European Archives of Oto-Rhino-Laryngology
https://doi.org/10.1007/s00405-023-08219-y

LARYNGOLOGY

Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI)

Jerome R. Lechien[1,2,3,4,5,11] · **Antonino Maniaci**[1,6] · **Isabelle Gengler**[1,7] · **Stephane Hans**[1,4] · **Carlos M. Chiesa-Estomba**[1,2,8] · **Luigi A. Vaira**[1,9,10]

Received: 26 July 2023 / Accepted: 30 August 2023

Abstract

Objectives To evaluate the reliability and validity of the Artificial Intelligence Performance Instrument (AIPI).

Methods Medical records of patients consulting in otolaryngology were evaluated by physicians and ChatGPT for differential diagnosis, management, and treatment. The ChatGPT performance was rated twice using AIPI within a 7-day period to assess test–retest reliability. Internal consistency was evaluated using Cronbach's α. Internal validity was evaluated by comparing the AIPI scores of the clinical cases rated by ChatGPT and 2 blinded practitioners. Convergent validity was measured by comparing the AIPI score with a modified version of the Ottawa Clinical Assessment Tool (OCAT). Interrater reliability was assessed using Kendall's tau.

Results Forty-five patients completed the evaluations (28 females). The AIPI Cronbach's alpha analysis suggested an adequate internal consistency ($\alpha = 0.754$). The test–retest reliability was moderate-to-strong for items and the total score of AIPI ($r_s = 0.486$, $p = 0.001$). The mean AIPI score of the senior otolaryngologist was significantly higher compared to the score of ChatGPT, supporting adequate internal validity ($p = 0.001$). Convergent validity reported a moderate and significant correlation between AIPI and modified OCAT ($r_s = 0.319$; $p = 0.044$). The interrater reliability reported significant positive concordance between both otolaryngologists for the patient feature, diagnostic, additional examination, and treatment subscores as well as for the AIPI total score.

Conclusions AIPI is a valid and reliable instrument in assessing the performance of ChatGPT in ear, nose and throat conditions. Future studies are needed to investigate the usefulness of AIPI in medicine and surgery, and to evaluate the psychometric properties in these fields.

Keywords Medicine · Surgery · Otolaryngology · Head neck · ChatGPT · Chatbot · Artificial · GPT · Instrument · Tool · Intelligence · Performance · Comparison · Diagnosis · Treatment

Introduction

A chatbot is an electronic system that has been developed to simulate conversations by responding to keywords or sentences. Chatbots are commonly used in various marketing or messaging platforms and websites [1, 2]. In November 2022, OpenAI (Open AI, San Francisco, USA) launched the Chatbot Generative Pre-trained Transformer (ChatGPT), which uses algorithms to respond to questions poses by the users [2]. Since then, many studies have been conducted to assess the performance of ChatGPT in different areas, such as law, business, or medical school exams, scientific manuscript revisions, or in some clinical fields [3–5]. Given to its large database, most experts agreed with the potential usefulness of ChatGPT as an adjunctive instrument in clinical practice, research, or administrative tasks [5]. However, this technology should be investigated for its capabilities and potential risks [6]. From a clinical point of view, the reliability of the current version of ChatGPT (v.4.0) in the diagnosis and the management of real clinical cases appears to be limited [7]. In a recent case series, practitioners subjectively reported that ChatGPT cannot discern the superiority

Jerome R. Lechien and Antonino Maniaci have contributed equally and joined as co-first authors. Carlos M. Chiesa-Estomba and Luigi A. Vaira have contributed equally and joined as co-senior authors.

Extended author information available on the last page of the article

Published online: 12 September 2023

of some additional examinations over others, while it cannot make the diagnosis of some atypical conditions in patients with complex medical or surgical histories (distracting information) [7]. The assessment of the performance of artificial intelligence (AI) chatbots is currently limited by the lack of valid and reliable clinical instruments for the evaluation of the performance of the chatbot. The current performance instruments are only validated for Human and cannot be used for artificial intelligence software because of lack of communication, empathy, and family management.

The objective of this study was to investigate the reliability and validity of the Artificial Intelligence Performance Instrument (AIPI).

