Journal of Global Research in Computer Sciences ISSN: 2229-371X

A Commentary on Computerized Diagnostic Decision Support Systems-A Comparative Performance Study of Isabel Pro *versus* ChatGPT4

Joe M Bridges*

School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, United States of America

Commentary

Received: 27-Jun-2024, Manuscript No. GRCS-24-140027; **Editor assigned:** 01-Jul--2024, Pre QC No. GRCS-24-140027 (PQ); **Reviewed:** 15-Jul-2024, QC No. GRCS-24-140027; **Revised:** 22-Jul-2024, Manuscript No. GRCS-24-140027 (R); **Published:** 29-Jul-2024, DOI: 10.4172/2229-371X.15.2.001

***For Correspondence:**

Joe M Bridges, School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, United States of America

E-mail: joe.bridges@uth.tmc.edu

Citation: Bridges JM. A Commentary on Computerized Diagnostic Decision Support Systems-A Comparative Performance Study of Isabel Pro *versus* ChatGPT4. J Glob Res Comput Sci. 2024;15:001

ABSTRACT

This paper compares the diagnostic performance of the commercially available diagnostic decision support system, Isabel Pro, to the OpenAI generative pre-trained artificial intelligence system, ChatGPT4. The study used 201 cases, each with a confirmed diagnosis, using identical inputs, requesting a differential diagnosis listing, and comparing the ranking of the correct diagnosis by each system using Mean Reciprocal Rank (MRR) and Recall at Rank for ranks 1, 5, 10, 20, 30 and 40. ChatGPT4 was requested to provide a complete reference citation for each diagnosis returned in its differential. An MRR of 1.0 would imply the correct diagnosis presented as the first-ranked diagnosis in all cases. ChatGPT4 returned an MRR of 0.428, while Isabel Pro returned an MRR of 0.389. ChatGPT4 outperformed on Recall at Ranks 1, 5, and 10, while Isabel Pro outperformed at ranks 20, 30, and 40. The 201 cases were insufficient to conclude that the systems were equivalent. The concerning issue for the clinical use of ChatGPT4 is "What reference substantiates the correct diagnosis?" ChatGPT4 fabricated over 12% of the references cited and almost 70% of the DOI. The study concludes that while the promise of artificial intelligence is high, the fabrication of references will limit the clinical use of these models until they achieve absolute accuracy.

Keywords: Artificial intelligence; Diagnosis; Computer assisted; Isabel pro; ChatGPT4

DESCRIPTION

Very few technical innovations have seen such rapid usage and adoption growth, as have the large language models, especially the generative pre-trained models such as OpenAI's ChatGPT. Beginning in March, 2023, the growth has been nothing short of explosive [1]. Much has been made of the potential for using these systems in medicine, diagnosis included [2-5]. Each article notes the need for more extensive validation of ChatGPT's diagnostic accuracy and the need to substantiate the basis for a given diagnosis. This study used two computerized diagnostic decision support systems-Isabel Pro, a commercially available system developed by Isabel Healthcare, Ltd., and ChatGPT4, the large language model developed by OpenAI. Isabel Pro employs a proprietary search engine that addresses a proprietary database of highly regarded medical references material, such as the Merck Manual Professional and Cochrane Reports and medical textbooks. Isabel Pro's database is updated monthly. ChatGPT4 is trained on the Common Crawl, an extensive, publicly available text data set. At present, the training includes all items through January, 2022. The study employed 201 cases, 175 from those published in the New England Journal of Medicine and 36 from the library of Dr. Charles P. Friedman, University of Michigan Medical School. Each case had a confirmed diagnosis. The dataset for this study was more extensive than any previous set of cases by a factor of three or four, covering a wide range of disease conditions, medical specialties, and patient demographics. The study entered exactly identical input for each system and requested 40 differentials, a listing substantially longer than any previous study. The Research Question in this study was "Given that studies have shown a statistically significant improvement in clinicians' diagnostic accuracy using Isabel Pro [6,7], does the large language model ChatGPT4 produce a greater number of accurate diagnoses ranked higher in presentation than Isabel Pro?" While ChatGPT4 was slightly better in MRR (0.428 *versus* 0.389) and in the top 10 presentation of diagnosis rankings (69% *versus* 65%), the most significant concern noted by this study was the unknown process used by ChatGPT4 to produce its differential diagnosis ranking, especially given the significant number of reference fabrications. Both systems failed to diagnose several cases, roughly 13% for each. Isabel Pro is a finely crafted system that is easy to use, fast, and references the best medical reference sources [8]. ChatGPT4 is noticeably slower, frequently requiring requests to continue the listing. Improving diagnostic accuracy is a vital need in today's clinical practice, with estimates of diagnostic accuracy being 95% in the United States, implying about 12 million diagnostic errors annually, with half likely resulting in patient harm [9]. The most challenging job humans undertake, medical diagnosis, requires that we expend all possible effort to improve diagnostic accuracy. Computerized diagnostic decision support systems are a promising method to help clinicians improve diagnostic accuracy [10]. Artificial intelligence shows great promise, but is unlikely to be widely used by clinicians until the "Black Box" nature of its process is revealed and the fabrication of references resolves in favor of absolute accuracy.

REFERENCES

1. Wachter RM, et al. Will generative artificial intelligence deliver on its promise in health care? JAMA. 2023;331:65-69.
2. Liu X, et al. Evaluating Chatgpt as an adjunct for analyzing challenging case. Blood. 2023;142:7273.
3. Horiuchi D, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2023;66:73-79.
4. Hailu R, et al. ChatGPT-assisted diagnosis: Is the future suddenly here? STAT. 13 February, 2023.

5. Dave T, et al. Chatgpt in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
6. Sibbald M, et al. Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel. BMJ Qual Saf. 2021;31:426-433.
7. Bridges JM. Evaluation, validation, and implementation of a computerized diagnostic decision support system in primary care. Houston: University of Texas Health Science Center at Houston, United States of America. 2022.
8. Riches N, et al. The effectiveness of electronic Differential Diagnoses (DDX) generators: A systematic review and meta-analysis. PLOS One. 2016;11:e0148991.
9. Singh H, et al. Types and origins of diagnostic errors in primary care settings. JAMA. 2013;173:418.
10. Graber ML. Reaching 95%: Decision support tools are the surest way to improve diagnosis now. BMJ Qual Saf. 2022;31:415-418.