Scripta
Spring 2024, Vol. 3, No. 1, 45-49
ISSN: 2769-9242

ChatGPT's Viability in Medical Diagnostics

Karthik Kuppuswamy *

Abstract

Despite the unprecedented rate of scientific advancement in recent years, misdiagnosis remains a devastating, fatal, and all too frequent occurrence within the medical profession. The seemingly inexplicable lack of correlation between the evolution of the medical field and a decrease in misdiagnosis rates can be primarily attributed to the fact that the intrinsic fallibility of humans, exacerbated by the pressure of working conditions, cannot be completely remedied by any technology, no matter how advanced. However, recent research discussing the potential incorporation of ChatGPT into medical diagnostics has yielded promising results. Through the analysis of a variety of modern literature discussing the potential applications of ChatGPT in the medical field, this article will assess the viability of ChatGPT in producing ethical, accurate medical diagnostics, considering factors like ease of implementation, past usages of machine learning, and the algorithmic soundness of the model itself. The findings indicate that numerous refinement measures must be implemented prior to the consideration of ChatGPT in diagnostics, mainly due to unreliable medical information in current databases and training required to effectively utilize ChatGPT in a specialized context.

Introduction

According to Zheqing Zhang, an administrator in the Medical Science Division at Oxford College, patients with uncommon illnesses will, on average, receive three misdiagnoses and consult five doctors before being diagnosed accurately [1]. Logically, a pattern recognition algorithm should theoretically function at a much higher degree of precision due to its objectivity and capacity for processing and categorizing large amounts of data almost instantly. However, almost all machine learning models lack a crucial component that renders them virtually unusable for medical diagnostics in their current state: coherence and ease of communication. Currently, no technology has demonstrated the level of coherence or usability to justify its integration into medical diagnostics. The development of ChatGPT by OpenAI in 2022, however, presents a promising step towards achieving these criteria. ChatGPT has redefined the relationship between technology

*Junior Student, Fulton Science Academy.
Contact: kkuppuswamy@fultonscienceacademy.org

and the professional domain. Whereas technology had previously served as a supplement to increase productivity, ChatGPT possesses enough autonomy to match, and sometimes even outperform, experienced professionals in numerous fields. Despite this progress, integrating AI faces significant hurdles. Medical diagnostics, being riddled with privacy and security barriers, remains resistant to the advent of artificial intelligence and ChatGPT. However, this newfound technology is actively undergoing comprehensive study to determine its viability as a diagnostic tool, prompting questions about both its applicability and its ethical implications. Although some may argue that the immediate integration of ChatGPT into the medical diagnostic process for uncommon diseases would reduce misdiagnosis and increase efficiency, further refinement prior to its implementation is necessary due to its dependence on a limited and sometimes unreliable database and its reliance on meticulous, precise prompt engineering techniques that require specialized training to develop.

Methods

A comprehensive literature review was conducted to determine the efficacy of ChatGPT in medical diagnostics, as well as machine learning in general. Sources were primarily obtained from electronic databases, specifically Google Scholar. Advantages, limitations, and methods for ChatGPT's optimization were identified, and collected data was utilized to evaluate ChatGPT's diagnostic competence.

Assessing Viability

Advocated of ChatGPT's integration into medical diagnostics emphasize its extraordinary potential, highlighting its capacity to engage with patients in a manner almost indistinguishable from a human doctor and the past successes of machine learning in diagnostics necessitate its immediate integration into medical diagnostics. These proponents often reference machine learning's success in medical diagnostics in the past as a precedent for ChatGPT's implementation. Loredana Caruccio, an assistant professor at the University of Salerno, echoes this perspective by describing machine learning's widespread use "in intelligent diagnosis systems, with notable success in areas such as cancer diagnosis, cardiovascular disease, and medical image analysis" [2]. Given that the listed areas of medical diagnostics

are often some of the most problematic for doctors, these successes appear to set a strong precedent for ChatGPT's efficacy and applicability. Additionally, ChatGPT demonstrated an "overall median humanness" of 7 on a scale from 5-9, reinforcing its ability to provide coherent, human outputs and seemingly strengthening the case for ChatGPT's immediate implementation [3]. However, this argument overlooks machine learning's more auxiliary role in the diagnostic system.

The main weakness of this argument stems from the historical fact that the role of machine learning in diagnoses was purely supplementary, so machine learning success in the past cannot be extrapolated to guarantee ChatGPT's success. According to Caruccio, traditional machine learning approaches employed feature engineering methodologies, preprocessing techniques, and parameter tuning to ensure the accuracy of their results. These approaches effectively limit the data channeled into producing a response, ensuring that data is reliable and relevant. ChatGPT, on the other hand, is a pre-trained predictive model trained with massive chunks of data that performs natural language processing tasks, compiling the data into coherent responses independently of human intervention [2]. ChatGPT's proposed utility to diagnostics involves far more independence and far less regulation from medical professionals, increasing the risk of unremedied error.