Methods

Development of AIPI

The AIPI was developed by the AI Study Group of the Young-Otolaryngologists of the International Federation of Otorhinolaryngological Societies (YO-IFOS), which includes board-certified otolaryngologists and head and neck surgeons. Three experts (J.R.L., L.A.V., S.H.) surveyed the literature on clinical instruments assessing the performance of physicians (e.g., resident, fellow) or medical students in clinical practice. Experts used the following keywords: 'Performance'; 'Tool'; 'Instrument'; 'Achievement'; 'Success'; 'Diagnosis'; 'Management'; and 'Treatment'. The following search databases were used: PubMed, Scopus, and Cochrane Library. The most widely used clinical tools described in the literature consider the following performance outcomes: history; symptoms; physical examinations; differential diagnosis; additional examinations; treatments; communication; time of management; documentation; and technical therapeutic features [8–11] Based on these outcomes, experts developed the AIPI, which includes 9 items assessing to medical and surgical history; symptoms; physical examination; diagnosis; additional examinations; management plan, and treatments (Fig. 1). The scoring of items was defined to be less subjective as possible, avoiding the use of Likert-scale. The final AIPI score ranges from 0 to 20 with a score of 20 indicating excellent clinical case management by the AI, while a score of 0 is associated with inadequate management. AIPI may be subdivided into the 4 following sub-scores associating common items: patient feature score (/6), diagnosis score (/7), additional examination score (/5), and treatment score (/3). AIPI provides a comprehensive approach to clinical cases, intended for use not only in otolaryngology but also in general medicine and surgery.

Outcomes of Artificial Intelligence Performance Instrument (AIPI)	**Practitioner evaluation**			Item score	Subscores
1. Consideration of medical and surgical history in the AI management:	Fully (2)	Partly (1)	Not (0)	…../2	Patient
2. Consideration of symptoms of patients in the AI management	Fully (2)	Partly (1)	Not (0)	…../2	feature score
3. Consideration of physical findings reported by practitioner(s)	Fully (2)	Partly (1)	Not (0)	…../2	…../6
4. The differential diagnoses provided by AI are:	Complete and plausible (3) Incomplete but plausible (2) Incomplete and not plausible for one or several (1) Absent (0)			…../3	
5. The primary diagnosis of AI was:	Correct (3) Plausible (2) Not plausible (1) Absent (0)			…../3	Diagnosis
6. The management plan of AI included potential physical/additional examinations for determining the diagnosis		Yes (1)	No (0)	…../1	score …../7
7. The additional examinations proposed by AI are/include	All pertinent and necessary examinations (3) All pertinent but partialy necessary examinations (2) An association of pertinent, necessary, and inadequate examinations (1) An association of inadequate examinations (0)			…../3	Additional Examination
8. AI identified the most relevant additional examination to perform first	Yes (1) No, AI provided a list without stratification (0)			…../1	Score …../5
9. The treatments proposed by AI are/include	All pertinent and necessary therapeutic findings (3) All pertinent but incomplete therapeutic findings (2) An association of pertinent, necessary, and inadequate therapeutic findings (1) No adequate therapeutic approach (0)			…../3	
				Total AIPI	…../20

Fig. 1 Artificial Intelligence Performance Instrument. AIPI score ranges from 0 (inadequate management) to 20 (adequate management)

Setting and clinical cases

Fifty clinical cases of outpatients consulting in the Departments of Otolaryngology-Head and Neck Surgery of CHU Saint-Pierre (Brussels, Belgium) and the Dour Medical Center (Dour, Belgium) were prospectively recruited in July 2023. The patient medical records needed to be complete regarding history, symptoms, physical examination description, differential diagnosis, potential additional examinations, and treatments. Incomplete clinical cases were excluded. Specifically, the consultation findings of a single otolaryngologist were recorded in a database to be used for the assessment of the internal validity. Then, these consultation findings were controlled by two senior otolaryngologists to conform with the current guidelines, and, therefore, considered as the standard (adequate management) for the assessment of the ChatGPT performance (Fig. 2). The guidelines consisted of the scientific position paper/recommendations of the European and American Societies in Otolaryngology-Head and Neck Surgery.

