LARYNGOLOGY

ChatGPT-4 accuracy for patient education in laryngopharyngeal reflux

Jerome R. Lechien[1,2,3,4] · Thomas L. Carroll[5] · Molly N. Huston[6] · Matthew R. Naunheim[1,7,8]

Abstract

Introduction Chatbot Generative Pre-trained Transformer (ChatGPT) is an artificial intelligence-powered language model chatbot able to help otolaryngologists in practice and research. The ability of ChatGPT in generating patient-centered information related to laryngopharyngeal reflux disease (LPRD) was evaluated.

Methods Twenty-five questions dedicated to definition, clinical presentation, diagnosis, and treatment of LPRD were developed from the Dubai definition and management of LPRD consensus and recent reviews. Questions about the four aforementioned categories were entered into ChatGPT-4. Four board-certified laryngologists evaluated the accuracy of ChatGPT-4 with a 5-point Likert scale. Interrater reliability was evaluated.

Results The mean scores (SD) of ChatGPT-4 answers for definition, clinical presentation, additional examination, and treatments were 4.13 (0.52), 4.50 (0.72), 3.75 (0.61), and 4.18 (0.47), respectively. Experts reported high interrater reliability for sub-scores (ICC = 0.973). The lowest performances of ChatGPT-4 were on answers about the most prevalent LPR signs, the most reliable objective tool for the diagnosis (hypopharyngeal-esophageal multichannel intraluminal impedance-pH monitoring (HEMII-pH)), and the criteria for the diagnosis of LPR using HEMII-pH.

Conclusion ChatGPT-4 may provide adequate information on the definition of LPR, differences compared to GERD (gastroesophageal reflux disease), and clinical presentation. Information provided upon extra-laryngeal manifestations and HEMII-pH may need further optimization. Regarding the recent trends identifying increasing patient use of internet sources for self-education, the findings of the present study may help draw attention to ChatGPT-4's accuracy on the topic of LPR.

Keywords ChatGPT · Artificial intelligence · Chatbot · Reference · Otolaryngology · Head neck surgery · Laryngopharyngeal · Reflux

✉ Jerome R. Lechien
Jerome.Lechien@umons.ac.be

1 Research Committee, Young Otolaryngologists of the International Federation of Otorhinolaryngological Societies (IFOS), Paris, France

2 Division of Laryngology and Broncho-Esophagology, Department of Otolaryngology-Head Neck Surgery, EpiCURA Hospital, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium

3 Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, School of Medicine, Phonetics and Phonology Laboratory (UMR 7018 CNRS, Université Sorbonne Nouvelle/Paris 3), Paris, France

4 Polyclinique Elsan de Poitiers, Poitiers, France

5 Division of Otolaryngology-Head and Neck Surgery, Brigham and Women's Hospital, Department of Otolaryngology-Head and Neck Surgery, Harvard Medical School, Boston, MA, USA

6 Department of Otolaryngology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA

7 Department of Otolaryngology-Head and Neck Surgery, Harvard Medical School, Boston, MA, USA

8 Division of Laryngology, Massachusetts Eye and Ear, Boston, MA, USA

Introduction

The Chatbot Generative Pre-trained Transformer (Chat-GPT, OpenIA, CA, USA) is an artificial intelligence-powered language model that can responds to both simple and complicated questions in science [1, 2]. The version 3.5 is free-access, which makes ChatGPT a new source of medical information for patients. However, ChatGPT information is not reviewed by practitioners; this is problematic due to potential misinformation or information that may conflict with a practitioner's recommendations [3]. The aim of the present study was to evaluate the ability of ChatGPT to generate patient-centered, accurate information surrounding laryngopharyngeal reflux disease (LPRD).

Methods

Twenty-five questions were developed regarding the definition, clinical presentation, diagnostics, and treatment of LPRD from the recent Dubai definition and diagnostic criteria of LPRD consensus [International Federation of Otorhinolaryngological Societies (IFOS)] [4].

