Cureus
Part of SPRINGER NATURE

Open Access Original Article

Diagnostic Performance of GPT-4o and Claude 3 Opus in Determining Causes of Death From Medical Histories and Postmortem CT Findings

Review began 07/26/2024
Review ended 08/07/2024
Published 08/20/2024

DOI: 10.7759/cureus.67306

Masanori Ishida [1], Wataru Gonoi [2], Keisuke Nyunoya [2], Hiroyuki Abe [3], Go Shirota [2], Naomasa Okimoto [2], Kotaro Fujimoto [2], Mariko Kurokawa [2], Motoki Nakai [1], Kazuhiro Saito [1], Tetsuo Ushiku [3], Osamu Abe [2]

1. Department of Radiology, Tokyo Medical University Hospital, Tokyo, JPN 2. Department of Radiology, The University of Tokyo Hospital, Tokyo, JPN 3. Department of Pathology, The University of Tokyo Hospital, Tokyo, JPN

Corresponding author: Masanori Ishida, mytsy1007@gmail.com

Abstract

Introduction: This study evaluates the diagnostic performance of the latest large language models (LLMs), GPT-4o (OpenAI, San Francisco, CA, USA) and Claude 3 Opus (Anthropic, San Francisco, CA, USA), in determining causes of death from medical histories and postmortem CT findings.

Methods: We included 100 adult cases whose postmortem CT scans were diagnosable for the causes of death using the gold standard of autopsy results. Their medical histories and postmortem CT findings were compiled, and clinical and imaging diagnoses of both the underlying and immediate causes of death, as well as their personal information, were carefully separated from the database to be shown to the LLMs. Both GPT-4o and Claude 3 Opus generated the top three differential diagnoses for each of the underlying or immediate causes of death based on the following three prompts: 1) medical history only; 2) postmortem CT findings only; and 3) both medical history and postmortem CT findings. The diagnostic performance of the LLMs was compared using McNemar's test.

Results: For the underlying cause of death, GPT-4o achieved primary diagnostic accuracy rates of 78%, 72%, and 78%, while Claude 3 Opus achieved 72%, 56%, and 75% for prompts 1, 2, and 3, respectively. Including any of the top three differential diagnoses, GPT-4o's accuracy rates were 92%, 90%, and 92%, while Claude 3 Opus's rates were 93%, 69%, and 93% for prompts 1, 2, and 3, respectively. For the immediate cause of death, GPT-4o's primary diagnostic accuracy rates were 55%, 58%, and 62%, while Claude 3 Opus's rates were 60%, 62%, and 63% for prompts 1,2, and 3, respectively. For any of the top three differential diagnoses, GPT-4o's accuracy rates were 88% for prompt 1 and 91% for prompts 2 and 3, whereas Claude 3 Opus's rates were 92% for all three prompts. Significant differences between the models were observed for prompt two in diagnosing the underlying cause of death (p = 0.03 and <0.01 for the primary and top three differential diagnoses, respectively).

Conclusion: Both GPT-4o and Claude 3 Opus demonstrated relatively high performance in diagnosing both the underlying and immediate causes of death using medical histories and postmortem CT findings.

Categories: Pathology, Forensic Medicine, Radiology
Keywords: cause of death, postmortem computed tomography, large language model, gpt-4o, claude 3 opus, artificial intelligence

Introduction

Large language models (LLMs) are neural network models trained on vast amounts of text data, demonstrating their high performance in natural language processing tasks [1]. They can be utilized across various fields, including medicine [2]. Regarding the diagnostic capabilities of LLMs in radiology, it has been reported that these models can provide a list of highly probable differential diagnoses and their respective accuracies based solely on textual information from clinical histories and imaging findings [3-5]. Thus, the potential use of language artificial intelligence (AI) in diagnostic imaging has been revealed. Recently, various manufacturers have developed their own LLMs, which have undergone multiple version upgrades. In the area of death investigation, the diagnostic abilities of AI models using radiological images have not yet been investigated. With the shortage of forensic doctors, pathologists, and radiologists specializing in death investigation, the clinical implementation of AI-based diagnostic aids is expected, and studies of deep learning models using postmortem CT images have been reported [6]. Therefore, in this study, we employed the latest versions of GPT-4o (OpenAI, San Francisco, CA, USA) [7] and Claude 3 Opus (Anthropic, San Francisco, CA, USA) [8] to assess their diagnostic performance in determining the cause of death using medical histories and imaging findings.