The data of the consultation were presented to ChatGPT without mentioning the human differential diagnoses,

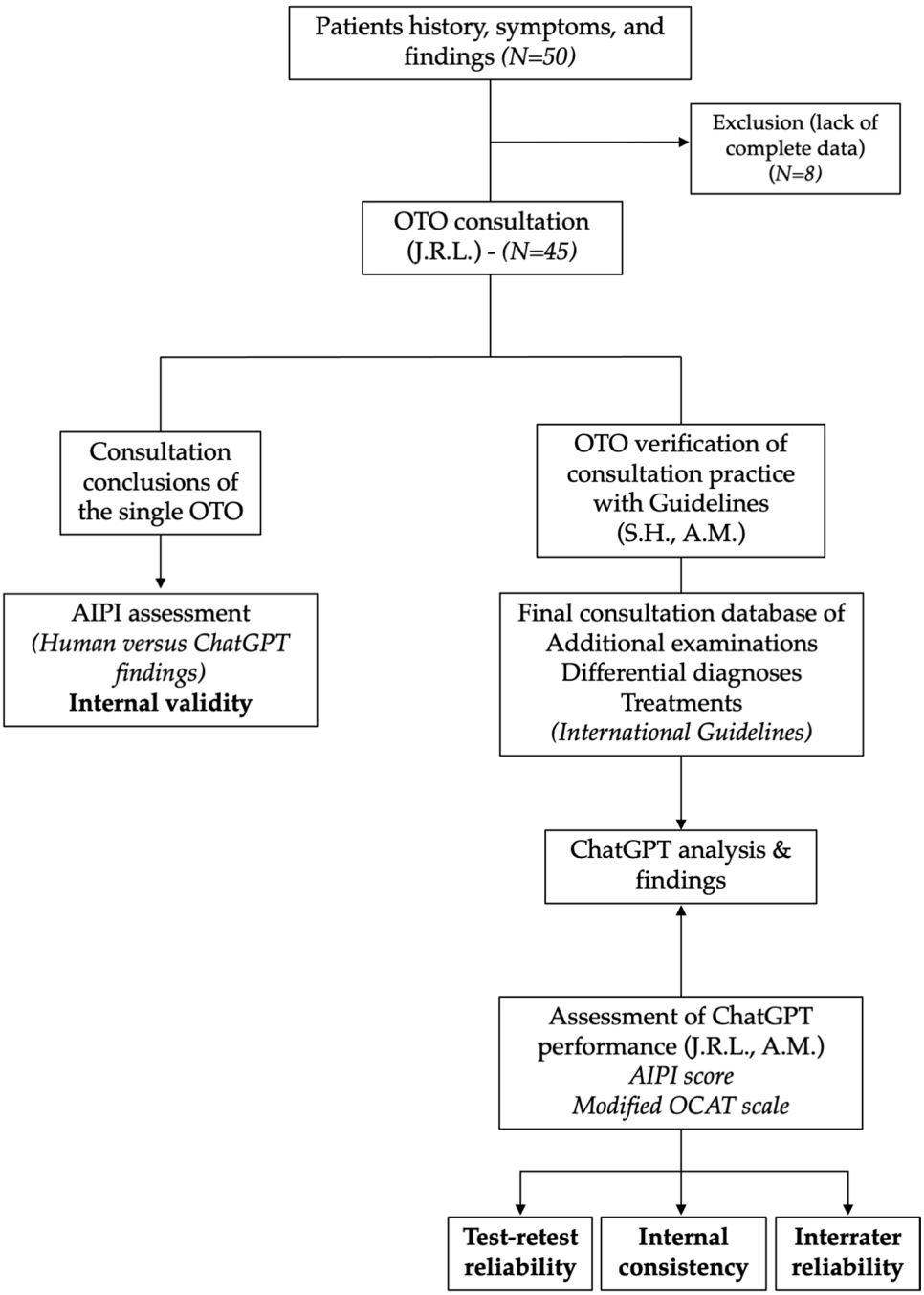

Fig. 2 Chart flow. *OCAT* Ottawa Clinic Assessment Tool, *OTO* otolaryngologist

additional examinations, and treatments. ChatGPT was interrogated for differential diagnoses (What are your differential diagnoses?), additional examinations (What are your additional examinations to find the diagnosis?), and potential therapeutic approach(es) (What are your treatment(s) for the primary diagnosis?). The ChatGPT findings were collected in a database and compared with the practitioner's findings by a panel of two blinded physicians.

The local ethics committee approved the study protocol (CHUSP, n°BE0762023230708). The patient consented to participate.

Statistical methods

Statistical analyses were performed through the Statistical Package for the Social Sciences for Windows (SPSS version 24,0; IBM Corp, Armonk, NY, USA). A level of significance of $p < 0.05$ was used. For correlation analyses, coefficients were considered as low, moderate, and strong for $r_s < 0.30$, 0.30–0.60, and $r_s > 0.60$, respectively. Several psychometric properties were assessed.

Intra- and interrater reliabilities

Internal consistency was measured with Cronbach's alpha. The ChatGPT findings were scored twice with the AIPI within 7 days to assess test–retest reliability (Spearman analysis). The judges' concordance (interrater reliability) was measured through a comparison of the AIPI of two blinded practitioners with Kendall's W (coefficient of concordance; Fig. 2).

Convergent and internal validities

A correlation analysis between scores of AIPI and the diagnostic, management, and treatment items of the Ottawa Clinical Assessment Tool (OCAT) [8] was conducted to measure the convergent validity (Spearman correlation coefficient). OCAT is a valid clinical instrument used to evaluate the performance of residents or fellow-in-training. The OCAT score was rated by two blinded otolaryngologists (C.C., J.R.L.). For each item, otolaryngologists used a 5-point Likert scale ranging from 1 (unprepared to do, inappropriate management) to 5 (can be independent, adequate management) [8]. A total score of the three items was measured to be compared with the AIPI total score.

The internal validity of AIPI was assessed by a comparison of AIPI scores for ChatGPT and the baseline practitioner management (Mann–Whitney *U* test). Precisely, the data of the senior practitioner (J.R.L.) who received the patients were kept in a data depositary and they were judged with the AIPI score to evaluate the internal validity (single human versus ChatGPT; Fig. 2).

Results

Forty-five patients completed the consultation (Fig. 2). There were 28 females and 17 males, respectively. The mean age was 48.0 ± 16.4 years. The primary diagnosis was made in all patients (Table 1). ChatGPT was interrogated for all patient cases. Symptoms, physical examination, history, additional examination, differential diagnosis, and treatment findings of patients are available in Appendices 1 and 2.

Cronbach's alpha analysis suggested an adequate internal consistency ($\alpha = 0.754$). The mean item and total scores of AIPI are reported in Table 2. The AIPI total score and all AIPI subscores assessing the practice of a single otolaryngologist in the consultation were significantly higher than the AIPI total score of ChatGPT, which supports an adequate internal validity (Table 2). The test–retest reliability was moderate-to-high for sub- and total scores of AIPI (Table 3). The convergent validity reported a low-to-moderate and significant association between AIPI and the modified OCAT score ($r_s = 0.319$; $p = 0.045$). The results of the correlation analysis between AIPI and selected OCAT items (differential diagnoses, management plan, and treatment) were detailed in Appendix 3. The physical examination score of ChatGPT was correlated with all OCAT items and total scores. There was a significant association between the differential diagnosis subscore of AIPI and the differential diagnosis score of OCAT ($r_s = 0.569$, $p = 0.001$).

The interrater reliability reported significant positive concordance coefficients between both otolaryngologists for the patient feature, diagnostic, differential diagnosis, and treatment subscores as well as for the AIPI total score (Table 4). The accuracy of ChatGPT in the management of clinical cases is available in Table 5. According to both judges (J.R.L., A.M.), the differential diagnoses and the primary diagnosis of ChatGPT were judged as incomplete and not plausible in 31–42% and 27–29% of cases, respectively (Table 5). Judges reported that additional examinations proposed by ChatGPT were associated with pertinent, necessary, and inadequate examinations in 62–67% of cases. The first and the second judge believed that ChatGPT identified the most relevant additional examination in 24% and 33% of cases, respectively. Regarding treatments, judges reported that ChatGPT proposed an association of pertinent, necessary, and inadequate therapeutic findings in 56% and 60% of cases, while the therapeutic findings were considered pertinent and incomplete in 16% of cases, respectively.