The Dubai consensus is a Delphi-based international guideline conducted by IFOS and including 48 experts in LPRD from five continents. The consensus consists of 38 statements regarding definition, pathophysiology, and differences compared to gastroesophageal reflux disease (GERD) ($N=3$); prevalence/incidence of LPR ($N=1$); contributing factors ($N=2$) and associations with other otolaryngological conditions ($N=3$); symptoms ($N=4$); findings ($N=4$); and additional examinations ($N=21$; Appendix 1) [4]. In addition to the Dubai consensus paper, four key recent publications (reviews and meta-analyses) published in the past 5 years were considered in the establishment of questions and expected responses [5–8]. These papers were studies from the IFOS group and report the highest citations and research interest score (RIS; ResearchGate) metrics in the LPRD field. The four papers included a state-of-the art review dedicated to the epidemiology, clinical presentation and management of LPRD (268 citations; 67 citations/year; RIS: 81.2) [5]; a systematic review and meta-analysis of LPRD treatments (149 citations; 30 citations/year; RIS: 58.1) [6], a best practice paper in treatment of LPRD (2 citations; 2 citations/year; RIS: 3.0) [7]; and a contemporary review of the management of recalcitrant LPRD (40 citations; 13 citations/year; RIS: 10.5) [8]. The questions are reported in Table 1. The institutional review board of CHU Saint-Pierre was not required for this study (ref.CHUST23).

Chatbot input and output

Questions regarding definition, clinical presentation, diagnosis, and treatment of LPRD were entered through the ChatGPT-4 interface accessible via the API (https://chat.openai.com). An example input was: "*Could you provide reliable information concerning laryngopharyngeal reflux disease, also called extraesophageal reflux disease? My doctor…*". The chatbot responses in terms of accuracy, comprehensiveness, and 'similarity to a response they would give' were analyzed through 5-point Likert scales by four blinded, board-certified otolaryngologists, fellowship-trained in laryngology [3]. The following definitions of the Likert scores were considered: 1 = very inaccurate, very incomprehensive, or very dissimilar to laryngologist response; 2 = inaccurate, incomprehensive, or dissimilar to laryngologist response, 3 = somewhat accurate, somewhat comprehensive, or somewhat similar to laryngologist response; 4 = accurate, comprehensive, or similar to laryngologist response; and 5 = very accurate, very comprehensive, or very similar to laryngologist response. As proposed by Davis et al. [3], an output graded between 1 and 3 was defined as a response less than suitable for patient use.

Statistical analyses

Statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (SPSS version 22,0; IBM Corp, Armonk, NY, USA) and Microsoft Excel (version 2302; Microsoft, Redmond, WA, USA). The interrater reliability was evaluated with interrater reliability coefficients for the ChatGPT-4 sub-scores: definition (/20), clinical presentation (/25), additional examinations (/30), and treatment (/50) (Table 1). The consistency was considered as low, moderate and strong for k < 0.40, 0.40–0.60, and k > 0.60, respectively.

Results

The accuracy of ChatGPT-4 responses is available in Table 1. The mean Likert scores of ChatGPT-4 definition, clinical presentation, additional examinations, and treatments were 4.13 ± 0.52, 4.50 ± 0.73, 3.75 ± 0.61, and 4.18 ± 0.47, respectively. Judges reported high interrater reliability (Fig. 1). The ChatGPT-4 responses associated with the lower scores were related to hypopharyngeal-esophageal multichannel intraluminal impedance-pH monitoring (HEMII-pH) criteria, the most prevalent signs of LPR, and the most reliable objective examinations to make an LPR diagnosis. More specifically, ChatGPT-4