Materials And Methods

This case-control study was approved by the Institutional Review Board of the University of Tokyo Hospital (approval no. 2076). Written informed consent was obtained from the families of the deceased. All procedures conformed to the ethical principles outlined in the Declaration of Helsinki. In this study, we included 100 cases of adult in-hospital natural deaths that occurred between May 2014 and June 2019, whose postmortem CT was diagnosable for the causes of death with the gold standard of autopsy results.

The compiled medical histories and postmortem CT findings of all adult cases were blinded to the autopsy results. Postmortem CT reports were written independently by two board-certified diagnostic radiologists at our tertiary referral center. In the interpretation, all organs and structures of the postmortem CT scan (the average time elapsed from death to postmortem CT is six hours and 40 minutes) from the head to the lower extremities were systematically evaluated. When antemortem CT images were available, they were referred to while interpreting the postmortem CT images and writing the report. Postmortem CT reports written in Japanese were double-checked and finalized by consensus, then collected and used for input. Medical histories were prepared by the attending physician for postmortem CT and pathological autopsy. The medical histories were prepared in a manner similar to writing discharge summaries. Medical histories include present illness at admission, past history (pre-existing or pathologically diagnosed diseases), social life history (including tobacco use, alcohol consumption, occupation, etc.), and family history. Detailed hematological, pathological, biochemical, and radiological reports were not included in the medical histories.

An overview of this study is presented in Figure *1*. We used GPT-4o (OpenAI), released on May 13, 2024, and Claude 3 Opus (Anthropic), released on March 4, 2024, to list a primary diagnosis and two differential diagnoses of the underlying and immediate cause of death, respectively. When showing the data (text only) to the LLMs, medical and imaging diagnoses of the underlying and immediate causes of death, as well as personal information, were carefully separated from the database. In defining causes of death, the following two distinct levels can be distinguished: the underlying cause of death (the primary underlying disease that initiates the events leading to death) and the immediate cause of death (the final disease or condition resulting in death) [9,10]. Because all medical histories and postmortem CT reports were in Japanese, textual data without a fixed format was passed to the models without translation. The LLMs were asked to present two differential diagnoses of the cause of death in addition to the primary diagnosis in order of likelihood (Figure *2*).

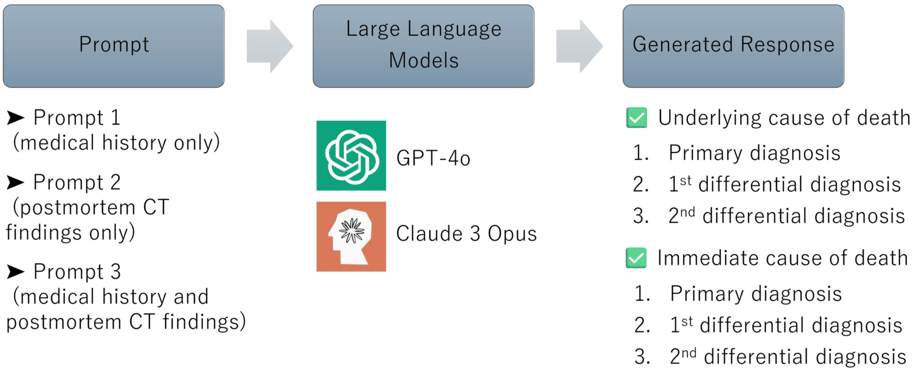
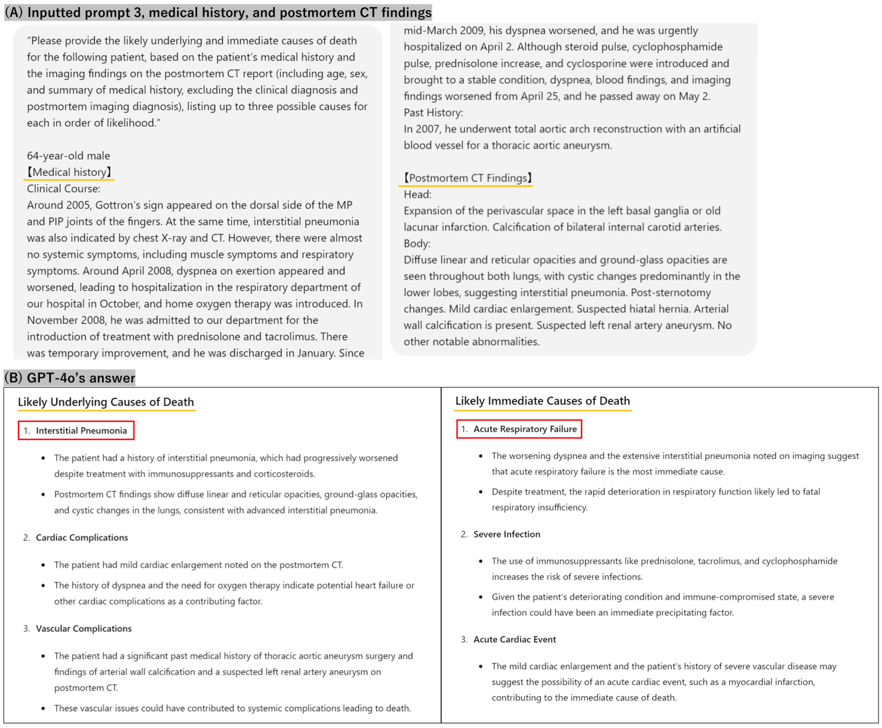

Prompt
Large Language Models
Generated Response
➤ Prompt 1 (medical history only)
➤ Prompt 2 (postmortem CT findings only)
➤ Prompt 3 (medical history and postmortem CT findings)
GPT-4o
Claude 3 Opus
✓ Underlying cause of death
1. Primary diagnosis
2. 1st differential diagnosis
3. 2nd differential diagnosis
✓ Immediate cause of death
1. Primary diagnosis
2. 1st differential diagnosis
3. 2nd differential diagnosis

FIGURE 1: Overview of the study



(A) Inputted prompt 3, medical history, and postmortem CT findings

"Please provide the likely underlying and immediate causes of death for the following patient, based on the patient's medical history and the imaging findings on the postmortem CT report (including age, sex, and summary of medical history, excluding the clinical diagnosis and postmortem imaging diagnosis), listing up to three possible causes for each in order of likelihood."

64-year-old male
【Medical history】
Clinical Course:
Around 2005, Gottron's sign appeared on the dorsal side of the MP and PIP joints of the fingers. At the same time, interstitial pneumonia was also indicated by chest X-ray and CT. However, there were almost no systemic symptoms, including muscle symptoms and respiratory symptoms. Around April 2008, dyspnea on exertion appeared and worsened, leading to hospitalization in the respiratory department of our hospital in October, and home oxygen therapy was introduced. In November 2008, he was admitted to our department for the introduction of treatment with prednisolone and tacrolimus. There was temporary improvement, and he was discharged in January. Since mid-March 2009, his dyspnea worsened, and he was urgently hospitalized on April 2. Although steroid pulse, cyclophosphamide pulse, prednisolone increase, and cyclosporine were introduced and brought to a stable condition, dyspnea, blood findings, and imaging findings worsened from April 25, and he passed away on May 2.
Past History:
In 2007, he underwent total aortic arch reconstruction with an artificial blood vessel for a thoracic aortic aneurysm.

【Postmortem CT Findings】
Head:
Expansion of the perivascular space in the left basal ganglia or old lacunar infarction. Calcification of bilateral internal carotid arteries.
Body:
Diffuse linear and reticular opacities and ground-glass opacities are seen throughout both lungs, with cystic changes predominantly in the lower lobes, suggesting interstitial pneumonia. Post-sternotomy changes. Mild cardiac enlargement. Suspected hiatal hernia. Arterial wall calcification is present. Suspected left renal artery aneurysm. No other notable abnormalities.

(B) GPT-4o's answer

Likely Underlying Causes of Death

1. **Interstitial Pneumonia**
   - The patient had a history of interstitial pneumonia, which had progressively worsened despite treatment with immunosuppressants and corticosteroids.
   - Postmortem CT findings show diffuse linear and reticular opacities, ground-glass opacities, and cystic changes in the lungs, consistent with advanced interstitial pneumonia.
2. **Cardiac Complications**
   - The patient had mild cardiac enlargement noted on the postmortem CT.
   - The history of dyspnea and the need for oxygen therapy indicate potential heart failure or other cardiac complications as a contributing factor.
3. **Vascular Complications**
   - The patient had a significant past medical history of thoracic aortic aneurysm surgery and findings of arterial wall calcification and a suspected left renal artery aneurysm on postmortem CT.
   - These vascular issues could have contributed to systemic complications leading to death.

