






BMJ Open Quality Performance evaluation of ChatGPT in detecting diagnostic errors and their contributing factors: an analysis of 545 case reports of diagnostic errors

Yukinori Harada ¹, Tomoharu Suzuki,² Taku Harada,^{1,3} Tetsu Sakamoto,¹ Kosuke Ishizuka,⁴ Taiju Miyagami ⁵, Ren Kawamura ¹, Kotaro Kunitomo,⁶ Hiroyuki Nagano,⁷ Taro Shimizu ¹, Takashi Watari ⁸

To cite: Harada Y, Suzuki T, Harada T, *et al.* Performance evaluation of ChatGPT in detecting diagnostic errors and their contributing factors: an analysis of 545 case reports of diagnostic errors. *BMJ Open Quality* 2024;**13**:e002654. doi:10.1136/bmjopen-2023-002654

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjopen-2023-002654>).

Received 17 October 2023
Accepted 28 May 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Yukinori Harada;
yuki.gym23@gmail.com

ABSTRACT

Background Manual chart review using validated assessment tools is a standardised methodology for detecting diagnostic errors. However, this requires considerable human resources and time. ChatGPT, a recently developed artificial intelligence chatbot based on a large language model, can effectively classify text based on suitable prompts. Therefore, ChatGPT can assist manual chart reviews in detecting diagnostic errors.

Objective This study aimed to clarify whether ChatGPT could correctly detect diagnostic errors and possible factors contributing to them based on case presentations.

Methods We analysed 545 published case reports that included diagnostic errors. We imputed the texts of case presentations and the final diagnoses with some original prompts into ChatGPT (GPT-4) to generate responses, including the judgement of diagnostic errors and contributing factors of diagnostic errors. Factors contributing to diagnostic errors were coded according to the following three taxonomies: Diagnosis Error Evaluation and Research (DEER), Reliable Diagnosis Challenges (RDC) and Generic Diagnostic Pitfalls (GDP). The responses on the contributing factors from ChatGPT were compared with those from physicians.

Results ChatGPT correctly detected diagnostic errors in 519/545 cases (95%) and coded statistically larger numbers of factors contributing to diagnostic errors per case than physicians: DEER (median 5 vs 1, $p < 0.001$), RDC (median 4 vs 2, $p < 0.001$) and GDP (median 4 vs 1, $p < 0.001$). The most important contributing factors of diagnostic errors coded by ChatGPT were 'failure/delay in considering the diagnosis' (315, 57.8%) in DEER, 'atypical presentation' (365, 67.0%) in RDC, and 'atypical presentation' (264, 48.4%) in GDP.

Conclusion ChatGPT accurately detects diagnostic errors from case presentations. ChatGPT may be more sensitive than manual reviewing in detecting factors contributing to diagnostic errors, especially for 'atypical presentation'.

INTRODUCTION

Recent advances in artificial intelligence (AI) technology have accelerated its implementation in diagnostic processes in clinical practice and research on diagnostic processes. Since its introduction for public use in November

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Manual chart review using a standardised assessment tool is a reliable method for judging the presence or absence of diagnostic errors. However, manual chart reviews require significant human resources, which may limit studies of diagnostic excellence.

WHAT THIS STUDY ADDS

⇒ This study investigated the performance potential of ChatGPT in detecting diagnostic errors and the factors contributing to diagnostic errors by reviewing case presentation texts from case reports of diagnostic errors. ChatGPT correctly detected diagnostic errors in most cases. ChatGPT can also detect a larger number of contributing factors to diagnostic errors than physicians.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE, OR POLICY

⇒ This study suggests that ChatGPT may reduce the efforts and costs of manual chart reviews to judge diagnostic errors, enabling diagnostic excellence and diagnostic safety in healthcare.

2022, several studies have been conducted to evaluate the performance of ChatGPT in clinical diagnosis. ChatGPT is an AI chatbot developed based on large language models (generative pre-trained Transformer 3.5 and 4). ChatGPT produces accurate and detailed text-based responses to written prompts.

Positive data for ChatGPT have been reported in the field of clinical diagnosis. Previous studies have shown that ChatGPT can answer questions testing medical knowledge correctly at a pass level for national medical license examinations.^{1,2} Some authors have suggested that ChatGPT can be used to support clinical decisions.³ Some studies have shown that ChatGPT exhibits high performance in developing the final diagnosis as the top differential diagnosis in common and

complex cases.⁴⁻⁸ In addition, a previous study suggested that the accuracy of ChatGPT's differential diagnosis increased as more clinical context was provided and that its accuracy was not associated with the patient's age, gender and case acuity.⁴ Therefore, although the diagnostic performance of ChatGPT in cases with higher risks of diagnostic errors, including uncommon diseases or atypical presentations,⁹⁻¹³ is still unknown,^{4,7} ChatGPT may have the ability to assess diagnostic processes based on the high accuracy in clinical diagnosis.

ChatGPT may be useful for research about diagnostic errors and diagnostic excellence. ChatGPT can be used for the classification of clinical texts by specifying definitions or rules for classification. Some studies have attempted to use ChatGPT for specific classifications based on clinical case descriptions and found that although the quality and internal reliability of classification by ChatGPT were not yet optimal,¹⁴⁻¹⁶ it outperformed humans in terms of classification speed.¹⁵ Reviewing the clinical charts and records of patients using standardised assessment tools is one of the most commonly used methods in research and quality improvement actions regarding diagnostic errors; however, this practice requires considerable human resources, effort and time. This problem may be resolved if ChatGPT assists humans in reviewing clinical charts and records. Therefore, a pilot study is needed to evaluate the potential of ChatGPT in assessing the diagnostic process by reviewing texts describing cases. However, as entering patient information into ChatGPT is not acceptable because of concerns about personal information security, case reports are suitable data resources for such studies because they provide concise case presentations, include confirmed diagnoses with a high level of certainty and have few concerns regarding personal information security problems.

We previously conducted a systematic review of case reports containing diagnostic errors¹⁷; we collected data about the final diagnosis, commonality of diseases, typicality of presentation and contributing factors of diagnostic errors in each case. Using the database, we conducted this study to evaluate the performance of ChatGPT to assess the diagnostic process by reviewing case descriptions.

METHODS

Study design

This study used ChatGPT and case reports that contained diagnostic errors.

Target case reports and data used

The precise selection of case reports is described in our previous study.¹⁷ In brief, we searched PubMed using search terms related to diagnostic errors, namely 'diagnostic errors', 'delayed diagnosis', 'misdiagnosis' and 'wrong diagnosis'. We retrieved case reports with diagnostic errors that described only one patient and were published until 31 December 2021 from eight countries:

Australia, Canada, Germany, Italy, Japan, the Netherlands, the UK and the USA. A total of 563 case reports of diagnostic errors were obtained after two-stage screening as follows. In the first step, two reviewers screened the case reports by reading the titles and abstracts, and in the second step, 2 of 11 reviewers screened the case reports by reading the full texts. We excluded 18 case reports written in languages other than English to avoid the heterogeneity of outputs from ChatGPT due to language differences. In total, 545 case reports were included in this study. We extracted the following data from these case reports from a previous systematic review: the final diagnoses, commonality of the final diagnoses (common or uncommon), typicality of presentation (typical or atypical), most important codes and all codes of Diagnosis Error Evaluation and Research (DEER),¹⁸ Reliable Diagnosis Challenges (RDC)¹⁹ and Generic Diagnostic Pitfalls (GDP) taxonomies.²⁰ The detailed process for generating these data has been previously described.¹⁷

ChatGPT use

We used ChatGPT (GPT-4, 3 August 2023) between 15 and 23 August 2023. The prompts used in this study were developed by referencing those used in a previous study.⁶ We tested the prototype prompts using five case reports of diagnostic errors not included in this study and finalised them after editing some parts to improve the outputs. We used a total of five prompts in the same chat per case. The first part instructs ChatGPT how to classify cases into four categories based on disease commonality and typicality of presentation: typical presentation of a common disease, atypical presentation of a common disease, typical presentation of an uncommon disease and atypical presentation of an uncommon disease. The criteria for determining common or uncommon diseases and typical or atypical presentations are also included in this part. The second part describes the case. After inputting the prompt, ChatGPT outputs the classification of disease commonality and the typicality of presentation (1=typical presentation of common disease; 2=atypical presentation of common disease; 3=typical presentation of uncommon disease; and 4=atypical presentation of uncommon disease). The third part asks ChatGPT to judge whether diagnostic errors occurred in the presented case (1=diagnostic errors occurred; 0=no diagnostic errors occurred) based on the definition of diagnostic errors as 'the failure to (a) establish an accurate and timely explanation of the patient's health problem(s) or (b) communicate that explanation to the patient'. The fourth part asks ChatGPT to output all relevant codes of the DEER, RDC and GDP taxonomies, as well as the most important codes of these taxonomies specific to the presented case by displaying the respective taxonomy lists. The fifth part asked ChatGPT to summarise the code output in the seventh part. Examples of prompts and responses by ChatGPT are provided in the online supplemental file 1.

Outcomes

We assessed the rate of diagnostic errors determined by ChatGPT as the primary outcome. As the secondary outcomes, we assessed the distribution of the breakdowns of the DEER, RDC and GDP taxonomies coded by ChatGPT; and the rates of common diagnosis and typical presentation judged by ChatGPT, as well as the distribution of the classification (typical presentation of common disease, atypical presentation of common disease, typical presentation of uncommon disease and atypical presentation of uncommon disease).

Statistical analysis

Continuous and ordinal data were presented as medians with IQRs and compared using the Mann-Whitney U test. Categorical data are presented as percentages and were compared using the χ^2 test. We calculated the inter-rater agreement between ChatGPT and humans using Cohen's kappa statistics for all outcomes. A p value < 0.05 was considered significant. All statistical analyses were conducted using R V.4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Patient and public involvement

It was not appropriate to involve patients or the public in the design, or conduct, or reporting, or dissemination plans of our research.

RESULTS

ChatGPT's assessment of the commonality of the final diagnosis and typicality of presentation

ChatGPT assessed that the final diagnosis was common in 120/544 (22.1%) and the presentation was typical in 251/544 (46.1%) of the cases. Overall, ChatGPT classified six cases (1.1%) into the typical presentation of common disease, 114 (21.0%) into the atypical presentation of common disease, 245 (45.0%) into the typical presentation of uncommon disease and 179 (32.9%) into the atypical presentation of uncommon disease group. The inter-rater agreement of human and ChatGPT was 0.46 (0.36–0.56) for the commonality of the final diagnosis, 0.08 (0.00–0.16) for the typicality of presentation and 0.13 (0.07–0.20) for the classification by commonality and typicality.

ChatGPT's assessment of diagnostic errors and factors contributing to diagnostic errors

ChatGPT detected that diagnostic errors occurred in 519/545 cases (95.0%). The number of factors contributing to diagnostic errors per case coded by ChatGPT (median 5; IQR 6 and 16) was statistically higher than those coded by humans (median 1; IQR 3 and 11) for DEER ($p < 0.001$), RDC (median 4; IQR 6 and 18 by ChatGPT and median 2; IQR 4 and 9 by humans; $p < 0.001$) and GDP (median 4; IQR 5 and 10 by ChatGPT and median 1; IQR 2 and 5 by humans; $p < 0.001$). The most common breakdowns for DEER, RDC and GDP coded by ChatGPT were 'failure/delay in considering the diagnosis' (510,

93.6%), 'atypical presentation' (513, 94.1%) and 'atypical presentation' (539, 98.9%), which were same as the most common breakdowns coded by human (tables 1–3).

The most important DEER, RDC and GDP breakdowns coded by ChatGPT were 'failure/delay in considering the diagnosis' (315, 57.8%), 'atypical presentation' (365, 67.0%) and 'atypical presentation' (264, 48.4%), while by humans were 'failure/delay in considering the diagnosis' (234, 42.9%), 'findings masking/mimicking another diagnosis' (108, 19.8%) and 'limitations of a test or exam finding not appreciated' (160, 29.4%). The inter-rater agreements for the most important breakdown between ChatGPT and humans were 0.15 (0.09–0.21) for DEER, 0.04 (0.01–0.07) for RDC and 0.12 (0.07–0.17) for GDP.

DISCUSSION

We found that first, ChatGPT (GPT-4) correctly detected diagnostic errors in 95% of case reports by reading only the case description. Second, there was a large discrepancy between ChatGPT and physicians in assessing the commonality of final diagnosis and typicality of presentation. Third, ChatGPT raised more contributing factors to diagnostic errors using DEER, RDC and GDP taxonomies than physicians, and compared with physicians, weighed more on atypical presentation as the most important contributing factor for diagnostic errors.

This study indicates that ChatGPT can support research on diagnostic errors using published case reports. ChatGPT detected the presence of diagnostic errors in 95% of the case reports by reading only case descriptions, suggesting that ChatGPT its high sensitivity for detecting diagnostic errors. Owing to this, ChatGPT can be a useful tool for collecting case reports that include diagnostic errors, facilitating their further study. Furthermore, ChatGPT can also be used to screen diagnostic errors in clinical practice by imputing a written summary of care for a patient, which can facilitate an effective feedback system for improving the diagnostic process in each institution through the timely detection of possible cases of diagnostic errors. The construction of effective feedback systems for the diagnostic process is recommended to improve clinical diagnosis.^{21–24} Detection of possible cases of diagnostic errors with less effort is a fundamental requirement for the development of these feedback systems.^{22–25} Even with the current medical record systems, the use of trigger events or some calculated scores to detect a high-risk population for diagnostic errors (eg, a clinical visit followed several days later by an unplanned hospitalisation or subsequent visit to the emergency department, patients with discrepancies in diagnosis between admission and discharge) has been proposed to screen for possible cases of diagnostic errors effectively.^{25–28} However, these triggers may miss some patients with diagnostic errors (low sensitivity).²⁵ Moreover, manual reviewing of all cases is time-consuming and impractical. Therefore, ChatGPT can be used to screen for possible cases of diagnostic errors by entering only a

Table 1 Diagnostic Error Evaluation and Research taxonomy (total count)

Category	ChatGPT	Human	P value
Access/presentation			
(A) Failure/delay in presentation	6 (1.1%)	20 (3.7%)	0.005
(B) Failure/denied care access	3 (0.6%)	2 (0.4%)	0.65
History			
(A) Failure/delay in eliciting critical history data	327 (60.0%)	29 (5.3%)	<0.001
(B) Inaccurate/misinterpretation	193 (35.4%)	32 (5.9%)	<0.001
(C) Failure in weighing	21 (3.9%)	33 (6.1%)	0.09
(D) Failure/delay to follow-up	31 (5.7%)	5 (0.9%)	<0.001
Physical examination			
(A) Failure/delay in eliciting critical physical examination finding	102 (18.7%)	18 (3.3%)	<0.001
(B) Inaccurate/misinterpreted	24 (4.4%)	41 (7.5%)	0.03
(C) Failure in weighing	1 (0.2%)	21 (3.9%)	<0.001
(D) Failure/delay to follow-up	4 (0.7%)	6 (1.1%)	0.52
Tests (laboratory/radiology)			
(A) Failure/delay in ordering needed test(s)	330 (60.6%)	164 (30.1%)	<0.001
(B) Failure/delay in performing ordered test(s)	7 (1.3%)	3 (0.6%)	0.20
(C) Error in test sequencing	2 (0.4%)	1 (0.2%)	0.56
(D) Ordering of wrong test(s)	22 (4.0%)	1 (0.2%)	<0.001
(E) Test ordered the wrong way	2 (0.4%)	0 (0.0%)	0.16
(F) Sample mixup/mislabelled (eg, wrong patient/test)	0 (0.0%)	0 (0.0%)	N/A
(G) Technical errors/poor processing of specimen/test	4 (0.7%)	16 (2.9%)	0.01
(H) Erroneous laboratory/radiology reading of test	72 (13.2%)	88 (16.1%)	0.17
(I) Failed/delayed reporting of result to clinician	40 (7.3%)	2 (0.4%)	<0.001
(J) Failed/delayed follow-up of (abnormal) test result	189 (34.7%)	4 (0.7%)	<0.001
(K) Error in clinician interpretation of test	257 (47.2%)	77 (14.1%)	<0.001
Assessment			
(A) Failure/delay in considering the diagnosis	510 (93.6%)	357 (65.5%)	<0.001
(B) Too little consideration/weight given to the diagnosis	421 (77.2%)	57 (10.5%)	<0.001
(C) Too much weight on competing/coexisting diagnosis	36 (6.6%)	52 (9.5%)	0.08
(D) Failure/delay to recognise/weigh urgency	50 (9.2%)	23 (4.2%)	0.001
(E) Failure/delay to recognise/weigh complication(s)	9 (1.7%)	29 (5.3%)	<0.001
Referral/consultation			
(A) Failure/delay in ordering referral	79 (14.5%)	40 (7.3%)	<0.001
(B) Failure/delay obtaining/scheduling ordered referral	13 (2.4%)	1 (0.2%)	0.001
(C) Error in diagnostic consultation performance	8 (1.5%)	6 (1.1%)	0.59
(D) Failure/delayed communication/follow-up of consultation	35 (6.4%)	6 (1.1%)	<0.001
Follow-up			
(A) Failure to refer patient to close/safe setting/monitoring	17 (3.1%)	7 (1.3%)	0.04
(B) Failure/delay in timely follow-up/rechecking of patient	101 (18.5%)	14 (2.6%)	<0.001
Unclear	0 (0.0%)	3 (0.6%)	0.08