One of the primary causes for this error is ChatGPT's database, which frequently contributes to partially or entirely incorrect diagnoses by providing irrelevant or vague information. In a study performed by Ryan King, a cardiologist at UC Irvine, both ChatGPT models generated accurate responses to the majority of general medical questions. However, GPT-3.5 produced responses containing inaccurate information "slightly more often than GPT-4 (10.8% vs 5.1%)," when responding to more specific questions about cardiology and gastroenterology, also generating one entirely incorrect response. King attributed GPT-4's superior performance to "the model's possible enhanced exposure to relevant data during training" [4]. This connection between "relevant training" and favorable outcomes demonstrates the need for medical-specific databases, especially for rare diseases, where information on the internet and in existing databases may be limited. Additionally, the decline of ChatGPT's performance when responding to more specific medical questions underlines its current unsuitability for rare disease diagnosis. Similarly, according to Tomoyuki Kuroiwa, an orthopedic surgeon at Tokyo Medical and Dental University,

a comprehensive review of symptom checkers indicated that the accuracy of predictive models increased when situational data, like seasonal and personal information, was inputted [5]. This trend indicates that more specific data is necessary to make accurate diagnoses, necessitating the training of ChatGPT with situational data and tailored databases.

In addition to database refinement, employing optimal prompt-creation techniques can also enhance accuracy and efficiency, necessitating prompt-specific training for medical professionals. To obtain accurate, comprehensible, and relevant outputs from AI models, the user must input a very specific set of demands. The way in which this set of demands is formulated and conveyed, known as prompt engineering, can drastically affect both the structure and content of the AI model's output. Unlike humans, AI lack the capacity to interpret subtle communication cues and implied meanings emphasizing the importance of concise, discrete prompt engineering. For instance, Harriet Walker, a researcher at the National Health Society Foundation Trust, conceded that when evaluating the performance of ChatGPT, the use of prompts involving words like "hepatocellular carcinoma" tended to yield improved responses and that entering "less technical vocabulary...might have affected the results" [6]. This correlation between technical vocabulary and the efficacy of ChatGPT further underlines the delicacy of prompt injection and the training required to maximize ChatGPT's utility in diagnostics. An example of prompt engineering to streamline AI integration, proposed by Caruccio, is "I have these symptoms: [S]. Which diagnosis do you think is the most accurate among [D] [H]?" This format is effective because narrowing down the diagnoses transforms the problem of diagnosis into a probabilistic problem, which AI models are excellent at solving. Additionally, by including suspected diagnoses, the rate of AI hallucinations, which are completely incorrect responses induced either by inaccurate source information or extrapolation by the AI model, decreases significantly. Numerous different structures have been proposed for various disorders, but their success relies heavily on appropriate training protocol for healthcare workers.

Conclusion

Although ChatGPT possesses enormous potential to streamline the medical diagnostic process for rare disease in the future, it currently presents far too many algorithmic inconsistencies and issues with data reliability to be

considered a viable tool for rare disease diagnosis. However, as with other machine learning models, ChatGPT could play a more supplemental role in diagnostics before its more widespread implementation. This approach would allow us to evaluate its usability and aptitude in a low-risk manner and, if appropriate, transition to a fully automated diagnostic system. For example, medical reporting, or the documentation of patient symptoms, would be well-suited to ChatGPT's strengths of objective reasoning and almost instantaneous processing. This application would indirectly benefit medical diagnostics by improving the accuracy of collected data that the diagnosis is dependent on [3]. Additionally, ChatGPT's ease of communication and readily comprehensible outputs enable it to facilitate patient education, particularly those with rare disease. As medicine-specific, reviewed databases are developed and ethical guidelines established, ChatGPT's potential utility in the medical profession will skyrocket. For now, however, experimenting with novel, unproven technology in such a high-risk, sensitive field is morally unjustifiable, illogical, and could eliminate any possibility of its implementation in the future.

References

1. Zhang, Z. Diagnosing rare diseases and mental well-being: a family's story. *Orphanet Journal of Rare Diseases* **18,** 45 (2023).
2. Caruccio, L. *et al.* Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Systems with Applications* **235,** 121186 (2024).
3. Liu, J., Wang, C. & Liu, S. Utility of ChatGPT in clinical practice. *Journal of Medical Internet Research* **25,** e48568 (2023).
4. King, R. C. *et al.* A multidisciplinary assessment of ChatGPTs knowledge of amyloidosis. *medRxiv,* 2023–07 (2023).
5. Kuroiwa, T. *et al.* The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. *Journal of Medical Internet Research* **25,** e47621 (2023).
6. Walker, H. L. *et al.* Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *Journal of Medical Internet Research* **25,** e47479 (2023).