis Lists and Final Diagnoses

We utilized our dataset from a previous study (Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokomasu K, Shimizu T. Diagnostic Performance of Generative Artificial Intelligences for a Series of Complex Case Reports. unpublished data, November 2023). From the PubMed search, we identified a total 557 case reports. We excluded the non-diagnosed cases (130 cases) and the pediatric cases, ages under 10 years old (35 cases). The exclusion criteria were based on the previous research for CDSS [31]. After the exclusion, we included 392 case reports. The case reports were brushed up as case descriptions to focus the diagnosis. The authors typically defined the final diagnoses. Through inputting into the case descriptions and systematic prompt, three generative AI systems – ChatGPT-4, Google Bard (currently Google Gemini), and LLaMA2 chatbot – generated the top ten differential-diagnosis lists. The utilized AI systems were not trained for any additional medical use or reinforced. The main investigator (TH) conducted the entire process, with validation provided by another investigator (YH). Through this process, this dataset included differential-diagnosis lists corresponding to case descriptions and final diagnoses from case reports in the *American Journal of Case Reports*. Detailed lists of differential diagnoses and their final diagnoses are shown in the Multimedia Appendix 1.

ChatGPT-4 Assessment of the Differential-diagnosis Lists

In selecting the generative AI systems for evaluation, we focused on ChatGPT-4 due to its distinct architectural frameworks and widespread usage in the field of healthcare research. ChatGPT-4, developed by OpenAI, is notable for its advanced natural language processing capabilities and extensive training dataset, making it particularly relevant for healthcare [32]. We utilized the August 3 version and September 25 version of ChatGPT-4 to evaluate differential-diagnosis lists. The access date was from September 11th, 2023, to October 6th, 2023. A structured prompt was crafted to ascertain whether ChatGPT-4 could identify the final diagnosis within a list and its position if present. The prompt required direct copying and pasting of the final diagnoses and differential-diagnosis lists from our dataset. We assessed the inclusion of the final diagnosis in the list (Yes = 1, No = 0) and its position. The prompt selection was a preliminary investigation. To ensure unbiased output, each session was isolated by deactivating chat history and training controls, and restarting ChatGPT-4 before every new evaluation. We obtained a single output from ChatGPT-4 for each differential-diagnosis list. The details of this structured prompt in the current study are expounded in Multimedia Appendix 2.

Physician assessment of the differential-diagnosis lists

For comparison, two independent physicians (KM and TS, Tetsu Sakamoto) also evaluated the differential-diagnosis lists. The presence of the final diagnosis within the differential-diagnosis lists was marked with a 1 or 0. A "1" was marked when the lists precisely and acceptably identified the

final diagnosis [33], further ranking it from 1-10 based on its placement. A "0" indicated its absence. Discrepancies between the evaluations of the two physicians were resolved by another physician (KT). Notably, the physicians were blinded to which AI generated the lists they assessed. We selected three independent physicians, specializing in general internal medicine. Selection was based on expertise in diagnostic processes and familiarity with AI technologies in healthcare. All physicians underwent a brief guidance session to familiarize themselves with the evaluation criteria and objectives of the study to ensure consistent assessment standards.

Outcome

The primary outcome was defined as the kappa coefficient for inter-rater agreement between ChatGPT-4 and the physicians' evaluations for the differential-diagnosis lists generated by three AI systems, including ChatGPT-4, Google Bard (currently Google Gemini), and LLaMA2 chatbot. The secondary outcomes were defined as the kappa coefficients for inter-rater agreement between ChatGPT-4 and the physicians' evaluations for the differential-diagnosis lists generated by each AI system. Additionally, another secondary outcome was defined as the ranking patterns between ChatGPT-4's evaluation and that of physicians.

Statistical Analysis

Analytical procedures were conducted using R version 4.2.2 (The R Foundation for Statistical Computing, Vienna, Austria). The agreement between different evaluations was quantified using Cohen's kappa coefficient through the irr package in R. Agreement strength was categorized as per Cohen's kappa benchmarks: values under 0.40 indicated poor agreement; values between 0.41 and 0.75 showed fair to good agreement; and values ranging from 0.75 to 1.00 denoted very good agreement [34]. The 95% CIs were used to quantify uncertainty. Additionally, we compared ranking patterns between ChatGPT-4's evaluation and that of physicians [35].