summary of cases to aid the detection of diagnostic errors in daily clinical practice. Therefore, this study proposes a new method to implement AI to improve diagnosis.

The use of ChatGPT for research or quality improvement for diagnostic safety has some issues. Accurate assessment of the diagnostic process and detection of the cause of diagnostic errors are vital in research and quality improvement actions for diagnostic safety. In this

study, there were large discrepancies between ChatGPT and physicians in assessing the commonality of disease, typicality of presentation and taxonomies of contributing factors to diagnostic errors, such as DEER, RDC and GDP. In particular, ChatGPT tended to judge more cases as atypical presentations and code more contributing factors per case than physicians. These results indicate that ChatGPT may consider normal variances as atypical

Table 2 Reliable Diagnosis Challenges taxonomy (total count)

Category	GPT	Human	P value
Challenging disease presentation			
Atypical presentation	513 (94.1%)	211 (38.7%)	<0.001
Non-specific symptoms and signs	82 (15.0%)	69 (12.7%)	0.25
Unfamiliar/outside specialty	161 (29.5%)	61 (11.2%)	<0.001
Findings masking/mimicking another diagnosis	383 (70.3%)	210 (38.5%)	<0.001
Red herring misleading findings	83 (15.2%)	91 (16.7%)	0.51
Rapidly progressive course	6 (1.1%)	21 (3.9%)	0.003
Slowly evolving blunting onset perception	54 (9.9%)	39 (7.2%)	0.10
Deceptively benign course	24 (4.4%)	8 (1.5%)	0.004
Patient factors			
Language/communication barriers	9 (1.7%)	14 (2.6%)	0.29
Signal: noise - patients with multiple other symptoms or diagnoses	24 (4.4%)	23 (4.2%)	0.88
Failure to share data (to be forthcoming with symptoms or their severity)	79 (14.5%)	10 (1.8%)	<0.001
Failure to follow-up	64 (11.7%)	4 (0.7%)	<0.001
Testing challenges			
Test not available due to geography, access, cost	58 (10.6%)	5 (0.9%)	<0.001
Logistical issues in scheduling, performing	31 (5.7%)	2 (0.4%)	<0.001
False positive/negative test limitations	54 (9.9%)	38 (7.0%)	0.08
Performance/interpretation failures	444 (81.5%)	145 (26.6%)	<0.001
Equivocal results/interpretation	99 (18.2%)	44 (8.1%)	<0.001
Test follow-up issues (eg, tracking pending results)	39 (7.2%)	5 (0.9%)	<0.001
Stressors			
Time constraints for clinicians and patients	32 (5.9%)	2 (0.4%)	<0.001
Discontinuities of care	137 (25.1%)	1 (0.2%)	<0.001
Fragmentation of care	36 (6.6%)	4 (0.7%)	<0.001
Memory reliance/challenges	20 (3.7%)	3 (0.6%)	<0.001
Broader challenges			
Recognition of acuity/severity	91 (16.7%)	43 (7.9%)	<0.001
Diagnosis of complications	16 (2.9%)	35 (6.4%)	0.01
Recognition of failure to respond to therapy	23 (4.2%)	28 (5.1%)	0.47
Diagnosis of the underlying aetiological cause	44 (8.1%)	149 (27.3%)	<0.001
Recognising misdiagnosis occurrence	77 (14.1%)	85 (15.6%)	0.50
Unclear	0 (0.0%)	3 (0.6%)	0.08

and non-significant variances of the diagnostic process as contributing factors to diagnostic errors. A previous study assessing the performance of fracture classification using ChatGPT also showed that although ChatGPT classified significantly faster than humans, its classification performance was inferior.¹⁵ In another study, ChatGPT evaluated patient neuro-examination descriptions using well-established neurological assessment scales; however, its accuracy was reduced when confronted with incomplete or vague descriptions.¹⁶ In addition, using GPT3.5 model, a previous study indicated that the performance of ChatGPT in processing medical text classification tasks with few samples may still be far from optimal.¹⁴ Therefore, well-formatted case descriptions and fine-tuning of

prompts with sufficient samples are needed to enhance the ability of ChatGPT to assess the details of the diagnostic process and the cause of diagnostic errors. Nevertheless, the high sensitivity of ChatGPT to detect atypical presentations and the contributing factors of diagnostic errors make it suitable for screening before manual assessment.