**Table 1** Questions submitted into ChaGPT-4

Question text	L1	L2	L3	L4	Av	SD
Definition and epidemiology						
1. Are GERD and LPR the same disease? If there are some differences, could you explain them?	4	4	4	4	4.00	0.00
2. What are the pathophysiological mechanisms underlying the development of upper aerodigestive tract mucosa inflammation in patients with LPR?	4	5	4	4	4.25	0.50
3. Is the diet important in the development of LPR? If yes, can you explain which foods/beverages are suspected to be associated with LPR?	5	5	5	4	4.75	0.50
4. What is the incidence of LPR symptoms in outpatients consulting in otolaryngology?	4	3	4	3	3.50	0.58
Definition and epidemiology score (/20)	17	17	17	15	16.50	1.00
Symptoms and findings						
5. What are the 10 most prevalent symptoms of LPR?	4	5	4	5	4.50	0.58
6. What are the 10 most prevalent signs of laryngopharyngeal reflux?	3	3	3	4	3.25	0.50
7. Do patients with laryngopharyngeal reflux have frequently heartburn or typical GERD-symptoms?	5	5	5	5	5.00	0.00
8. Is it important to use patient-reported outcome questionnaires in the management of LPR and why?	5	5	5	5	5.00	0.00
9. Is there a significant association between symptom and sign severities in LPR patients?	5	4	5	5	4.75	0.50
Symptom and finding score (/25)	22	22	22	24	22.5	1.00
Additional examinations						
10. Is gastrointestinal endoscopy mostly normal or abnormal in LPR patients?	4	4	4	4	4.00	0.00
11. What is the most reliable objective examinations for the diagnosis of LPR?	4	4	2	3	3.25	0.96
12. About pH-monitoring studies, what are the most important characteristics of the probe to ensure detection of LPR?	4	4	3	4	3.75	0.50
13. What are the criteria for objective diagnosis at the pH study?	3	4	2	3	3.00	0.82
14. What are the sensitivity and specificity of pepsin saliva tests (for example, Peptest)?	4	2	4	5	3.75	1.26
15. Is high-resolution manometry indicated in the management of LPR and for which situation?	5	5	5	4	4.75	0.50
Additional examination score (/30)	24	23	20	23	22.50	1.73
Treatments						
16. Could you name the several medications that may be used in the treatment of LPR?	5	5	3	3	4.00	1.15
17. Are proton pump inhibitors evidence-based in the management of LPR?	5	5	4	4	4.50	0.58
18. What are the characteristics of pharyngeal reflux events in terms of acidity, patient position of occurrence (that is, upright or supine), time of day occurrence (that is, daytime or nighttime)?	4	5	4	4	4.25	0.50
19. Are alginate drugs useful in the management of LPR? If yes, why?	5	5	4	5	4.75	0.50
20. Are the findings of HEMII-pH transposable to oropharyngeal pH monitoring (Restech)?	4	3	4	3	3.50	0.58
21. Could you cite 5 reliable differential diagnoses of LPR?	3	4	4	4	3.75	0.50
22. Is fundoplication associated with good postoperative results in the management of LPR?	3	4	3	4	3.50	0.58
23. What is the proportion of patients who do not have symptom improvement of relief after treatment?	3	5	5	5	4.50	1.00
24. What is the appropriate duration of medical treatment in laryngopharyngeal reflux?	4	5	4	4	4.25	0.50
25. When patients have adequate therapeutic responses, how long does it take for patients to improve?	4	5	5	5	4.75	0.50
Treatment score (/50)	40	46	40	41	41.75	2.87

This list of questions was established by the two authors considering the Dubai Criteria Consensus Statements

did not mention the existence of HEMII-pH and recommended making an objective LPR diagnosis using pH-only monitoring (without impedance) (Appendix 2). Similarly, weakly acidic or alkaline pharyngeal reflux events were not reported in physiological and therapeutic questions. The responses of ChatGPT-4 regarding pepsin saliva detection findings were flawed, according to the judges' evaluations.