Results

Overall Evaluation

This study involved three generative AI systems – ChatGPT-4, Google Bard (currently Google Gemini), and LLaMA2 chatbot – outputting differential-diagnosis lists for 392 case descriptions, resulting in a total of 1,176 lists. In 825 lists where physicians included a final diagnosis, ChatGPT-4 matched 636 lists and did not match 189 lists. Conversely, in 351 lists where physicians did not include a final diagnosis, ChatGPT-4 matched 330 lists and did not match 21 lists. Totally, ChatGPT-4's evaluations matched the physicians' evaluations in 966 out of 1,176 lists (82.1%). Cohen's kappa coefficient was 0.63 (95% CI: 0.56-0.69), indicating a fair to good agreement between ChatGPT-4 and the physicians' evaluations. ChatGPT-4 omitted the final diagnosis in 16.1% of cases (189/1,176), contrasting with physicians' evaluations that included these diagnoses. Table 1 shows ChatGPT-4's evaluations concurred with the physicians' evaluations. Table 2 details the kappa coefficient for inter-rater agreement between ChatGPT-4 and the physicians' evaluations. The representative input utilized in ChatGPT-4's evaluations is illustrated in Figure 2, and the corresponding output is shown in Figures 3. A formed dataset is shown in Multimedia Appendix 3.

Table 1. ChatGPT-4's evaluations concurred with the physicians' evaluations.

Variables (N=1,176)	ChatGPT-4		Total
	Inclusion of final diagnosis	Non-inclusion of final diagnosis	

Physicians				
	Inclusion of final diagnosis			
		636	189	825
	Non-inclusion of final diagnosis			
		21	330	351

Table 2. Kappa coefficient for inter-rater agreement between ChatGPT-4 and the physicians' evaluations for the differential-diagnosis lists.

Differential-diagnosis lists generator	Cohen's kappa coefficient (95%CI)	Strength of Agreement[a]	Number of differential-diagnosis list
All			
	0.63 (0.56-0.69)	Fair to Good	1,176
ChatGPT-4			
	0.47 (0.39-0.56)	Fair to Good	392
Google Bard[b]			
	0.67 (0.52-0.73)	Fair to Good	392
LLaMA2 Chatbot			
	0.63 (0.52-0.73)	Fair to Good	392

[a]Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. John Wiley & Sons; 2003. New York.
[b]Currently Google Gemini.

Evaluation for Each Generative Artificial Intelligence

The kappa coefficients for differential-diagnosis lists generated by ChatGPT-4, Google Bard (currently Google Gemini), and LLaMA2 chatbot were 0.47 (95% CI: 0.39-0.56), 0.67 (95% CI: 0.52-0.73), and 0.63 (95% CI: 0.52-0.73), respectively. All kappa coefficients indicated a fair to good agreement between ChatGPT-4 and the physicians' evaluations.

Comparison of Ranking Patterns Between ChatGPT-4 and Physicians

Both ChatGPT-4's evaluation and that of physicians showed a general trend of decreasing frequency as the rank increases. Figure 4 shows the comparisons of ranking patterns between ChatGPT-4 and physicians.

Evaluation Between Physicians

Physicians' evaluations (KM and TS, Tetsu Sakamoto) for the differential-diagnosis lists showed very good agreement, with concordance in 88.8% (1,044/1,176) of cases. The kappa coefficient was 0.75 (95% CI: 0.46-0.99).

Discussion

Principal Results

This experimental study highlights several key findings. First, ChatGPT-4's evaluations matched those of physicians in more than 82% of the cases, demonstrating fair to good agreement according to kappa coefficient values. These results imply that ChatGPT-4's accuracy in identifying the final diagnosis within differential-diagnosis lists is comparable to that of physicians. Unlike traditional CDSSs, generative AI systems, including ChatGPT-4, are capable of performing multiple roles in the diagnostic process, including formulating and assessing differential diagnoses. These capabilities highlight ChatGPT-4's potential to streamline diagnostics in clinical settings by expediting diagnostic feedback [36]. Our study design focuses on ChatGPT-4's ability to refine and to validate pre-existing diagnostic considerations as supplementary tools for medical diagnostics. This scenario is akin to real-world clinical settings where generative AI systems could verify and support physicians' final diagnostic decisions. By assessing the AI's accuracy in this context, we can better understand its potential role and limitations in practical medical applications. Furthermore, in medical education, generative AI tools like ChatGPT-4 can offer students valuable self-learning opportunities. They provide timely feedback in the form of final diagnoses [37], enabling them to cross-reference with reliable sources for verification [38].