This study's results about the high sensitivity of ChatGPT to detect diagnostic errors should be interpreted with caution in several aspects. First, sensitivity alone does not assure the performance of a tool to determine the presence or absence of a target outcome. Because this study used case reports exclusive to diagnostic errors, there was a selection bias in the used cases and we could not

Table 3 General Diagnostic Pitfall taxonomy (total count)

Category	GPT	Human	P value
(1) Failure to follow-up	183 (33.6%)	21 (3.9%)	<0.001
(2) Limitations of a test or examination finding not appreciated	424 (77.8%)	200 (36.7%)	<0.001
(3) Disease A repeatedly mistaken for disease B	34 (6.2%)	141 (25.9%)	<0.001
(4) Risk factors not adequately appreciated	15 (2.8%)	42 (7.7%)	<0.001
(5) Atypical presentation	539 (98.9%)	210 (38.5%)	<0.001
(6) Counter-diagnosis cues overlooked (eg, red flags)	299 (54.9%)	111 (20.4%)	<0.001
(7) Communication failures between primary care physician and specialist	32 (5.9%)	4 (0.7%)	<0.001
(8) Issues surrounding referral	68 (12.5%)	16 (2.9%)	<0.001
(9) Urgency not fully appreciated	347 (63.7%)	31 (5.7%)	<0.001
(10) Chronic disease presumed to account for new symptoms	117 (21.5%)	24 (4.4%)	<0.001
(11) Miscommunication related to laboratory ordering	41 (7.5%)	1 (0.2%)	<0.001
(12) Evolving symptoms not monitored	48 (8.8%)	8 (1.5%)	<0.001
Unclear	0 (0.0%)	19 (3.5%)	<0.001

also assess the specificity of ChatGPT to detect diagnostic errors. One might assume that the high sensitivity of ChatGPT in this study may reflect on the very low specificity (ie, ChatGPT may judge almost all cases as diagnostic errors). Therefore, future studies are needed to assess both the sensitivity and specificity of ChatGPT to detect diagnostic errors. However, case reports do not seem suitable resources for such kinds of studies. Case reports focused on diagnosis can be published only when they include some teaching points related to the diagnostic process; therefore, finding case reports free from diagnostic errors may be difficult. Second, published case reports usually include mostly relevant information without noise. Noise refers to any irrelevant or misleading data that diminishes the clarity of the clinical signal. Due to the increased noise in clinical charts from the real world, ChatGPT's proficiency to detect and classify diagnostic errors can decline in the real world. Considering these two issues, in the next steps, ChatGPT's performance in detecting and classifying diagnostic errors should be evaluated using 'live' medical records with or without diagnostic errors, compared with the judgement of human expert reviewers. Nevertheless, integrating ChatGPT into the real world to detect diagnostic errors may have another challenge: a lack of context-specific knowledge. Because contextual information such as available diagnostic resources (eg, human, equipment, time, cost) and patient perspectives are typically gone unrecorded in medical records. Consequently, ChatGPT may rely on the ideal diagnostic process and outcome as the reference standard to judge the presence or absence of diagnostic errors, potentially leading to overly sensitive error detection. Indeed, in this study, items related to fundamentally 'human' tasks in the diagnostic process, such as history taking and physical examinations, were more frequently coded as contributing factors to diagnostic errors by ChatGPT than by 'human' physician researchers. We assume that this result derives from the

fact that human researchers may judge the issues related to history-taking and physical examination case by case by moving the thresholds considering background contexts (eg, settings, time restriction) and clinical relevance to diagnostic errors. In contrast, ChatGPT may judge based only on 'texts' in the prompts that are not flexible, and the issues judged as contributing factors to diagnostic errors cannot be clinically relevant in some cases. To address this issue, similar to the approach suggested for human expert reviews of potential diagnostic errors,^{22 29} developing the practical method to input contextual knowledge into ChatGPT is crucial. This strategy would help mitigate the risk of excessively sensitive error judgements by ChatGPT. Until the method is developed, human experts' adjustment with contextual knowledge to the judgement by ChatGPT remains necessary. Another solution may be adding more detailed explanations to taxonomies such as DEER, RDC and GDP to tell ChatGPT what types of issues are clinically relevant to diagnostic errors in the real world.