Discussion

The rapid development of intelligent chatbots and their easy availability for patients and physicians make urgent the conduction of clinical studies dedicated to the assessment of

Table 1 Patient symptoms

Outcomes	Patients ($N=45$)
Age (mean, SD)	48.0 ± 16.4
Gender (*N*, %)	
Female	28 (62.2)
Male	17 (37.8)
Primary diagnosis	
Laryngopharyngeal Reflux Disease	5 (11.1)
Laryngopharyngeal carcinoma	3 (6.7)
Presbycusis	3 (6.7)
Vocal fold polyp	2 (4.4)
Unilateral or bilateral vocal cord paralysis	2 (4.4)
Chronic otitis media	2 (4.4)
Eustachian tube dysfunction	2 (4.4)
Vocal fold hemorrhage	1 (2.2)
Vocal fold scarring	1 (2.2)
Bacterial laryngitis	1 (2.2)
Reinke edema	1 (2.2)
Bamboo nodes (vocal folds)	1 (2.2)
Glottis insufficiency	1 (2.2)
Laryngeal primary hypersensitivity	1 (2.2)
Iatrogenic laryngitis	1 (2.2)
Laryngocele	1 (2.2)
Iatrogenic laryngeal superior nerve injury	1 (2.2)
Psychogenic dysphonia	1 (2.2)
Cervical arthrodesis inducing iatrogenic dysphagia	1 (2.2)
Eagle syndrome	1 (2.2)
Esophageal scleroderma (CREST syndrome)	1 (2.2)
Recurrent tonsil infection	1 (2.2)
Salivary lymphoepithelial cyst	1 (2.2)
Salivary lithiasis	1 (2.2)
Supraglottic laryngeal carcinoma (resistant to radiation)	1 (2.2)
Second laryngeal carcinoma	1 (2.2)
Pharyngeal syphilitic ulceration	1 (2.2)
Postviral olfactory dysfunction	1 (2.2)
Rheumatoid polyarthritis	1 (2.2)
Bilateral ear external duct stenosis	1 (2.2)
Benign paroxysmal vertigo	1 (2.2)
Allergic rhinitis	1 (2.2)
Nasal septum hematoma	1 (2.2)

SD standard deviation

chatbot performance. The evaluation of the performance of medical students, residents, or other practitioner categories must include the practitioner's consideration of medical and surgical history, symptoms, and physical examination to propose a list of differential diagnoses, which will be studied through potential additional examinations [12, 13]. Many clinical instruments have been developed to reliably judge practitioner's performance [9–11]. However, according to the differences between Humans and machine assessment, the use of current validated human-based clinical instruments may be inadequate, leading our group to develop AIPI, which is only dedicated to IA performance assessment.

The psychometric analyses support that AIPI is a valid and reliable clinical instrument for rating the performance of ChatGPT in the management of real clinical cases. The internal consistency, test–retest reliability, interrater reliability, and internal validity reported adequate values, which corroborate the findings of other clinical performance assessment tools [8–11]. In many studies, the practitioner performances were assessed with the mini-clinical evaluation exercise (Mini-CEX), which is a formative assessment tool designed to provide feedback on practitioner skills [10, 14, 15]. The test–retest reliability of Mini-CEX ranged from 0.24 to 0.76, while studies reported good interrater reliability with an intra-class correlation coefficient (ICC) ranging from 0.57 to 0.83 [10, 15]. Similar ICC values were found for the APTA clinical performance instrument, which is dedicated to the assessment of the performance of physical therapists or assistants [9]. Indeed, the Task Force for the Development of Student Clinical Performance Instruments reported adequate internal consistency ($\alpha > 0.70$) and good intraclass coefficients (ICC) for the APTA performance assessment in physical examination (ICC = 0.30), management plan (ICC = 0.49), or selection of additional tests/measurements (ICC = 0.61), which are similar outcomes than those found in AIPI [9]. Moreover, the APTA coefficients for test–retest reliability ranged from 0.81 to 0.96 [9], which corroborates the results obtained for AIPI items, sub- and total scores. In the present study, we used OCAT items for the assessment of convergent validity. Our choice was made despite the possibilities of similar AI clinical instruments in the literature. Rekman et al. showed that OCAT scores were significantly better in experienced residents compared to not experienced residents, suggesting a high internal validity [9]. In the present study, we observed that AIPI sub- and total scores were significantly higher in Humans compared to ChatGPT clinical case evaluation. The internal validity analysis was particularly interesting, because we observed that the consideration of symptoms and physical scores for the establishment of differential diagnoses were significantly similar between senior otolaryngologists and ChatGPT. In practice, the judges reported that ChatGPT differential diagnoses and primary diagnoses were plausible in 58–69%, and 56–71% of cases, respectively, while only 22% of treatments were judged as pertinent and necessary. These findings may suggest that the current version of ChatGPT functions more as an electronic encyclopedia providing a potential list of differential diagnoses and additional examinations, rather than a virtual practitioner considering the patient features. The proposition of a neck MRI in a patient with a pacemaker