Second, ChatGPT-4 failed to identify the final diagnosis in 16% of differential-diagnosis lists, even though these diagnoses were recognized by the evaluating physicians. Notably, despite achieving very good agreement among physicians, ChatGPT-4 did not reach similar levels of concordance. This discrepancy highlights potential areas for improving the system's ability to interpret and analyze complex medical data. This discrepancy arises primarily from ChatGPT-4's reliance on textual patterns and word associations within the provided differential-diagnosis lists. Unlike physicians, who utilize a comprehensive medical knowledge base and clinical experience, an inherent limitation in generative AI systems like ChatGPT-4 is their reliance on existing data patterns and textual association. To mitigate these discrepancies, continuous development in generative AI systems for healthcare is needed. Additionally, future research should focus on enhancing the medical training of these systems. This will enhance the generative AI systems' diagnostic feedback, making it more adaptable to real clinical settings.

Third, regarding evaluation at what rank in the differential-diagnosis list was the final diagnosis found, both ChatGPT-4 and physicians exhibited a trend of decreasing frequency. This suggests ChatGPT-4's diagnosis ranking shows a similar trend of physicians' diagnosis ranking. Moreover, all three generative AI systems, including ChatGPT-4, Google Bard (currently Google Gemini), and LLaMA2 chatbot, prioritized the most likely diagnoses at the top of the list, leading to a natural decrease in frequency as less probable diagnoses are ranked lower. Therefore, generative AI systems showed the potential not only to generate differential-diagnosis lists for clinical cases but also to evaluate these lists as feedback.

Fourth, an examination of the differential-diagnosis lists generated by three different AI systems showed the overlap in the 95%CI for the kappa coefficients across the three AI platforms. One might hypothesize that ChatGPT-4 would exhibit improved performance when evaluating differential-diagnosis lists it generated itself. However, observed results may stem from the inherent variability in generative AI outputs, including ChatGPT-4. This inherent variability underscores the challenge of maintaining a consistent standard of accuracy and reliability in the outputs from generative AI systems. Even when evaluating differential-diagnosis lists generated by itself, ChatGPT-4's performance did not markedly surpass that of lists generated by other AI systems. Additionally, the observed performance differences may be partially due to version inconsistencies. The generation of differential-diagnosis lists utilized an earlier version of ChatGPT-4 (March 24). Subsequent

evaluations employed later versions (August 3 and September 25). Different versions of generative AI systems can exhibit varied capabilities and outputs, potentially impacting the accuracy and consistency of diagnostic evaluations. This highlights the need for ongoing updates and versions alignment in clinical AI applications to maintain reliability.

Limitations

This study has several limitations. First, ChatGPT-4's role was limited to identifying the final diagnosis within the differential-diagnosis list. The current binary evaluation method has not been a well-established approach in evaluating diagnostic performance by other CDSSs. Another study employed a five-grade level of accuracy for a variety number of differentials [39]. Investigating more complex outcomes, such as quantitative evaluations and additional clinical suggestions, might yield different results. Second, our inputs to ChatGPT-4 consisted only of the final diagnoses and the differential-diagnosis list, without the case descriptions that generated these lists. Further research should examine what types of input enhance AI systems' performance the most. Third, there was a non-negligible risk associated with generative AI systems, including ChatGPT-4, regarding their capacity to inadvertently learn from and replicate the information contained with publicly available case reports. Fourth, the dataset was sourced from a single case reports journal and generated by three AI systems. Future research would benefit from using real-world scenarios [40]. Expanding the dataset to include a more diverse range of AI systems is also advisable.

Regarding limitations for generative AI systems like ChatGPT-4, there is currently no approval for their use as CDSSs. Furthermore, ChatGPT-4 operates as a fee-based application, which could potentially limit its accessibility to the wider public. Additionally, the reliability of generative AI systems can vary based on the input data it was trained on. If it is not exposed to diverse clinical scenarios during its training, it may not be as effective in real-world diagnostic situations [41]. Moreover, while AI tools can assist, they do not replace the nuanced judgments and decision-making processes of human physicians [42, 43]. Additionally, the rapid evolution of AI means that our findings may become outdated as Google Bard and LLaMA2 were updated to the new LLM model, Google Gemini and LLaMA3, respectively [17, 44]. Lastly, over-reliance on AI without critical review could lead to diagnostic errors [45].