This study had several limitations. First, we used prompts only once for each case; imputing the same prompts another time could have produced different results. However, humans may also produce different outputs when performing the same tasks iteratively. Therefore, this limitation does not reduce the value of this study; rather, it indicates that a double check by ChatGPT or humans is needed to validate ChatGPT responses on diagnostic errors and their contributing factors. Second, ChatGPT assessed the presence or absence of diagnostic errors and their contributing factors based only on case descriptions and definitions of diagnostic errors, commonality of disease, typicality of presentation and the three taxonomies. In contrast, physician researchers judge the presence or absence of diagnostic errors and their contributing factors based on the entire case report, including the abstract, introduction, discussion and conclusions. The difference between ChatGPT and

human assessments can be attributed to this. Considering this discrepancy, the ability of ChatGPT to correctly detect 96% of diagnostic errors and identify additional contributing factors for diagnostic errors based only on case descriptions its utility as a screening tool. Third, it is unclear how many case reports included in this study were used for training of ChatGPT. Therefore, ChatGPT could have generated the correct diagnosis for the cases that were part of its pre-training ChatGPT. Fourth, although GPT-4 is a multimodal AI model that is inherently capable of inputting images and tables, this study excluded them. Therefore, in cases where images and tables provided key information in the diagnosis, the quality of ChatGPT could have been low.

CONCLUSIONS

ChatGPT can be a useful tool to screen possible cases with diagnostic errors and shortlist factors contributing to diagnostic errors in researching diagnostic errors using case reports. However, owing to the limitation of ChatGPT, such as judging normal variances as abnormal, its responses must be validated by a physician.

Author affiliations

¹Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Shimotsuga-gun, Tochigi, Japan

²Urasoe General Hospital, Urasoe, Okinawa, Japan

³Nerima Hikarigaoka Hospital, Nerima-ku, Tokyo, Japan

⁴Yokohama City University School of Medicine Graduate School of Medicine, Yokohama, Kanagawa, Japan

⁵Department of General Medicine, Faculty of Medicine, Juntendo University, Bunkyo-ku, Tokyo, Japan

⁶NHO Kumamoto Medical Center, Kumamoto, Kumamoto, Japan

⁷Department of General Internal Medicine, Tenri Hospital, Tenri, Nara, Japan

⁸Integrated Clinical Education Center, Kyoto University Hospital, Kyoto, Kyoto, Japan

X Takashi Watari @wataritari1

Contributors YH conceptualised this study. All authors contributed to data collection. YH drafted the manuscript, and all authors contributed to revision. YH acted as the guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The data sets used in the current study will be made available from the corresponding author upon request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Yukinori Harada <http://orcid.org/0000-0001-6042-7397>