Discussion

Laryngopharyngeal reflux is one of the most prevalent conditions in otolaryngology, accounting for 10% to 30% of outpatients visits in general otolaryngology [9, 10]. According to the high prevalence and the impact of LPR on patient quality of life, LPR is a perfect candidate and
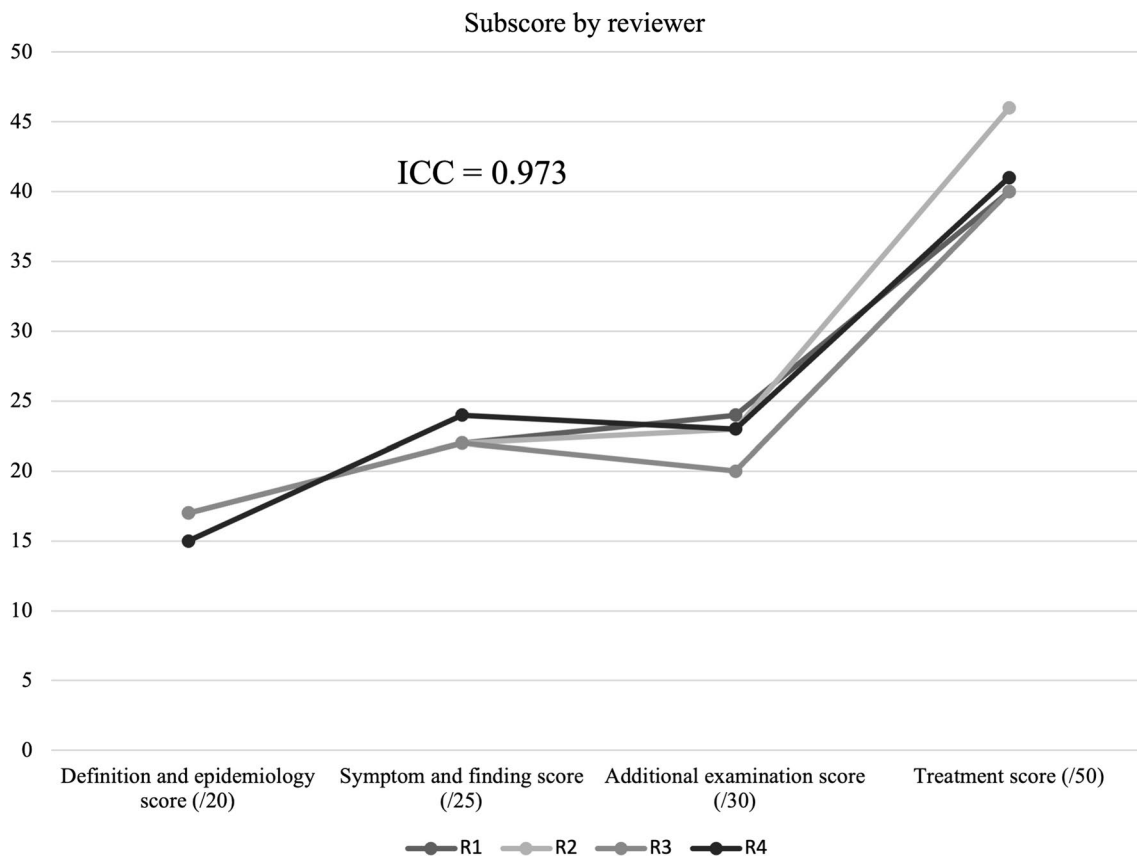
Fig. 1 Interrater reliability. ICC = intraclass correlation

common fodder for internet research by patients who may be concerned about their symptoms [11, 12] The findings of the present study support that the responses of ChatGPT-4 are accurate, comprehensive, or similar to a laryngologist's response in most categories, including definition, epidemiology, clinical presentation and treatments. While diagnosis of LPR remains a controversial topic in otolaryngology and gastroenterology, the recent studies and consensus papers confirm the usefulness and reliability of hypopharyngeal-esophageal multichannel intraluminal impedance-pH monitoring (HEMII-pH) for the diagnosis. HEMII-pH can provide objective findings with data on the full esophageal column and further elucidate if a reflux event reaches the pharynx [4, 5, 13]. ChatGPT-4 does not appear to have knowledge of HEMII-pH because only dual-probe pH monitoring (acid only detection without impedance) was mentioned, likely due to an assumption that LPR is mainly acidic in nature. The lack of awareness about HEMII-pH may explain the fact that alginate and magaldrate therapies were not mentioned as potential therapies for LPR. The accuracy of ChatGPT in patient education was similarly investigated in other prevalent otolaryngological diseases including rhinoplasty recovery [14], sleep apnea [15], thyroid nodules

