EVALUATING CHATGPT'S RESPONSES TO FREQUENTLY ASKED QUESTIONS REGARDING POLYCYSTIC OVARY SYNDROME

Lauren Pace, MD[1], Nicholas Kummer, BS[2], Fernando Bril, MD[3], Pardis Hosseinzadeh, MD, MS[4], Ricardo Azziz, MD, MBA, MPH[1,3,5,6].
[1]Dept. of Ob/Gyn, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA
[2], Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA,
[3]Dept. of Medicine, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA
[4]Dept. of Gynecology and Obstetrics, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[5]Dept. of Healthcare Organization & Policy, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA
[6]Dept. of Health Policy, Management, and Behavior, School of Public Health, University at Albany, SUNY, Rensselaer, NY, USA

Background: As natural language processing (NLP) models have become more mainstream, their potential applications in healthcare have been increasingly explored. The popular Artificial Intelligence (AI)-powered tool Chat-Generative Pre-Trained Transformer (ChatGPT) has been demonstrated to have utility in patient acquisition of healthcare related information[1,2]. More specifically, it has demonstrated the capability to provide appropriate responses to fertility and reproduction-related queries[3,4]. Polycystic ovary syndrome (PCOS) is a common, and yet poorly understood, condition among reproductive-age females with numerous implications for lifestyle, health, and fertility.

Objective: Our study seeks to evaluate the responses of the NLP ChatGPT to 27 common patient queries regarding PCOS.

Materials and Methods: In October 2023, 27 distinct queries related to the diagnosis and background, treatment, and fertility implications of PCOS were submitted to ChatGPT. Two fellowship-trained reproductive endocrinologists and one fellowship trained medical endocrinologist evaluated each response on a 1-7 Likert scale across five categories: thoroughness, accuracy, readability, clinical applicability, and misinformation. Statistics were evaluated and agreement between raters was determined by Intraclass Correlation Coefficient (ICC) two-way random effects model and "single measures" unit. Calculations were made with IBM SPSS Statistics 29.

Results: Across all raters, the category with the highest average score was readability with a mean of 6.68 (with a maximum of 7 representing the highest ease of readability. Clinical applicability was next highest with an average score of 6.30, and thoroughness was similarly rated at 6.23. The lowest score was in the category of accuracy at 6.01. Misinformation was rated, on average, at 1.51 (with a minimum of 1 representing no misinformation, and 2 representing very little to nearly no misinformation). Questions related to fertility in PCOS demonstrated the highest thoroughness and accuracy and the lowest rates of misinformation. There was significant ($p<0.001$) absolute agreement between all raters via ICC two-way random effects model and "single measures" unit, ICC=0.806 [0.688-0.875].

Conclusions: Chat-GPT 3.5 demonstrates that NLP models have potential utility in providing thorough, accurate, and easily readable responses to potential patient questions, specifically as they pertain to PCOS. The clinical applicability of responses is generally favorable and rates of

misinformation provided on the subject are low, indicating a fair amount of reliability in the responses. Of questions posed to the software, those related to fertility in PCOS patients performed the best across three of five evaluated categories. The potential for ChatGPT to be a clinical practice adjunct has yet to be fully evaluated, but its tendency to provide appropriate responses to common patient questions in this area provides potential insight to its utility as an easily accessible patient education tool.

Financial Support: None

References:

1. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq*. Published online February 28, 2023:rs.3.rs-2566942. doi:10.21203/rs.3.rs-2566942/v1

2. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Evaluating ChatGPT’s Accuracy in Lung Cancer Prevention and Screening Recommendations. *Radiol Cardiothorac Imaging*. 2023;5(4):e230115. doi:10.1148/ryct.230115

3. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertil Steril*. 2023;120(3, Part 2):575-583. doi:10.1016/j.fertnstert.2023.05.151

4. AI and machine learning can successfully diagnose polycystic ovary syndrome. National Institutes of Health (NIH). Published September 18, 2023. Accessed October 29, 2023. https://www.nih.gov/news-events/news-releases/ai-machine-learning-can-successfully-diagnose-polycystic-ovary-syndrome