Journal of Trends in Medical Investigation

J Trends Med Invest 2025;1(2):49-53
https://doi.org/10.64512/jtmi.2025.11

Original Article

Diagnostic performance of ChatGPT in detecting gastrointestinal tract perforation on chest radiographs: a comparative study

Merve Tokoçin[1], Onur Tokoçin[2]

[1]*Department of General Surgery, İstanbul Bağcılar Training and Research Hospital, İstanbul, Türkiye*
[2]*Department of Emergency Medicine, Faculty of Medicine, Arel University, Bahçelievler, İstanbul, Türkiye*

ABSTRACT

Background: Gastrointestinal (GI) tract perforation is a surgical emergency requiring rapid diagnosis, often via chest radiography. Artificial intelligence (AI), including large language models like ChatGPT, has potential to enhance medical imaging but its efficacy in detecting GI perforation is unclear. We compared the diagnostic accuracy of ChatGPT 3.5 and 4 with human experts in interpreting chest radiographs for GI perforation.

Methods: This retrospective study, approved by the Arel University Hospital Ethics Committee (E-52857131-050.06.04-455896), analyzed 504 chest radiographs from patients diagnosed with GI perforation between 2010 and 2021. Radiographs were classified into three groups: definite GI perforation, suspicious requiring further imaging, or no perforation. Two clinicians (emergency medicine specialist and general surgeon) independently evaluated radiographs, followed by ChatGPT 3.5 and 4 using a standardized prompt. Diagnostic accuracy was assessed with chi-square tests, and decision-making times with Student's t-test ($p<0.05$ for significance).

Results: Of 504 patients (11.1% female, mean age 45.4 years), human evaluators correctly classified 80.1% of radiographs, compared with 3.9% for ChatGPT 3.5 and 5.9% for ChatGPT 4 ($p<0.001$). ChatGPT models were faster ($p<0.001$) but failed to interpret 94.1–96.1% of radiographs, often recommending clinical consultation.

Conclusion: General-purpose ChatGPT models lack the accuracy for reliable GI perforation diagnosis on chest radiographs. Specialized AI models, trained on medical imaging datasets, are needed to improve diagnostic precision and support clinical workflows.

Keywords: AI in healthcare, artificial intelligence, emergency radiology, radiology AI

Introduction

Gastrointestinal (GI) tract perforation is a life-threatening condition necessitating urgent surgical intervention. Chest radiography, widely used to detect pneumoperitoneum, is limited by high workloads and human error, which can delay diagnosis. GI perforations, such as those caused by peptic ulcers, appendicitis, or diverticulitis, carry significant morbidity and mortality if not diagnosed promptly, with mortality rates ranging from 10–30% depending on the underlying cause and time to intervention (1). In busy emergency departments, where clinicians often face high patient volumes and time constraints, rapid and accurate diagnosis is critical to

✉ Merve Tokoçin ▪ mervetokocin@gmail.com

Received: 23.07.2025 ▪ Accepted 25.08.2025

improving patient outcomes. The complexity of interpreting subtle radiological signs, such as free intraperitoneal air, underscores the need for tools that can enhance diagnostic efficiency without compromising accuracy.

Artificial intelligence (AI), particularly large language models like ChatGPT (OpenAI), has shown promise in medical imaging by reducing diagnostic time and errors. ChatGPT, developed by OpenAI, is a conversational AI model based on the GPT architecture, designed to process and generate human-like text and, in later versions, interpret visual inputs such as images. Since its release, ChatGPT has been adopted globally across diverse fields, including education, customer service, content creation, and healthcare, due to its ability to process vast datasets and provide contextually relevant responses (2,3). In healthcare, its applications range from answering medical queries to assisting with clinical documentation, but its role in diagnostic imaging remains underexplored. This study evaluates the diagnostic performance of ChatGPT 3.5 and 4 in detecting GI perforation on chest radiographs compared with human experts, assessing accuracy, speed, and clinical applicability.

Methods

Study Design and Participants

This retrospective study, approved by the Arel University Hospital Non-Invasive Ethics Committee (E-52857131-050.06.04-455896), included 504 patients diagnosed with GI perforation between January 2010 and December 2021 at Arel University Hospital. Patients were identified from emergency department records, and posteroanterior chest radiographs were retrieved. Informed consent was obtained for image use in research.

Radiographs were classified into three groups:

1. Definite GI perforation (no further imaging needed).

2. Suspicious for GI perforation (additional imaging required).

3. No GI perforation (further imaging mandatory) (Figure 1).

Human Evaluation

Two experienced clinicians (an emergency medicine specialist and a general surgeon) independently classified radiographs without access to ChatGPT outputs or each other's assessments. A consensus diagnosis was established for discordant cases via discussion.

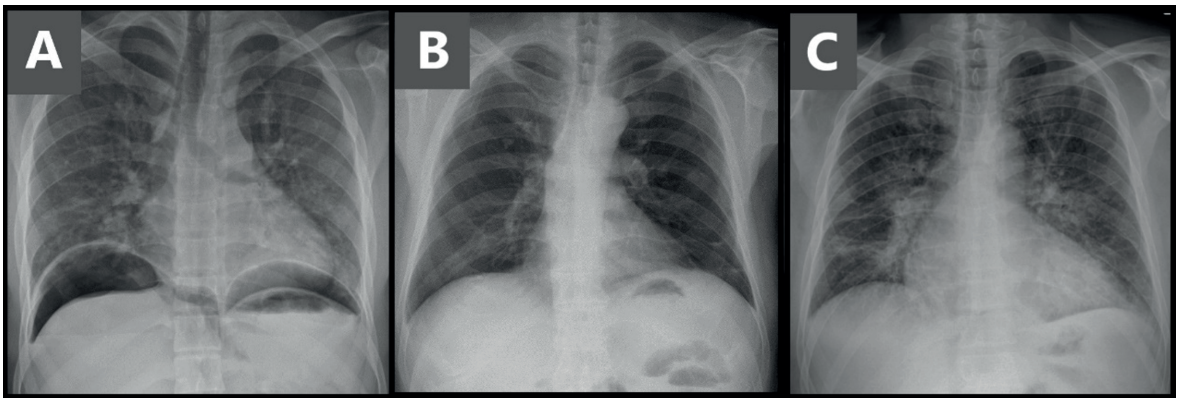

Figure 1. Cheast graphs were divided into 3 groups: a) GI perforation is certain; b) GI perforation is suspicious; c) no GI perforation

AI Evaluation

Radiographs were uploaded to ChatGPT 3.5 and 4 using a single authenticated account. The standardized prompt was: "Examine the chest radiograph and identify any pneumoperitoneum." Each image was input once without modification to ensure consistency. ChatGPT outputs were categorized into the three groups based on their responses or lack thereof (e.g., inability to interpret) (Table 1).

Statistical Analysis

Diagnostic accuracy was calculated as the proportion of correct classifications compared with the human consensus. Chi-square tests assessed differences in correct classification rates ($p<0.05$ for significance). Decision-making times were compared using Student's t-test ($p<0.001$). Confidence intervals (95% CI) were calculated for accuracy estimates, and Cohen's d was used for time comparisons. Analyses were conducted using SPSS version 25.0 (IBM Corp., Armonk, NY, USA).

Results

Patient Characteristics

Of 504 patients, 56 (11.1%) were female, and the mean age was 45.4 years (range 18–86). All patients had confirmed GI perforation via chest radiography (n=394, 78.2%), abdominal tomography (n=94, 18.7%), or diagnostic laparotomy (n=16, 3.2%).

Human Evaluation

Initial human evaluation classified 394 (78.2%) radiographs as definite GI perforation (Group 1), 94 (18.7%) as suspicious (Group 2), and 16 (3.2%) as no perforation (Group 3). Secondary review, conducted without time constraints, reclassified 413 (81.9%; 95% CI 78.3–85.2) to Group 1, 77 (15.3%; 95% CI 12.2–18.8) to Group 2, and 14 (2.8%; 95% CI 1.5–4.6) to Group 3.

