Letter to the Editor

Data Contamination in AI Evaluation

Alaeddin Acar, MD

Department of Neurosurgery, Kulu State Hospital, Konya, Turkey

Corresponding Author:
Alaeddin Acar, MD
Department of Neurosurgery, Kulu State Hospital
No 4, 139518 Street, Dinek, Kulu
Konya, 42770
Turkey
Phone: 90 542 472 37 23
Email: alaeacar@gmail.com

Related Articles:
Comment on: https://medinform.jmir.org/2025/1/e68409/
Comment in: http://medinform.jmir.org/2025/1/e82057/

(JMIR Med Inform 2025;13:e80987) doi: 10.2196/80987

KEYWORDS

artificial intelligence; large language model; ChatGPT; emergency medicine; clinical performance examination; history taking; clinical reasoning; empathy; patient experience

This letter is regarding the recent publication of the article titled "Clinical Performance and Communication Skills of ChatGPT Versus Physicians in Emergency Medicine: Simulated Patient Study" by Park et al [1]. The study makes a significant contribution to the growing field of artificial intelligence (AI) evaluation in medicine, and I congratulate the authors on their valuable work. However, I would like to highlight a potential methodological limitation in the written examination portion of the study. The authors state that their examination questions were taken from a 2018 textbook, *100 Cases in Emergency Medicine and Critical Care* [2]. The AI model they tested, ChatGPT (OpenAI), was trained on huge amounts of public text from the internet, which likely included this textbook. This means ChatGPT may have seen exactly the same questions and answers during its training.

This problem is known as "data contamination." If the AI has already seen the test questions, its high scores might show good memory, not good medical reasoning. This makes the comparison to human doctors, who were seeing the questions for the first time, unfair. The study found that ChatGPT performed much better than doctors on this written test, but this result could be due to this methodological limitation.

Other researchers in the field are aware of this problem and take steps to avoid it. For example, a study by Busch et al [3] on radiology used private, members-only cases that were not likely in the AI's training data to minimize this risk. Another study by Noda et al [4] on a Japanese medical examination used questions from an examination that took place after the AI's training data cut-off date.

These studies show the importance of using new and unseen questions when testing AI. Because the study by Park et al [1] did not use this approach, I believe the results of their written examination should be viewed with caution. Future studies must use methods like those in the Busch et al [3] and Noda et al [4] papers to ensure a fair and valid test of AI's abilities.

Acknowledgments

Google Gemini was used for language editing.

Conflicts of Interest

None declared.

References

1. Park C, An MH, Hwang G, Park RW, An J. Clinical performance and communication skills of ChatGPT versus physicians in emergency medicine: simulated patient study. JMIR Med Inform. Jul 17, 2025;13:e68409. [FREE Full text] [doi: 10.2196/68409] [Medline: 40674718]
2. Shamil E, Ravi P, Mistry D. 100 Cases in Emergency Medicine and Critical Care. Boca Raton, FL. CRC Press; 2018.

3. Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L. Integrating text and image analysis: exploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res. May 01, 2024;26:e54948. [doi: 10.2196/54948] [Medline: 38691404]
4. Noda M, Ueno T, Koshu R, Takaso Y, Shimada MD, Saito C, et al. Performance of GPT-4V in answering the Japanese otolaryngology board certification examination questions: evaluation study. JMIR Med Educ. Mar 28, 2024;10:e57054. [FREE Full text] [doi: 10.2196/57054] [Medline: 38546736]

Abbreviations

AI: artificial intelligence

Edited by A Iannaccio; this is a non–peer-reviewed article. Submitted 20.Jul.2025; accepted 20.Aug.2025; published 29.Sep.2025.

Please cite as:
Acar A
Data Contamination in AI Evaluation
JMIR Med Inform 2025;13:e80987
URL: https://medinform.jmir.org/2025/1/e80987
doi: 10.2196/80987
PMID: