The Importance of Documenting Chatbot Performance in the Management of Specific and Rare Conditions for Patients

Dear Editor,

We reviewed the correspondence entitled "*AI Chatbots in Treatment Decision-making for Acquired Bilateral Vocal Fold Paralysis: Correspondence.*"[1] The authors discussed the methodology of our paper entitled "*Evaluating the Potential of AI Chatbots in Treatment Decision-making for Acquired Bilateral Vocal Fold Paralysis in Adults*,"[2] and they suggested that the lack of openness in the Chatbot's decision-making process, and the lack of information about the AI models' calibration and training are major flaws of the study. We thank the authors for sharing their opinion. In this letter, we wish to draw attention to few points.

We agree that the lack of information on the hyperparameters of ChatGPT-3.5, 4.0., and Llama 2.0 can make it difficult for researchers or health care providers to analyze the performance of Chatbots. This point does not only concern our study but all studies that assessed such Chatbots in providing medical or surgical information. However, this study was not conducted to better understand the functioning and performance of Llama 2.0 and ChatGPT-4.0. The study was conducted to evaluate the performance of Chatbots in providing medical or surgical information dedicated to the management of clinical vignettes. In this clinical way and from a pragmatical standpoint, the lack of knowledge of the hyperparameters does not discourage people, including patients, from getting information about the disease and seeking advice regarding treatment from Chatbots. This reality makes particular sense for such a rare condition as bilateral vocal fold paralysis (BVFP), which is characterized by the lack of consensus on the management and treatment of the condition.[3] Thus, the conclusion of our study could help patients with BVFP to be cautious when considering the potential outputs of Chatbots in regard to their condition.

This point is particularly important as ChatGPT-4.0 was found to report moderate-to-high accuracy in some common otolaryngological conditions.[4–6] For example, the largest cohort study of real clinical cases submitted to ChatGPT-4.0 recently suggested a strong consistency ($k > 0.600$) between otolaryngologists and ChatGPT-4 for the indication of some additional examinations (ie, upper aerodigestive tract endoscopy, positron emission tomography and computed tomography, audiometry, tympanometry, and psychophysical evaluations), while the primary diagnosis was correctly performed by ChatGPT-4 in 38% to 86% of cases depending on subspecialty.[4] In rhinology, ChatGPT-4 proposed plausible and correct primary diagnoses in 62.5% cases,[5] while in laryngology, ChatGPT had the highest performance for proposing a primary (90%) or the most plausible differential diagnoses (65%), and the therapeutic options (60%–68%).[6] These studies, with abstracts easily available through a Google search, may encourage patients to trust the Chatbot's performance in providing accurate clinical information. The conduction of such studies assessing the accuracy of Chatbots in the management of specific and rare conditions, or in providing medical or surgical information, carries a great value for clinical practice and patient use of Chatbots despite the lack of information on hyperparameters.

Declaration of Competing Interest

The authors have no conflict of interest.

Acknowledgments

NA.

Emilie A.C. Dronkers[*]
Ahmed Geneid[†]
Chadwan Al Yaghchi[*]
Jerome R. Lechien[‡,§]
[*]National Centre for Airway Reconstruction, Imperial College Healthcare NHS Trust, London, UK
[†]Department of Otolaryngology and Phoniatrics-Head and Neck Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland
[‡]Department of Anatomy and Experimental Oncology, Mons School of Medicine, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium
[§]Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, Paris Saclay University, Paris, France
Address correspondence and reprint requests to Jerome R. Lechien, Laboratory of Anatomy and Cell Biology, Faculty of Medicine, University of Mons (UMONS), Avenue du Champ de mars, 6, B7000 Mons, Belgium.
E-mail: Jerome.Lechien@umons.ac.be (J.R. Lechien).

https://doi.org/10.1016/j.jvoice.2024.04.032

Funding: None.

0892-1997

References

1. Daungsupawong H, Wiwanitkit V. AI chatbots in treatment decision-making for acquired bilateral vocal fold paralysis: correspondence. *J Voice.* 2024.
2. Dronkers EAC, Geneid A, Al Yaghchi C, et al. Evaluating the potential of AI chatbots in treatment decision-making for acquired bilateral vocal fold paralysis in adults. *J Voice.* 2024. https://doi.org/10.1016/j.jvoice.2024.02.020.

3. Lechien JR, Hans S, Mau T. Management of bilateral vocal fold paralysis: a systematic review. *Otolaryngol Head Neck Surg.* 2024;170:724–735. https://doi.org/10.1002/ohn.616.
4. Lechien JR, Naunheim MR, Maniaci A, et al. Performance and consistency of ChatGPT-4 versus otolaryngologists: a clinical case series. *Otolaryngol Head Neck Surg.* 2024. https://doi.org/10.1002/ohn.759.
5. Radulesco T, Saibene AM, Michel J, et al. ChatGPT-4 performance in rhinology: a clinical case series. *Int Forum Allergy Rhinol.* 2024. https://doi.org/10.1002/alr.23323.
6. Lechien JR, Georgescu BM, Hans S, et al. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. *Eur Arch Otorhinolaryngol.* 2024;281:319–333. https://doi.org/10.1007/s00405-023-08282-5.