[16], oropharyngeal cancer information [3], and sialendoscopy clinical findings [17]. In rhinoplasty recovery, the accuracy of information provided by ChatGPT was judged as satisfactory, enhancing patient education and alleviating emotional distress by providing general information and reassurance during the recovery process [14]. In sleep apnea, the information provided by ChatGPT was considered highly reliable without identified incorrect or dangerous information [15]. The accuracy of ChatGPT in thyroid nodule education was lower, however, compared to other otolaryngological fields. Indeed, Campbell et al. reported that ChatGPT responses were judged as correct in 69.2% of questions [16]. The poor results of ChatGPT were supported in oropharyngeal oncology. Davis et al. [3] observed that ChatGPT responses may not educate patients to an appropriate degree, and could outright misinform them, reading at a more difficult grade level than is recommended for patient material. In sialendoscopy and salivary gland diseases, authors reported a moderate-to-high agreement between experts and ChatGPT-3.5 responses [17]. The comparison of our findings with those of the literature remains difficult for several reasons. First, in most cases, authors did not specify the used ChatGPT version. ChatGPT-4 has been found to be more effective than

ChatGPT-3.5 in providing references in otolaryngology [18]. Second, the accuracy of ChatGPT may improve over time with new versions and information. To date, we have little information about the hyperparameters of GPT-4 that are set before the learning process begins which can affect how well a model trains. Third, the accuracy and performance of ChatGPT may be influenced by the disease itself—the accuracy of the information of a more controversial condition may be poorer than that of a more well-understood condition.

This study is the first to investigate the accuracy of Chat-GPT-4 in providing clinical information related to LPR. The use of recent Dubai consensus and critical systematic reviews as an information source as well as the assessment of accuracy by four blinded LPR experts is an additional strength. The lack of evaluation of ChatGPT-3.5 accuracy is the primary limitation of the present study. Moreover, the reliability of ChatGPT-4 responses over time was not investigated through regenerated responses but recent studies suggested that ChatGPT-4, unlike ChatGPT-3.5, reports high test–retest reliability.

Conclusion

ChatGPT-4 may provide overall adequate information about the definition of LPR, the difference between LPR and GERD, and symptoms associated with LPR. However, the lack of awareness of the usefulness of HEMII-pH testing, objective criteria diagnosis, weakly acid and alkaline reflux, and alginate use are major flaws. According to recent trends identifying increasing patient use of internet sources for self-education, and the high prevalence of LPR in primary care practice, otolaryngology, pulmonology and gastroenterology, accuracy of ChatGPT with respect to LPR is important as many patients are likely to use this tool in the future.