Taiju Miyagami <http://orcid.org/0000-0002-4893-2224>

Ren Kawamura <http://orcid.org/0000-0002-5632-3218>

Taro Shimizu <http://orcid.org/0000-0002-3788-487X>

Takashi Watari <http://orcid.org/0000-0002-9322-8455>

REFERENCES

- 1 Takagi S, Watari T, Erabi A, *et al*. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023;9:e48002.
- 2 Kung TH, Cheatham M, Medenilla A, *et al*. Performance of Chatgpt on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.
- 3 Liu J, Wang C, Liu S. Utility of Chatgpt in clinical practice. *J Med Internet Res* 2023;25:e48568.
- 4 Rao A, Pang M, Kim J, *et al*. Assessing the utility of Chatgpt throughout the entire clinical Workflow: development and usability study. *J Med Internet Res* 2023;25:e48659.
- 5 Hirose T, Harada Y, Yokose M, *et al*. Diagnostic accuracy of differential-diagnosis lists generated by Generative Pretrained transformer 3 Chatbot for clinical vignettes with common chief complaints: A pilot study. *Int J Environ Res Public Health* 2023;20:3378.
- 6 Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023;330:78–80.
- 7 Berg HT, van Bakel B, van de Wouw L, *et al*. Chatgpt and generating a differential diagnosis early in an emergency Department presentation. *Ann Emerg Med* 2024;83:83–6.
- 8 Shea Y-F, Lee CMY, Ip WCT, *et al*. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. *JAMA Netw Open* 2023;6:e2325000.
- 9 Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care—A systematic review. *Fam Pract* 2008;25:400–13.
- 10 Newman-Toker DE, Peterson SM, Badihian S, *et al*. Diagnostic errors in the emergency Department: A systematic review. *Agency for Healthcare Research and Quality (AHRQ)* 2022.
- 11 Matulis JC, Kok SN, Dankbar EC, *et al*. A survey of outpatient internal medicine clinician perceptions of diagnostic error. *Diagnosis (Berl)* 2020;7:107–14.
- 12 Goyder CR, Jones CHD, Heneghan CJ, *et al*. Missed opportunities for diagnosis: lessons learned from diagnostic errors in primary care. *Br J Gen Pract* 2015;65:e838–44.
- 13 Newman-Toker DE, Wang Z, Zhu Y, *et al*. Rate of diagnostic errors and serious Misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the “big three” *Diagnosis* 2021;8:67–84.
- 14 Chen Q, Sun H, Liu H, *et al*. An extensive benchmark study on biomedical text generation and mining with Chatgpt. *Bioinformatics* 2023;39:btad557.
- 15 Russe MF, Fink A, Ngo H, *et al*. Performance of Chatgpt, human Radiologists, and context-aware Chatgpt in identifying AO codes from Radiology reports. *Sci Rep* 2023;13:14215.
- 16 Chen TC, Kaminski E, Koduri L, *et al*. Chat GPT as a neuro-score Calculator: analysis of a large language model's performance on various neurological exam grading scales. *World Neurosurg* 2023;179:e342–7.
- 17 Harada Y, Watari T, Nagano H, *et al*. Diagnostic errors in uncommon conditions: A systematic review of case reports of diagnostic errors. *Diagnosis (Berl)* 2023;10:329–36.
- 18 Schiff GD, Hasan O, Kim S, *et al*. Diagnostic error in medicine: analysis of 583 physician-reported errors. *Arch Intern Med* 2009;169:1881–7.
- 19 Schiff GD. Finding and fixing diagnosis errors: can triggers help *BMJ Qual Saf* 2012;21:89–92.
- 20 Schiff GD, Volodarskaya M, Ruan E, *et al*. Characteristics of disease-specific and generic diagnostic pitfalls: A qualitative study. *JAMA Netw Open* 2022;5:e2144531.



- 21 Giardina TD, Shahid U, Mushtaq U, *et al.* Creating a learning health system for improving diagnostic safety: pragmatic insights from US health care organizations. *J Gen Intern Med* 2022;37:3965–72.
- 22 Fernandez Branson C, Williams M, Chan TM, *et al.* Improving diagnostic performance through feedback: the diagnosis learning cycle. *BMJ Qual Saf* 2021;30:1002–9.
- 23 Lane KP, Chia C, Lessing JN, *et al.* Improving resident feedback on diagnostic reasoning after Handovers: the LOOP project. *J Hosp Med* 2019;14:622–5.
- 24 Meyer AND, Singh H. The path to diagnostic excellence includes feedback to Calibrate how Clinicians think. *JAMA* 2019;321:737–8.
- 25 Singh H, Bradford A, Goeschel C. Operational measurement of diagnostic safety: state of the science. *Diagnosis (Berl)* 2021;8:51–65.
- 26 Mahajan P, Pai C-W, Cosby KS, *et al.* Identifying trigger concepts to screen emergency Department visits for diagnostic errors. *Diagnosis (Berl)* 2021;8:340–6.
- 27 Perry MF, Melvin JE, Kasick RT, *et al.* The diagnostic error index: A quality improvement initiative to identify and measure diagnostic errors. *J Pediatr* 2021;232:257–63.
- 28 Murphy DR, Meyer AN, Sittig DF, *et al.* Application of electronic trigger tools to identify targets for improving diagnostic safety. *BMJ Qual Saf* 2019;28:151–9.
- 29 Bradford A, Shahid U, Schiff GD, *et al.* Development and usability testing of the agency for Healthcare research and quality common formats to capture diagnostic safety events. *J Patient Saf* 2022;18:521–5.