Comparison with Prior Work

In our previous study involving ChatGPT-4 [30], we observed a very good agreement with physicians in identifying final diagnosis within the differential-diagnosis lists, achieving a 95.9% agreement rate (236 out of 246 lists, kappa = 0.86). In contrast, the current study demonstrated a fair to good agreement rate of 82.1% (966 out of 1,176 lists, kappa = 0.63). Despite employing the same evaluation methods in both studies, the observed decrease in agreement can be attributed to several factors: the source of case reports (GIM-published versus a broader range of case reports), the generators of differential diagnoses (physicians/ChatGPT-3/ChatGPT-4 versus ChatGPT-4/Google Bard [currently Gemini]/LLaMA2 chatbot), and the volume of lists assessed (246 lists versus 1,176 lists).

Future Directions

Future studies explore the potential of integrating ChatGPT-4 and similar AI systems into real-world clinical settings. This could involve developing interfaces that allow these AI systems to directly interact with electronic health records, providing real-time diagnostic feedback to physicians. Additionally, research could focus on tailoring these AI systems for specialized medical fields, where their ability to process vast amounts of data could significantly aid in complex case analysis. Another vital area for future research is the ethical implications of AI in medicine [43], particularly in patient

data privacy, AI decision transparency, and the impact of AI-assisted diagnostics on physician-patient relationships.

Furthermore, further research should also investigate the optimal use of AI technologies, including the exploration of both chatbot interfaces and Application Programming Interface (API) functionalities. A more detailed examination of API settings, such as adjustable parameters including temperature and Top P, could be invaluable. This investigation would provide clearer guidelines on when and how to use different AI tools effectively, considering both scientific evidence and effectiveness.

Moreover, our future research will focus on refining the evaluation of AI-generated differential diagnoses by incorporating more sophisticated and validated psychometric methods as the next diagnostic step. We propose to adopt methodologies for assessing the quality of differential diagnoses. This approach will allow us not only to compare AI-generated outputs with those from physicians but also to treat it as a form of Turing test—evaluating whether AI can match or surpass human performance in diagnostic tasks without being distinguishable from them [46].

Conclusions

ChatGPT-4 demonstrated a fair to good agreement in identifying the final diagnosis from differential-diagnosis lists, comparable to physicians for case report series. By reliably identifying diagnoses, ChatGPT-4 can provide on-time feedback by comparing final diagnoses with differential-diagnosis lists. Therefore, this study suggests that generative AI systems have a potential to assist physicians in the diagnostic process by providing reliable and efficient feedback, thereby contributing to improved clinical decision-making and medical education. However, it is imperative to recognize that these findings are based on experimental study. Real-world scenarios could present unique challenges, and further validations in diverse clinical environments are essential before broad implementation can be recommended.

Acknowledgements

Contributors: TH, YH, KM, TS (Tetsu Sakamoto), KT, 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, KM, TS (Tetsu Sakamoto), KT, 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

CDSS: Clinical Decision Support System
AI: Artificial Intelligence
LLM: Large Language Model
GIM: General Internal Medicine

API: Application Programming Interface

Multimedia Appendix 1

The differential-diagnosis generated by three artificial intelligences utilized in this study and the final diagnosis.

Multimedia Appendix 2

Structured prompt utilized in the current study.

Multimedia Appendix 3

Formed dataset utilized in the current study.

References

1. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015.
2. Graber M. Diagnostic errors in medicine: a case of neglect. Jt Comm J Qual Patient Saf. 2005;31(2):106-13.
3. 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.
4. Graber ML, Franklin N, Gordon R. Diagnostic Error in Internal Medicine. Archives of internal medicine. 2005;165(13):1493-9.
5. Schiff GD, Hasan O, Kim S, Abrams R, Cosby K, Lambert BL, et al. Diagnostic Error in Medicine: Analysis of 583 Physician-Reported Errors. Archives of internal medicine. 2009;169(20):1881-7.
6. Singh H, Connor DM, Dhaliwal G. Five strategies for clinicians to advance diagnostic excellence. Bmj. 2022;376:e068044.
7. Meyer AND, Singh H. The Path to Diagnostic Excellence Includes Feedback to Calibrate How Clinicians Think. JAMA. 2019;321(8):737-8.
8. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Medical education. 2008;42(5):468-75.
9. Mamede S, Schmidt HG. Reflection in Medical Diagnosis: A Literature Review. Health Professions Education. 2017;3(1):15-25.
10. 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.
11. Rubins D, McCoy AB, Dutta S, McEvoy DS, Patterson L, Miller A, et al. Real-Time User Feedback to Support Clinical Decision Support System Improvement. Appl Clin Inform. 2022;13(5):1024-32.
12. Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. New England Journal of Medicine. 2023;388(13):1201-8.
13. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 2023;25:e48568.
14. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education. 2023;23(1):689.
15. Collins C, Dennehy D, Conboy K, Mikalef P. Artificial intelligence in information systems research: A systematic literature review and research agenda. International Journal of