Likely Immediate Causes of Death

1. **Acute Respiratory Failure**
   - The worsening dyspnea and the extensive interstitial pneumonia noted on imaging suggest that acute respiratory failure is the most immediate cause.
   - Despite treatment, the rapid deterioration in respiratory function likely led to fatal respiratory insufficiency.
2. **Severe Infection**
   - The use of immunosuppressants like prednisolone, tacrolimus, and cyclophosphamide increases the risk of severe infections.
   - Given the patient's deteriorating condition and immune-compromised state, a severe infection could have been an immediate precipitating factor.
3. **Acute Cardiac Event**
   - The mild cardiac enlargement and the patient's history of severe vascular disease may suggest the possibility of an acute cardiac event, such as a myocardial infarction, contributing to the immediate cause of death.

FIGURE 2: Question for prompt 3 and the response samples

(A) Manually typed question for prompt three, medical history, and postmortem CT findings; (B) Response generated by GPT-4o

The language used in the study is Japanese, but this example is shown in English. In this case, the underlying cause of death was interstitial pneumonia, and the immediate cause of death was respiratory failure, consistent with diagnoses from autopsy and CT.

The three prompts

Analyses were conducted under the following three prompts after inputting the definitions of the underlying and immediate causes of death above. Prompt 1: "Please provide the likely underlying and immediate causes of death for the following patient, based on the patient's medical history (including age, sex, and summary of medical history, excluding the clinical diagnosis), listing up to three possible causes for each in order of increasing likelihood." Prompt 2: "Please provide the likely underlying and immediate causes of death for the following patient, based on the imaging findings on the postmortem CT report (including age and sex but excluding the summary of the patient's medical history, clinical diagnosis, and postmortem imaging diagnosis), listing up to three possible causes for each in order of likelihood." Prompt 3: "Please provide the likely underlying and immediate causes of death for the following patient, based on the patient's medical history and the imaging findings on the postmortem CT report (including age, sex, and summary of medical history, excluding the clinical diagnosis and postmortem imaging diagnosis), listing up to three possible causes for each in order of likelihood."

Each condition was examined using a separate thread in an independent session to prevent previous inputs from influencing subsequent ones. The underlying and immediate causes of death generated by LLMs were determined to be consistent with the postmortem CT and autopsy diagnoses. The accuracy of the LLMs' primary diagnosis and two differential diagnoses each for the underlying and immediate causes of death were determined by a consensus of two board-certified diagnostic radiologists with 17 to 18 years of experience in postmortem imaging. When there was a difference in expression or choice of words between the pathologist's diagnosis by autopsy and the AI's diagnosis by LLMs, it was considered accurate if it was determined to be synonymous. Ambiguous answers or those lacking sufficient elements were considered incorrect.

McNemar's test was used to assess differences in correct response rates for the primary and three major differential diagnoses under prompts 1, 2, and 3. We also examined whether there was a difference between the two LLMs in the number of cases in which the top three differential diagnoses could not be listed by either prompt. Two-sided p-values <0.05 were considered statistically significant. Statistical analyses were performed using SPSS Statistics version 29.0 (IBM Corp., Armonk, NY, USA).

Results

The classification of diseases for the underlying and immediate causes of death is shown in Table 1 and Table 2, respectively. The diagnostic accuracy of each model is summarized in Table 3. For the underlying cause of death, GPT-4o achieved primary diagnostic accuracy rates of 78%, 72%, and 78% for prompts 1, 2, and 3, respectively. Claude 3 Opus achieved 72%, 56%, and 75%, respectively. When considering any of the top three differential diagnoses, GPT-4o's accuracy rates were 92%, 90%, and 92%, while Claude 3 Opus's rates were 93%, 69%, and 93% for prompts 1, 2, and 3, respectively. For the immediate cause of death, GPT-4o's primary diagnostic accuracy rates were 55%, 58%, and 62%, compared with Claude 3 Opus's 60%, 62%, and 63% for prompts 1, 2, and 3, respectively. For any of the top three differential diagnoses, GPT-4o's accuracy rates were 88% for prompt 1 and 91% for prompts 2 and 3, whereas Claude 3 Opus's were 92% for all three prompts.

