Accepted and to be published in Journal of Medical Research and Innovation

Cite as Billah M, Amin S, Barua O. Assessment of 'Florence' in Addressing Inquiries on Nicotine Replacement Therapy**.** J Med Res Innov. 2024;8(1):e000293. DOI: 10.32892/jmri.293

Title: Assessment of 'Florence' in Addressing Inquiries on Nicotine Replacement Therapy

Authors
Meer Sadad Billah, DO[1]; Oishi Barua, MBBS[2]; Samia Amin, MBBS/MD, PhD [3*]

Affiliations
1. NYIT College of Osteopathic Medicine, United States
2. Chattogram Maa-O-Shishu Hospital Medical College, Bangladesh
3. Macquarie University, Australia

***Corresponding author**
Samia Amin
Email: samia.amin@hdr.mq.edu.au

Author contribution
MSB: Methodology, Writing review and editing
OB: Methodology, Writing review and editing
SA: Conceptualization, Methodology, Writing original draft, Supervision

Funding: None

Conflict of interest: None

Dear editor

Artificial intelligence (AI) assisted chatbots, or conversational agents are new digital tools that mimic instantaneous human conversation. As AI assistants become more prevalent, evaluating their accuracy and consistency in providing health information is important. Evidence suggests that role of chatbots in smoking cessation is promising particularly in participant's engagement.[1] The World Health Organization has launched a digital health worker '*Florence*' (a virtual human), powered by AI as its newest resource for providing the general population with accurate health information on COVID-19 vaccines and treatments, mental health, including smoking cessation.[2]

Background

Nicotine replacement therapy (NRT) is a widely recommended approach for smoking cessation to manage withdrawal symptoms associated with quitting smoking, such as irritability, cravings,

and mood swings.[3] It is recommended to use these products under the guidance of a healthcare professional.[3] However, it is evident that people often search for information about addiction help-seeking queries from AI assistants.[4] It is critical to understand the role and reliability of '*Florence*,' especially in smoking cessation. The objective of this research was to elucidate whether '*Florence*' provides evidence-based information in response to common NRT questions.

Methodology

A rigorous, methodical process was followed to develop an effective evaluation scale.[5] The evaluation scale was developed to comprehensively assess the performance of an AI system across 3 parameters. In the first parameter, the AI evaluated on 'voice recognition' and 'question understanding'. Voice recognition is scored on a scale from 0 to 2, where 0 represents a failure to differentiate between male and female voices, 1 indicates inconsistent recognition, and 2 signifies reliable recognition. Similarly, 'question understanding' is assessed on the same scale, with 0 denoting a lack of understanding, 1 representing inconsistent understanding, and 2 indicating consistent comprehension of questions. The second parameter was 'consistency in answers between researchers', the AI's performance is measured by answer consistency. Scores range from 0 to 2, where 0 signifies completely different answers between researchers, 1 suggests somewhat different answers, and 2 denotes identical answers. The third parameter, 'accuracy of answers', evaluated the AI's precision in providing correct responses. The scale ranges from 0 to 2, with 0 indicating completely inaccurate answers, 1 representing somewhat accurate answers with significant errors, and 2 signifying entirely accurate responses. The overall assessment is derived from the total score, where a cumulative score of 0-2 indicates 'poor' performance, 3-4 reflects 'fair' performance, 5-6 signifies 'good' performance, and 7-8 represents 'excellent' performance.

The scoring guidance was provided to support consistency across evaluators. We pilot tested this matrix before main data collection by collecting 20 questions from 'Quora' platform related to smoking cessation. Finally, the team had then read through the ACS FAQ webpage responses to evaluate consistency and accuracy and compare them with the responses from the '*Florence*.'

Fifty-six NRT questions were obtained from the American Cancer Society website.[6] Two researchers independently queried '*Florence*' and recorded the responses over a two-week period in January 2024. Responses were compared to the American Cancer Society answers to evaluate accuracy and between researchers to assess consistency. An 8-point rating scale was used across 3 evaluation parametres: voice recognition, question understanding, answer consistency, and accuracy.

Results

Out of 56 NRT questions asked, 11 questions (19.6%) were answered with excellent accuracy and depth of knowledge, demonstrating a strong command of the topics covered. Total 44 questions (78.6%) were rated as fair performance. Responses to these questions had some minor flaws in accuracy, comprehensiveness of information, or depth of explanation. There is room for improvement to address gaps in knowledge. Only 1 question (1.8%) received a poor performance rating. Approximately one-fifth of responses met excellence criteria, over three-fourths still have space for improving quality in content, detail, precision, or accuracy.

The results indicate a mixed performance of '*Florence*' in addressing NRT-related queries. The identified gaps in knowledge, as evidenced by the 'fair' performance ratings, underscore the need for continuous improvement in the AI system. Addressing these gaps could enhance the quality of content, detail, precision, and overall accuracy of responses. As the field of AI-assisted chatbots in health information provision evolves, ongoing evaluations and refinements are essential to ensure these tools meet the highest standards in accuracy and reliability. This research contributes valuable insights that can guide future enhancements in AI-assisted health information tools, ultimately benefiting individuals seeking reliable guidance in their health journeys.

This study's strengths lie in its comprehensive evaluation methodology, real-world application, and actionable insights for improvement. Yet, limitations such as potential biases in comparison sources and subjectivity in evaluations emphasize the need for careful consideration in interpreting '*Florence*'s' performance.

Conclusion

In summary, while 'Florence' excelled in linguistic processing like speech and question comprehension, supplemental training focused on strengthening NRT knowledge itself would help address shortcomings in consistency, precision, completeness, and depth when answering domain-specific questions. Targeted improvement tuning both language mastery and core subject matter competencies could boost overall performance from fair to excellent across evaluations.

References

1. Whittaker, R., Dobson, R., & Garner, K. (2022). Chatbots for smoking cessation: scoping review. Journal of Medical Internet Research, 24(9), e35556

2. World Health Organization: *Meet Florence 2.0, She can give you advice on a healthier lifestyle and mental health*. Available on: https://www.who.int/campaigns/Florence

3. Wadgave, U., & Nagesh, L. (2016). Nicotine Replacement Therapy: An Overview. International journal of health sciences, 10(3), 425–435.

4. Nobles, A. L., Leas, E. C., Caputi, T. L., Zhu, S. H., Strathdee, S. A., & Ayers, J. W. (2020). Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. NPJ digital medicine, 3, 11. https://doi.org/10.1038/s41746-019-0215-9

5. DeVellis, R. F., & Thorpe, C. T. (2021). Scale development: Theory and applications. Sage publications.

6. American Cancer Society. *Nicotine Replacement Therapy to help you quit tobacco*. Available on: https://www.cancer.org/cancer/risk-prevention/tobacco/guide-quitting-smoking/nicotine-replacement-therapy.html