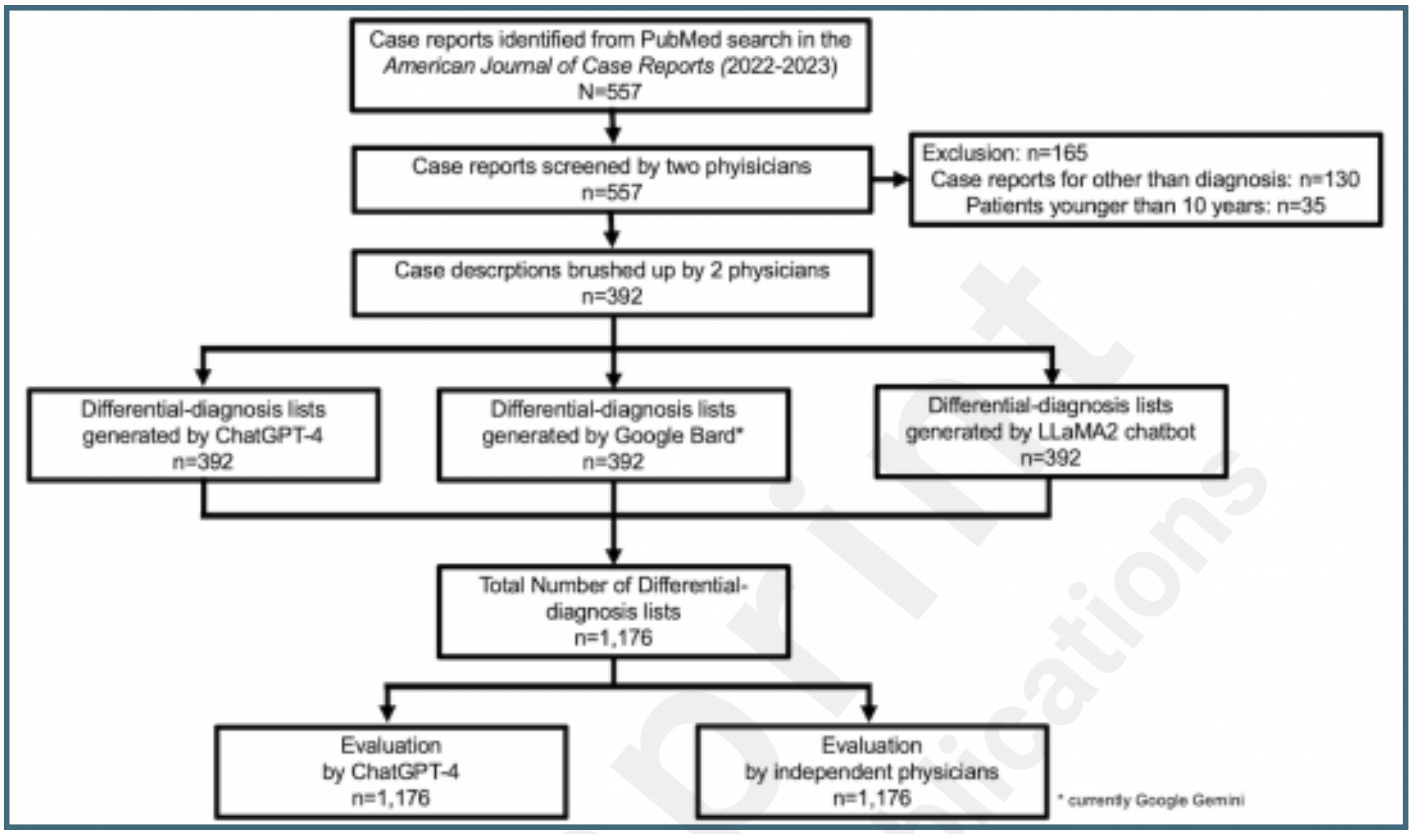
Information Management. 2021;60:102383.
16. Patrizio A. Google Bard: TechTarget; 2023 [Available from: https://www.techtarget.com/searchenterpriseai/definition/Google-Bard.
17. Sundar Pichai DH. Introducing Gemini: our largest and most capable AI model. 2023. [Available from: https://blog.google/technology/ai/google-gemini-ai/#sundar-note
18. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.
19. OpenAI. GPT-4 Technical Report2023 March 01, 2023:[arXiv:2303.08774 p.].
20. 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.
21. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47(1):33.
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. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-Generated Differential Diagnosis Lists for Complex Case-Derived Clinical Vignettes: Diagnostic Accuracy Evaluation. JMIR Med Inform. 2023;11:e48808.
24. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative Evaluation of Diagnostic Accuracy Between Google Bard and Physicians. Am J Med. 2023. Nov;136(11):1119-1123.e18.
25. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330(1):78-80.
26. Price RB, Vlahcevic ZR. Logical principles in differential diagnosis. Ann Intern Med. 1971;75(1):89-95.
27. Branson CF, Williams M, Chan TM, Graber ML, Lane KP, Grieser S, et al. Improving diagnostic performance through feedback: the Diagnosis Learning Cycle. BMJ Quality & Safety. 2021;30(12):1002-9.
28. Rosner BI, Zwaan L, Olson APJ. Imagining the future of diagnostic performance feedback. Diagnosis. 2023;10(1):31-7.
29. Filiberto AC, Leeds IL, Loftus TJ. Editorial: Machine Learning in Clinical Decision-Making. Front Digit Health [Internet]. 2021 2021; 3:[784495 p.]. Available from: http://europepmc.org/abstract/MED/34870273
https://www.frontiersin.org/articles/10.3389/fdgth.2021.784495/pdf
https://doi.org/10.3389/fdgth.2021.784495
https://europepmc.org/articles/PMC8636718
https://europepmc.org/articles/PMC8636718?pdf=render.
30. Mizuta K, Hirosawa T, Harada Y, Shimizu T. Can ChatGPT-4 evaluate whether a differential diagnosis list contains the correct diagnosis as accurately as a physician? Diagnosis 2024 Mar 12. doi: 10.1515/dx-2024-0027. Online ahead of print.
31. 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.
32. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:230313375. 2023.
33. 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.
34. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions: john wiley &

sons; 2003. New York.
35. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans Inf Syst. 2010;28(4):Article 20.
36. Hattie J, Timperley H. The power of feedback. Review of educational research. 2007;77(1):81-112.
37. Chamberland M, Setrakian J, St-Onge C, Bergeron L, Mamede S, Schmidt HG. Does providing the correct diagnosis as feedback after self-explanation improve medical students diagnostic performance? BMC Medical Education. 2019;19(1):194.
38. Abd-alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ. 2023;9:e48291.
39. 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.
40. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online Symptom Checkers: Recommendations for a Vignette-Based Clinical Evaluation Standard. J Med Internet Res. 2022;24(10):e37408.
41. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40.
42. Karches KE. Against the iDoctor: why artificial intelligence should not replace physician judgment. Theoretical Medicine and Bioethics. 2018;39(2):91-110.
43. WHO. Ethics and governance of artificial intelligence for health: WHO guidance. 2021.
44. Meta AI. Build the future of AI with Meta Llama 3 2024 [Available from: https://llama.meta.com/llama3/.
45. Passi S, Vorvoreanu M. Overreliance on AI Literature Review. Microsoft Technical Report MSR-TR-2022-12. Microsoft Corporation.
46. Pinar Saygin A, Cicekli I, Akman V. Turing test: 50 years later. Minds and machines. 2000;10(4):463-518.