Type of pathology	No. of cases (n = 100)
Solid malignancy	43
Cardiovascular	27
Hematological malignancy	18
Respiratory disease	5
Autoimmune disease	2
Liver disease	2
Infectious disease	1
Metabolic disorder	1
Neurologic disease	1

TABLE 1: Underlying causes of death by type of pathology

Type of pathology	No. of cases (n = 100)
Respiratory failure	43
Heart failure	16
Multiorgan failure	16
Intracranial hemorrhage	5
Cardiac tamponade	4
Liver failure	4
Circulatory failure	3
Hemorrhagic shock	3
Brain stem metastasis	2
Septic shock	2
Cerebral infarction	1
Hypoxic encephalopathy	1

TABLE 2: Immediate causes of death by type of pathology

Causes	Accuracy (%)		McNemar's test (p-value)
	GPT-4o	Claude 3 Opus	GPT-4o vs. Claude 3 Opus
Underlying cause of death			
Primary diagnosis by prompts 1, 2, and 3	78%, 72%, 78%	72%, 56%, 75%	0.21/*0.03/0.549
Top three differential diagnoses by prompts 1, 2, and 3	92%, 90%, 92%	93%, 69%, 93%	1.000/*<0.01/1.000
Immediate cause of death			
Primary diagnosis by prompts 1, 2, and 3	55%, 58%, 62%	60%, 62%, 63%	0.441/0.617/1.000
Top three differential diagnoses by prompts 1, 2, and 3	88%, 91%, 91%	92%, 92%, 92%	0.344/1.000/1.000

TABLE 3: Diagnostic accuracy of large language models

* A p-value of <0.05 was considered statistically significant.

McNemar's test demonstrated significant differences in determining the underlying cause of death between the two models for prompt 2 (p = 0.03 and <0.01 for determining the primary diagnosis and the top three diagnoses, respectively). No significant differences in determining the underlying cause of death between the models were observed for prompts 1 and 3. Additionally, no significant differences in determining the immediate cause of death between the models were observed for prompts 1, 2, and 3. The results are summarized in Table 3. There was also no significant difference between the two LLMs in the number of cases in which the top three differential diagnoses could not be listed by either prompt (p = 0.625 and <0.25 for determining underlying and immediate causes of death, respectively (Table 4)). The underlying causes of death for which GPT-4o was not listed are as follows: two cases of dilated cardiomyopathy, one case of dermatomyositis, one case of liver cirrhosis, one case of hepatocellular carcinoma, and one case of pulmonary tuberculosis. Of these, the same two cases of dilated cardiomyopathy and one case of pulmonary tuberculosis were also not listed by Claude 3 Opus as the underlying cause of death. In addition, one case of esophageal tumor was not listed as an underlying cause of death by Claude 3 Opus. The immediate causes of death for which GPT-4o was not listed are as follows: two cases of multiple organ failure and one case of peritonitis.

	GPT-4o vs. Claude 3 Opus	McNemar's test (p-value)
Underlying cause of death	6/100 vs. 4/100	0.625
Immediate cause of death	3/100 vs. 0/100	0.25

TABLE 4: Comparison of the number of cases in which the top three differential diagnoses are not listed by any of the prompts between the two models

Discussion

To the best of our knowledge, this is the first study to use generative AI to make cause-of-death diagnoses. In this study, we evaluated the diagnostic performance of the language AI models, GPT-4o and Claude 3 Opus, in determining the cause of death based on three types of input information: medical history only, postmortem CT findings only, and medical history with postmortem CT findings. The results showed that GPT-4o and Claude 3 Opus demonstrated relatively high performance in diagnosing both the underlying and immediate causes of death by inputting postmortem CT findings in addition to the medical histories. It could be said that the usefulness of AI in cause-of-death determination has been demonstrated, potentially reflecting the diagnostic performance of AI models.

In determining the top three differential diagnoses, both underlying and immediate causes of death were estimated with levels of accuracy of around 90% for all prompts, except for the method of diagnosing the underlying cause of death in Claude 3 Opus with prompt 2. It can be said that GPT-4o can diagnose with a high degree of accuracy for differential diagnosis, but Claude 3 Opus cannot achieve the same level.

However, the accuracy of the primary diagnosis of the immediate cause of death for both GPT-4o and Claude 3 Opus is moderate (about 50% to 60%). Among the prompts, the accuracy was higher when both history and postmortem CT findings were inputted and slightly higher when only postmortem CT findings were inputted compared to when only medical history was inputted. This suggests that when a generative AI diagnoses the cause of death, having both history and CT findings is generally more useful than having just one. The additional information provided in this study was also helpful.

The diagnostic accuracy rates of GPT-4o and Claude 3 Opus were similar, except for prompt 2, in diagnosing the underlying cause of death. The difference in prompt 2's results might be because not all of the relevant medical history is reflected in a report of a single imaging study, and the patient's clinical course is crucial in understanding the cause of death. For prompt 2, in diagnosing the underlying cause of death, Claude 3 Opus demonstrated significantly inferior diagnostic performance compared with GPT-4o when given only the postmortem CT findings as input, without the textual information of medical histories. It is difficult to determine the reasons for the differences in diagnostic performance between the models observed in this study. In the estimation of underlying and immediate causes of death, prompt 3 for both GPT-4o and Claude 3 Opus had similar or higher diagnostic accuracy than prompts 1 and 2. This suggests that both medical history and postmortem CT findings would be better for cause-of-death determination by LLMs.

Regarding previous studies, there is a paper evaluating the diagnostic performance of GPT-4 and Claude 3 Opus in answering radiology's diagnosis. The questions are based solely on textual information from clinical histories and imaging findings. With the input of text information from clinical histories and imaging findings, GPT-4, the previous version of GPT-4o, correctly answered 54% of the questions [3]. The accuracy of primary diagnoses obtained by Claude 3 Opus from clinical histories plus imaging findings was 55.3% [11]. These percentages obtained in both studies were similar, suggesting that the abilities of the two models were similar. Additionally, there is a pre-print article comparing the diagnostic performance of GPT-4o and Claude 3 Opus, and the diagnostic accuracy of Claude 3 Opus was superior to that of GPT-4o for reasons that are still unclear [12].

In this study, there was no difference in the number of cases that could not be listed as the top three diagnoses in the two models. There may be strengths and weaknesses in reasoning between the models; however, it is unclear what quirks exist in the models. There was no significant difference in the number of cases in which a differential diagnosis was not made, indicating that the use of either model is currently acceptable.

Although the diagnostic performance of LLMs is high, the basis for their decisions may not coincide with that of medical specialists in a reasonable number of cases. Even LLMs with high diagnostic performance in medical imaging may not have a high degree of concordance between the rationale for the decision and the findings that the specialist emphasized. Furthermore, the information and prompts provided to LLMs were found to be important. This raises concerns about the validity of cause-of-death diagnoses in LLMs, as there is a risk that inappropriate information may lead to unexpected diagnoses in practice.

Nevertheless, this study has some limitations. First, there is a bias in the distribution of included cases. A small number of cases had hemorrhagic lesions (12/100) that could be detected by postmortem CT. It could be considered a pilot study in terms of the limited number of cases and possible bias in cause-of-death groups. More cases are expected to be examined. Second, there is a potential bias arising from the use of non-standardized medical histories and postmortem CT reports and the variability in the information provided by different physicians. The degree to which the medical history is summarized, i.e., whether it is detailed or more concise, varies depending on the physician who treated the patient. While it is undeniable that the prepared histories did not vary from physician to physician and included information essential for postmortem CT interpretation and pathological autopsy as directed, it remains unclear whether a history that does not directly lead to death or one more focused on information related to death is preferable when assessing the cause of death by LLMs, which were diagnosed using information available on the Internet in this study. Third, some cases resulted in differences in cause-of-death representation between the radiologist and the pathologist. We cannot deny the possibility that the same pathological condition may have been assumed but expressed differently when verbalized. Additionally, although the gold standard for diagnosing the cause of death used in this study was a pathology diagnosis by autopsy, it is unlikely that differences in cause-of-death diagnoses among pathologists, including the possibility of multiple coexisting causes of death, do not exist. Fourth, the LLMs' outputs are likely to differ depending on whether the input is in Japanese or English.

When assisting in determining the cause of death based on medical histories and postmortem CT findings (at least at present), it cannot be said that LLMs can replace radiologists. However, our findings suggest that large-language AI models, in their current state, maybe more effectively utilized as a supportive tool for radiologists in formulating differential cause-of-death diagnoses based on medical histories and postmortem CT findings rather than as a standalone solution for replacing radiologists in the image interpretation process. Version upgrades can lead to improvements in the diagnostic performance of the LLMs [11-13]. Therefore, it is desirable to continue conducting research and evaluations in the future.