Table 2 ChatGPT performance

AIPI outcomes	ChatGPT	OTO (CT)	*p* value
1. Medical and Surgical History	1.53 ± 0.76	1.88 ± 0.33	0.045
2. Symptoms	1.91 ± 0.29	1.96 ± 0.20	NS
3. Physical examinations	1.82 ± 0.39	1.96 ± 0.20	NS
Patient feature score	5.27 ± 0.89	5.81 ± 0.57	0.003
4. Differential diagnoses	2.13 ± 0.87	2.46 ± 0.51	NS
5. Primary diagnosis	2.18 ± 0.91	2.81 ± 0.40	0.003
6. Management plan	0.40 ± 0.49	0.88 ± 0.33	0.001
Diagnostic score	4.71 ± 1.87	6.15 ± 0.78	0.001
7. Additional examinations	1.31 ± 0.79	2.35 ± 0.49	0.001
8. Most relevant additional examination	0.51 ± 0.89	0.81 ± 0.40	0.002
Additional examination score	1.82 ± 1.47	3.15 ± 0.73	0.001
9. Treatment	1.60 ± 0.88	2.73 ± 0.45	0.001
10. AIPI total score	13.33 ± 3.75	17.84 ± 1.76	0.001

AIPI Artificial Intelligence Performance Instrument, *CT* control, *OTO* otolaryngologists

Table 3 Test–retest reliability

AIPI outcomes	r_s	*p* value
1. Medical and Surgical History	0.792	0.001
2. Symptoms	0.999	0.001
3. Physical examinations	0.999	0.001
Patient feature score	0.648	0.001
4. Differential diagnoses	0.750	0.001
5. Primary diagnosis	0.544	0.011
6. Management plan	0.596	0.004
Diagnostic score	0.741	0.001
7. Additional examinations	0.626	0.002
8. Most relevant additional examination	0.791	0.001
Additional examination score	0.850	0.001
9. Treatment	0.850	0.001
10. AIPI total score	0.486	0.035

AIPI Artificial Intelligence Performance Instrument

Table 4 Interrater reliability of AIPI

AIPI outcomes	Kendall	*p* value
1. Medical and Surgical History	0.409	0.005
2. Symptoms	0.261	NS
3. Physical examinations	0.190	NS
Patient feature score	0.268	0.045
4. Differential diagnoses	0.412	0.002
5. Primary diagnosis	0.563	0.001
6. Management plan	0.299	0.047
Diagnostic score	0.491	0.001
7. Additional examinations	0.191	NS
8. Most relevant additional examination	0.366	0.015
Additional examination score	0.338	0.009
9. Treatment	0.952	0.001
10. AIPI total score	0.538	0.001

The interrater reliability analysis was carried out with Kendall tau

NS non-significant

(patient number 19, Appendix 1) was a blatant example of this issue. The theoretical performance of ChatGPT in otolaryngology head and neck surgery was supported in two recent studies. Hoch et al. observed that ChatGPT correctly answered 57% of 2576 theoretical questions related to the otolaryngology subspecialties [16]. Chiesa-Estomba et al. investigated the level of agreement between ChatGPT and 10 international sialendoscopists aiming the capabilities of Chat-GPT to further improve the management of salivary gland disorders. The authors reported a significant agreement between ChatGPT and experts in the clinical decision-making process within the salivary gland clinic, which supports the theoretical performance of ChatGPT [17].

The clinical findings highlighted in the accuracy analysis (Table 5) are important for medical student, resident, and fellow students, because our results suggested that ChatGPT information/recommendations need to be considered with precautions, keeping in mind that the human discernment of the practitioner is not yet acquired by chatbot systems. The same may be applied to patients. Indeed, according to the mediatization of ChatGPT performance, it is conceivable that the number of patients who will use the chatbot system before a practitioner consultation will increase in the next few months [21]. The findings of the present study may support the development of information and prevention policies to avoid the misuse of AI by patients.