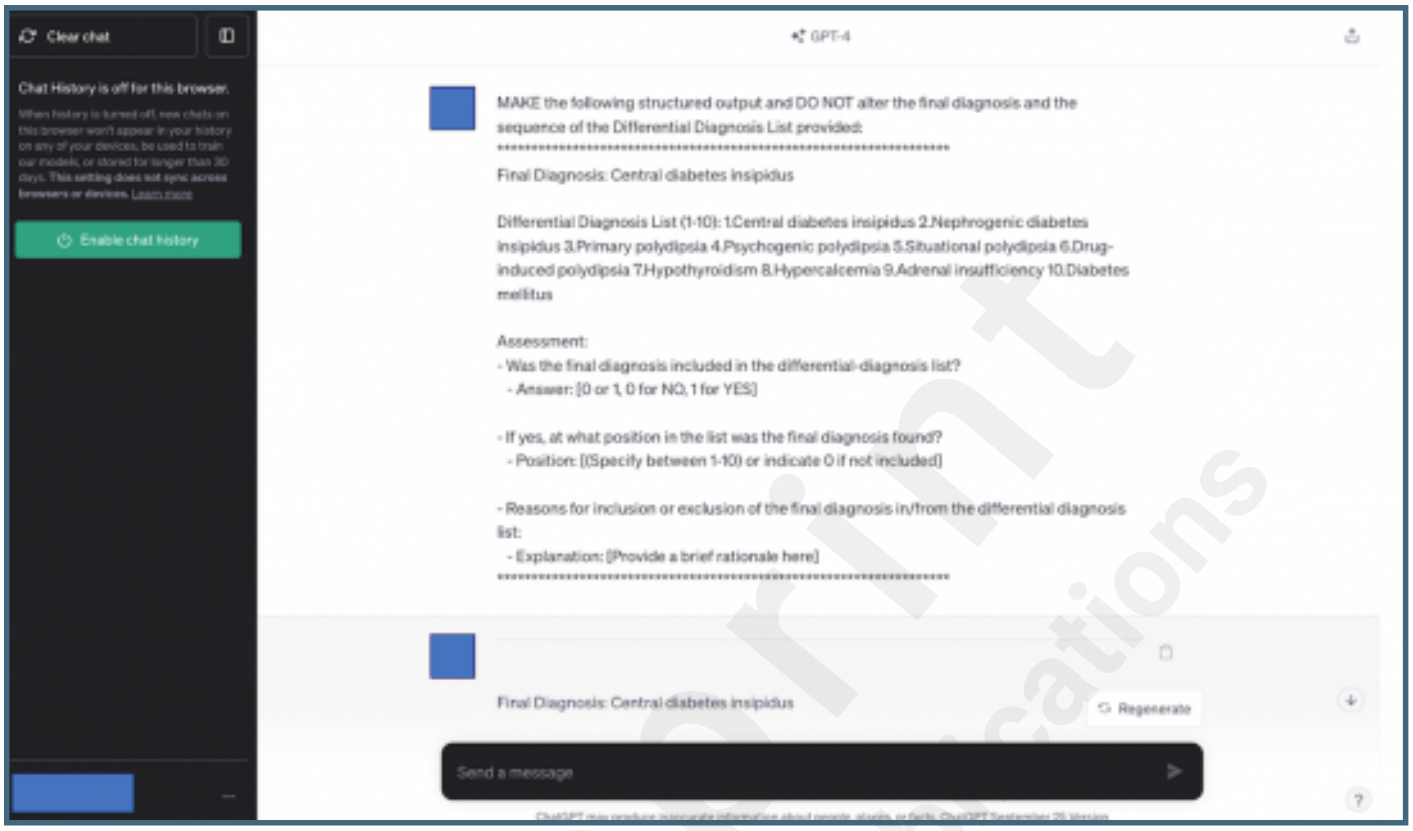
Supplementary Files

Figures

Study Flow Chart of Inclusion of Case Reports, Generation of Differential-diagnosis Lists, and Evaluation of the Lists.