Conclusions

This study demonstrated the usefulness of LLMs in determining the cause of death based on medical histories and postmortem CT findings. Both GPT-4o and Claude 3 Opus showed reasonable accuracy in identifying underlying and immediate causes of death, particularly when provided with detailed input data. While LLMs show promise in assisting with cause-of-death determination for medico-legal purposes, their integration into clinical practice should be approached cautiously, with a clear understanding of their capabilities and limitations. These tools have the potential to enhance the efficiency and accuracy of death investigations, particularly in contexts where specialist resources are limited, but should be used in conjunction with, rather than in place of, expert medical judgment.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Masanori Ishida, Wataru Gonoi

Acquisition, analysis, or interpretation of data: Masanori Ishida, Wataru Gonoi, Keisuke Nyunoya, Hiroyuki Abe, Go Shirota, Naomasa Okimoto, Kotaro Fujimoto, Mariko Kurokawa, Motoki Nakai, Kazuhiro Saito, Tetsuo Ushiku, Osamu Abe

Drafting of the manuscript: Masanori Ishida, Wataru Gonoi, Keisuke Nyunoya, Hiroyuki Abe, Go Shirota, Naomasa Okimoto, Kotaro Fujimoto, Mariko Kurokawa, Motoki Nakai, Tetsuo Ushiku, Osamu Abe

Critical review of the manuscript for important intellectual content: Motoki Nakai, Kazuhiro Saito, Tetsuo Ushiku, Osamu Abe

Supervision: Kazuhiro Saito, Tetsuo Ushiku, Osamu Abe

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. Institutional Review Board, The University of Tokyo Hospital issued approval 2076. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant No. 23K07202, gained by Masanori Ishida, and No. 20K07989, gained by Wataru Gonoi. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

Masanori Ishida and Wataru Gonoi contributed equally to the work and should be considered co-first authors.

References

1. GPTs are GPTs: an early look at the labor market impact potential of large language models . (2023). Accessed: June 24, 2024: https://arxiv.org/abs/2303.10130.
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ: AI in health and medicine. Nat Med. 2022, 28:31-38. 10.1038/s41591-021-01614-0
3. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, Miki Y: ChatGPT’s diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology. 2023, 308:e231040. 10.1148/radiol.231040
4. GPT-4 technical report. (2023). Accessed: June 24, 2024: https://arxiv.org/abs/2303.08774.
5. Horiuchi D, Tatekawa H, Shimono T, et al.: Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2024, 66:73-79. 10.1007/s00234-023-03252-4
6. Zeng Y, Zhang X, Wang J, et al.: Inconsistency between human observation and deep learning models: assessing validity of postmortem computed tomography diagnosis of drowning. J Imaging Inform Med. 2024, 37:1-10. 10.1007/s10278-024-00974-6
7. Hello GPT-4o. (2024). Accessed: May 25, 2024: https://openai.com/index/hello-gpt-4o/.
8. The Claude 3 model family: Opus, Sonnet, Haiku . (2023). Accessed: May 25, 2024: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
9. Nichols L, Saunders R, Knollmann FD: Causes of death of patients with lung cancer . Arch Pathol Lab Med. 2012, 136:1552-1557. 10.5858/arpa.2011-0521-OA
10. Inai K, Noriki S, Kinoshita K, et al.: Postmortem CT is more accurate than clinical diagnosis for identifying the immediate cause of death in hospitalized patients: a prospective autopsy-based study. Virchows Arch. 2016, 469:101-109. 10.1007/s00428-016-1937-6

11. Kurokawa R, Ohizumi Y, Kanzawa J, Kurokawa M, Kiguchi T, Gonoi W, Abe O: Diagnostic performance of Claude 3 from patient history and key images in diagnosis please cases. MedRxiv. 2024, 10.1101/2024.04.11.24305622

12. Sonoda Y, Kurokawa R, Nakamura Y, et al.: Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Jpn J Radiol. 2024, 10.1007/s11604-024-01619-y

13. Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J: Comparing GPT-3.5 and GPT-4 accuracy and drift in radiology diagnosis please cases. Radiology. 2024, 310:e232411. 10.1148/radiol.232411