AI Evaluation

ChatGPT 3.5 correctly classified 3.9% (20/504; 95% CI 2.4–6.0) of radiographs, and ChatGPT 4 classified 5.9% (30/504; 95% CI 4.0–8.4) correctly ($p<0.001$ vs. human evaluation). Both models failed to interpret 94.1–96.1% of radiographs, often responding with "consult a healthcare professional." Decision-making times were significantly faster for ChatGPT (mean 2.3 s for 3.5, 2.1 s for 4) than humans (mean 45.6 s; $p<0.001$, Cohen's $d=3.2$).

Statistical Findings

Chi-square tests confirmed that human classifications were significantly more accurate than ChatGPT ($\chi^2=392.4$, $p<0.001$). Time efficiency analysis showed ChatGPT's speed advantage ($t=28.7$, $p<0.001$) (Table 2).

Table 1. Diagnostic Accuracy and Time Efficiency of Human Experts and ChatGPT for GI Perforation on Chest Radiographs

Evaluator	Group 1: Definite GI Perforation (n, % [95% CI])*	Group 2: Suspected GI Perforation (n, % [95% CI])*	Group 3: No Perforation (n, % [95% CI])*	Non-Interpretable Rate (n, %)	Total Time (s)	Time Range (s)
Human (ER)	394 (78.2, 74.3–81.7)	94 (18.7, 15.3–22.4)	16 (3.2, 1.8–5.1)	0 (0.0)	10,206	60–720
Human (Home)	413 (81.9, 78.3–85.2)	77 (15.3, 12.2–18.8)	14 (2.8, 1.5–4.6)	0 (0.0)	14,952	60–780
ChatGPT 3.5	20 (4.0, 2.4–6.0)	0 (0.0, 0.0–0.7)	0 (0.0, 0.0–0.7)	484 (96.0)	1,159	48–432
ChatGPT 4	30 (6.0, 4.0–8.4)	0 (0.0, 0.0–0.7)	0 (0.0, 0.0–0.7)	474 (94.0)	1,075	48–420

Table 2. Correct Diagnoses by Human Experts and ChatGPT for GI Perforation on Chest Radiographs

Evaluator	Correct Diagnoses (n/504)	Accuracy (% [95% CI])	Non-Interpretable Rate (n, %)
Human (ER)	394	78.2 (74.3–81.7)	0 (0.0)
Human (Home)	413	81.9 (78.3–85.2)	0 (0.0)
ChatGPT 3.5	20	4.0 (2.4–6.0)	484 (96.0)
ChatGPT 4	30	6.0 (4.0–8.4)	474 (94.0)

Discussion

This study highlights the limitations of general-purpose ChatGPT models (3.5 and 4) in diagnosing GI perforation from chest radiographs, with accuracy rates of 3.9% and 5.9%, respectively, compared with 80.1% for human evaluators. The poor performance likely stems from the models' lack of specific training for medical imaging, as they are designed for general text and image processing rather than radiological interpretation. Specialized AI models, trained on curated medical datasets, have demonstrated superior performance in other imaging tasks, such as detecting pneumothorax or lung cancer, suggesting potential for improvement with tailored algorithms (4,5).

ChatGPT's speed advantage (2.1–2.3 s vs. 45.6 s) is notable but clinically irrelevant given its low accuracy. The standardized prompt used may have limited performance, as more specific or iterative prompts could enhance output quality. Additionally, the models' frequent inability to interpret radiographs and default to recommending clinical consultation underscores their unsuitability for standalone diagnostic use.

In medicine, ChatGPT has been explored in various applications beyond imaging, including clinical decision support, patient education, and medical documentation. For instance, studies have investigated its ability to answer medical queries, assist in drafting clinical notes, and even support medical education by generating practice questions or summarizing complex literature (2,6,7). In radiology, ChatGPT has been tested for tasks such as generating radiology reports and interpreting imaging findings, though with mixed results (8,9). A systematic review by Keshavarz et al. found that while ChatGPT shows promise in radiology report generation, its diagnostic accuracy for complex imaging tasks, such as identifying subtle abnormalities, remains limited due to insufficient training on specialized datasets (7). Similarly, Ahyad et al. reported that ChatGPT could reduce reporting time for radiologists but struggled with nuanced interpretations, aligning with our findings (8).

Literature involving ChatGPT in medical contexts highlights both its potential and limitations. For example, Xue et al. noted its utility in translational medicine for summarizing research findings but cautioned against overreliance due to potential inaccuracies (2). Fijačko et al. demonstrated that ChatGPT could pass certain medical exams, suggesting competence in knowledge-based tasks, but its performance in practical, image-based diagnostics remains inadequate (5). These studies collectively emphasize the need for specialized AI training to bridge the gap between general-purpose models and clinical applications. Our findings align with this, as ChatGPT's inability to reliably detect GI perforation underscores the necessity for domain-specific AI models in radiology (10).

Limitations include the use of general-purpose ChatGPT models, which are not optimized for radiology, and the retrospective design, which may not reflect real-time clinical challenges. Future research should explore fine-tuned AI models, incorporate diverse imaging modalities (e.g., CT), and assess iterative prompting strategies. Additionally, integrating ChatGPT with other AI tools, such as computer-aided

detection systems, could enhance its utility in clinical workflows (4). Ethical concerns, including data privacy and algorithmic bias, must also be addressed to ensure safe AI integration into clinical practice (11,12).

Conclusion

General-purpose ChatGPT models are not suitable for diagnosing GI perforation on chest radiographs due to low accuracy, despite faster processing times. Specialized AI models, developed with medical imaging expertise, are needed to enhance diagnostic precision and support clinicians. Addressing ethical and technical challenges will be crucial for AI's safe integration into medical workflows.

Ethical approval

The study was approved by Arel University Hospital Ethics Committee (number: E-52857131-050.06.04-455896).

Author contribution

The authors confirm contribution to the paper as follows: Study conception and design: MT, OT; data collection: MT, OT; analysis and interpretation of results: MT, OT; draft manuscript preparation: MT, OT. All authors reviewed the results and approved the final version of the manuscript.

Source of funding

The authors declare the study received no funding.

Conflict of interest

The authors declare that there is no conflict of interest.

REFERENCES

1. Patel V, Khan MN, Shrivastava A, et al. Artificial intelligence applied to gastrointestinal diagnostics: a review. J Pediatr Gastroenterol Nutr. 2020;70(1):4-11. [Crossref]
2. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. Clin Transl Med. 2023;13(3):e1216. [Crossref]
3. Hu M, Pan S, Li Y, Yang X. Advancing medical imaging with language models: a journey from N-grams to ChatGPT. arXiv. 2023:2304.04920.
4. Hwang EJ, Hong JH, Lee KH, et al. Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. Eur Radiol. 2020;30(7):3660-71. [Crossref]
5. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation. 2023;185:109732. [Crossref]
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40. [Crossref]
7. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. Diagn Interv Imaging. 2024;105(7-8):251-65. [Crossref]
8. Ahyad RA, Zaylaee Y, Hassan T, et al. Cutting edge to cutting time: can ChatGPT improve the radiologist's reporting? J Imaging Inform Med. 2025;38(1):346-56. [Crossref]
9. Shen OY, Pratap JS, Li X, Chen NC, Bhashyam AR. How does ChatGPT use source information compared with Google? a text network analysis of online health information. Clin Orthop Relat Res. 2024;482(4):578-88. [Crossref]
10. Shah P, Kendall F, Khozin S, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit Med. 2019;2:69. [Crossref]
11. Thondebhavi Subbaramaiah M, Shanthanna H. ChatGPT in the field of scientific publication - are we ready for it? Indian J Anaesth. 2023;67(5):407-8. [Crossref]
12. Ong CWM, Blackbourn HD, Migliori GB. GPT-4, artificial intelligence and implications for publishing. Int J Tuberc Lung Dis. 2023;27(6):425-6. [Crossref]