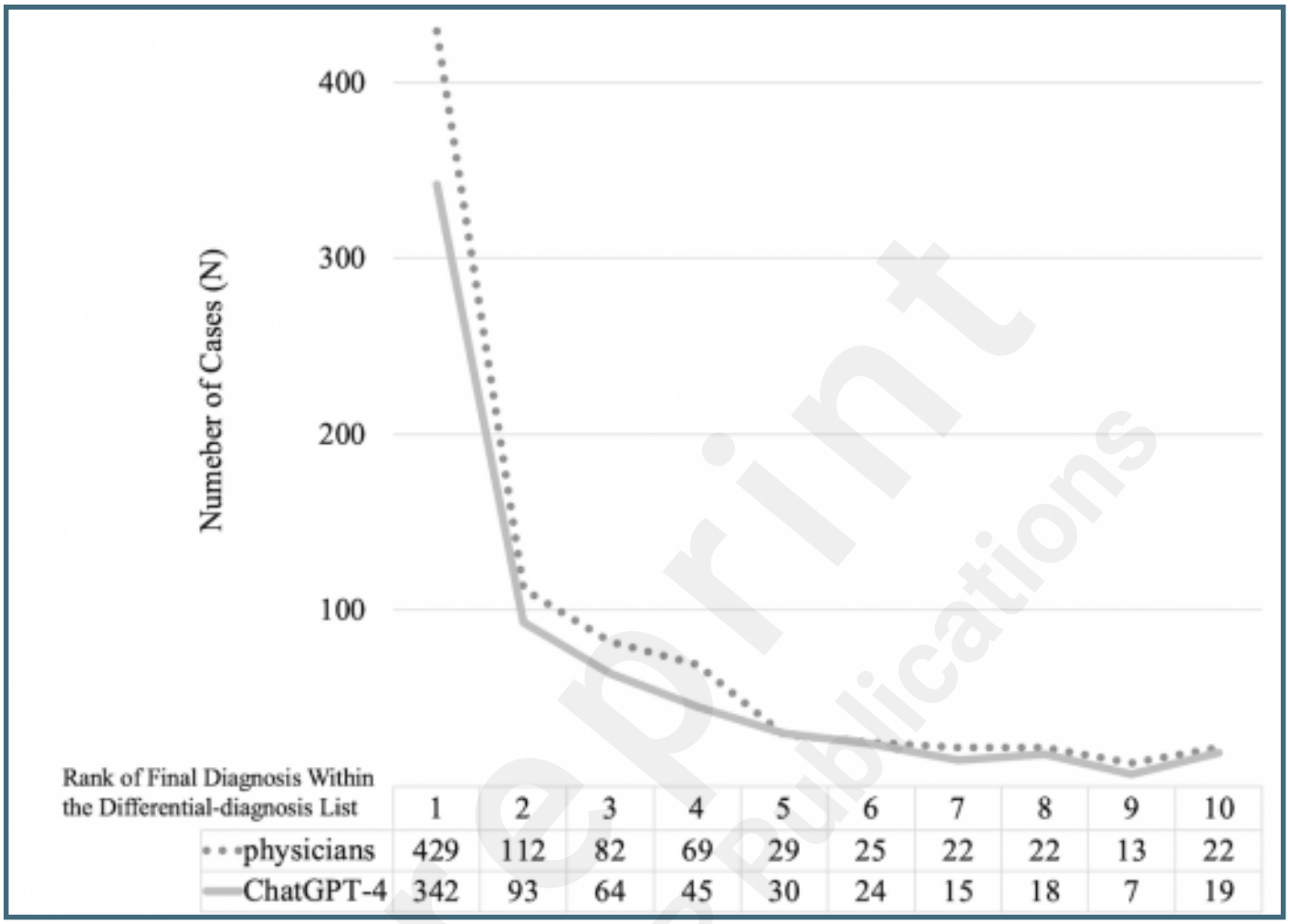


The Representative Input for the 4th Generation Chat Generative Pre-trained Transformer Generated to Evaluate Whether the Final Diagnosis was Included in the Differential-Diagnosis.



From the Input (Figure 2), the 4th Generation Chat Generative Pre-trained Transformer Generated the Representative Output of Evaluation.

Comparison of Ranking Patterns from Evaluation by the 4th Generation Chat Generative Pre-trained Transformer and Physicians.



Rank of Final Diagnosis Within the Differential-diagnosis List	1	2	3	4	5	6	7	8	9	10
physicians	429	112	82	69	29	25	22	22	13	22
ChatGPT-4	342	93	64	45	30	24	15	18	7	19

Multimedia Appendixes

The differential-diagnosis generated by three artificial intelligences utilized in this study and the final diagnosis.
URL: http://asset.jmir.pub/assets/3f860f2b565cd21c668603e16cc4b6cd.pdf

Structured prompt utilized in the current study.
URL: http://asset.jmir.pub/assets/a84d853efe3cfe38200d902d0a3ccad2.pdf

Formed dataset utilized in the current study.
URL: http://asset.jmir.pub/assets/1710e72467a37ad8bf3e803cbacddf54.xlsx

CONSORT (or other) checklists

Untitled.
URL: http://asset.jmir.pub/assets/537f7284dbff3af10f70d4d803ca2776.pdf

Related publication(s) - for reviewers eyes onlies

a revised version with tracked changes.
URL: http://asset.jmir.pub/assets/9efebd0c5094ce7e37869f43703760c2.pdf