the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Declarations

References

1. Briganti G (2023) How ChatGPT works: a mini review. Eur Arch Otorhinolaryngol. https://doi.org/10.1007/s00405-023-08337-7
2. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA et al (2023) Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. Otolaryngol Head Neck Surg. https://doi.org/10.1002/ohn.489
3. Davis RJ, Ayo-Ajibola O, Lin ME, Swanson MS, Chambers TN, Kwon DI, Kokot NC (2023) Evaluation of oropharyngeal cancer information from revolutionary artificial intelligence Chatbot. Laryngoscope. https://doi.org/10.1002/lary.31191
4. Lechien JR, Vaezi MF, Chan WW, Allen JE, Karkos PD, Saussez S et al (2023) The Dubai definition and diagnostic criteria of laryngopharyngeal reflux: the IFOS consensus. Laryngoscope. https://doi.org/10.1002/lary.31134
5. Lechien JR, Akst LM, Hamdan AL, Schindler A, Karkos PD, Barillari MR, Calvo-Henriquez C, Crevier-Buchman L, Finck C, Eun YG, Saussez S, Vaezi MF (2019) Evaluation and management of laryngopharyngeal reflux disease: state of the art review. Otolaryngol Head Neck Surg 160(5):762–782. https://doi.org/10.1177/0194599819827488
6. Lechien JR, Saussez S, Schindler A, Karkos PD, Hamdan AL, Harmegnies B, De Marrez LG, Finck C, Journe F, Paesmans M, Vaezi MF (2019) Clinical outcomes of laryngopharyngeal reflux treatment: a systematic review and meta-analysis. Laryngoscope 129(5):1174–1187. https://doi.org/10.1002/lary.27591
7. Kamal AN, Dhar SI, Bock JM, Clarke JO, Lechien JR, Allen J, Belafsky PC, Blumin JH, Chan WW, Fass R, Fisichella PM, Marohn M, O'Rourke AK, Postma G, Savarino EV, Vaezi MF, Carroll TL, Akst LM (2023) Best practices in treatment of laryngopharyngeal reflux disease: a multidisciplinary modified Delphi study. Dig Dis Sci 68(4):1125–1138. https://doi.org/10.1007/s10620-022-07672-9
8. Lechien JR, Muls V, Dapri G, Mouawad F, Eisendrath P, Schindler A, Nacci A, Barillari MR, Finck C, Saussez S, Akst LM, Sataloff RT (2019) The management of suspected or confirmed laryngopharyngeal reflux patients with recalcitrant symptoms: a contemporary review. Clin Otolaryngol 44(5):784–800. https://doi.org/10.1111/coa.13395
9. Kamani T, Penney S, Mitra I, Pothula V (2012) The prevalence of laryngopharyngeal reflux in the English population. Eur Arch Otorhinolaryngol 269(10):2219–2225. https://doi.org/10.1007/s00405-012-2028-1
10. Spantideas N, Drosou E, Karatsis A, Assimakopoulos D (2015) Voice disorders in the general Greek population and in patients with laryngopharyngeal reflux. Prevalence and risk factors. J

Voice 29(3):389.e27–32. https://doi.org/10.1016/j.jvoice.2014.08.006

11. Kang JW, Lee MK, Lee YC, Ko SG, Eun YG (2023) Somatic anxiety in patients with laryngopharyngeal reflux. Laryngoscope Investig Otolaryngol 8(5):1288–1293. https://doi.org/10.1002/lio2.1138

12. Wong MW, Hsiao SH, Wang JH, Yi CH, Liu TT, Lei WY, Hung JS, Liang SW, Lin L, Gyawali CP, Chen PR, Chen CL (2023) Esophageal hypervigilance and visceral anxiety contribute to symptom severity of laryngopharyngeal reflux. Am J Gastroenterol 118(5):786–793. https://doi.org/10.14309/ajg.0000000000002151

13. Salgado S, Borges LF, Cai JX, Lo WK, Carroll TL, Chan WW (2022) Symptoms classically attributed to laryngopharyngeal reflux correlate poorly with pharyngeal reflux events on multichannel intraluminal impedance testing. Dis Esophagus 36(1):doac041. https://doi.org/10.1093/dote/doac041

14. Capelleras M, Soto-Galindo GA, Cruellas M, Apaydin F (2023) ChatGPT and rhinoplasty recovery: an exploration of AI's role in post-operative guidance. Facial Plast Surg. https://doi.org/10.1055/a-2219-4901

15. Cheong RCT, Unadkat S, Mcneillis V, Williamson A, Joseph J, Randhawa P, Andrews P, Paleri V (2023) Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol. https://doi.org/10.1007/s00405-023-08319-9

16. Campbell DJ, Estephan LE, Sina E, Mastrolonardo EV, Alapati R, Amin DR, Cottrill E (2023) Evaluating ChatGPT responses on thyroid nodules for patient education. Thyroid. https://doi.org/10.1089/thy.2023.0491

17. Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, Sanchez-Barrueco A, Saga-Gutierrez C (2023) Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol. https://doi.org/10.1007/s00405-023-08104-8

18. Lechien JR, Briganti G, Vaira LA (2024) Accuracy of Chat-GPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. Eur Arch Otorhinolaryngol