The primary strength of the present study was its originality. Indeed, AIPI was developed in time, because the investigations of the ChatGPT performance in the management of real ear, nose, and throat clinical cases are still ongoing, and the use of a valid and reliable clinical instrument may

Table 5 Accuracy of ChatGPT Judged by Otolaryngologists

AIPI management outcomes	Judge 1 *N* (%)	Judge 2 *N* (%)
Differential diagnosis		
Complete or incomplete but plausible	26 (58)	31 (69)
Incomplete and not plausible	19 (42)	14 (31)
Absent	0 (0)	0 (0)
Primary diagnosis		
Correct or plausible	25 (56)	32 (71)
Not plausible	13 (29)	12 (27)
Absent	7 (15)	1 (2)
Additional examinations		
Pertinent and full or partial necessary	13 (29)	13 (29)
Association of pertinent, necessary, and inadequate	30 (67)	28 (62)
Association of inadequate examinations	2 (4)	4 (9)
The most relevant additional examination	11 (24)	15 (33)
Treatment		
Pertinent and necessary	10 (22)	10 (22)
Pertinent but incomplete	7 (16)	7 (16)
Association of pertinent, necessary, and inadequate	27 (60)	26 (58)
Inadequate	1 (2)	2 (4)

improve the research quality. Ear, nose, and throat symptoms and findings concern 10–55% of primary care consultations [18, 19] and up to 30% of visits to emergency departments [20]. Thus, AIPI may be used in other specialties, including general medicine or emergency, and, therefore, may be investigated for validity and reliability in other fields.

The primary limitation of this study was the low number of clinical cases and the low correlation coefficient in the convergent validity. The low convergent validity may be explained by the use of a modified version of OCAT, which was validated for human–practitioner performance only. However, our choice was limited, because there is no other AI performance tool available in the literature.

Conclusion

The AIPI is a reliable and valid AI performance tool that may be used to assess ChatGPT performance in clinical practice. The findings of the present study supported that ChatGPT appears more efficient in diagnosis, rather than in the selection of the most adequate additional examination and the proposition of pertinent and necessary therapeutic approaches. Future clinical studies are needed to assess the usefulness of AIPI in other medical fields regarding the high prevalence of ear, nose and throat disorders in medicine and surgery.

Appendix 1

See Table 6

Table 6 Clinical case features and ChatGPT results

						Otolaryngologist consultation findings		
N	G	Age	Symptoms	History/medication	Clinical examination	Additional examinations	Diagnosis	Treatment
1	F	33	Left cervical painful mass (3 months)	Asthma	Submandibular mass	Neck US, MRI and biology	Salivary lithiasis	NSAID, pilocarpine, Sialadenoscopy
2	M	65	Hearing loss Throat clearing, globus (6 months)	External ear stenosis, GERD	Bilateral total EED stenosis, laryngeal Inflammation	Audiometry (bone) Ear CT	Bilateral EED Stenosis acute Suspected LPR	Canaloplasty Diet, stress reduction, PPI/alginate
3	M	22	Left hearing loss, Tinnitus, throat clearing, Globus, cough (6 months)	Recurrent LPR Recurrent otitis media	Bilateral ear retraction pocket, laryngopharyngeal inflammation	Audiometry, tympanometry, nasopharyngeal pH testing	Chronic otitis Media, recurrent suspected LPR	Nasal saline irrigation, corticoids, diet, stress reduction, PPI/alginate
4	F	71	Sudden smell loss, globus, Dry eyes, sticky mucus,	COVID-19	Dry eyes, coated tongue, laryngopharyngeal	Psychophysical Evaluations	Postviral OD Suspected LPR	Olfactory cleft PRP Injection, diet, stress

Table 6 (continued)

N	G	Age	Symptoms	History/medication	Clinical examination	Otolaryngologist consultation findings		
						Additional examinations	Diagnosis	Treatment
			Throat clearing (7 months)		Inflammation			Reduction, PPI/alginate
5	M	39	Recurrent throat	Nasopharyngeal	Mulberry turbinate,	*Normal sinus CT*	Recurrent/	Drug change: Magaldrate
			Clearing, postnasal drip,	reflux (Restech)	and hypertrophy	*Nasopharyngeal*	chronic LPR	To alginate, continue
			Sticky mucus (> 3 years)		Laryngeal inflammation	*Reflux*		diet and stress reduction
6	M	75	Nasal Congestion,	Nasopharyngeal	Laryngopharyngeal	*Normal sinus CT*	Nasopharyngeal	Diet, stress reduction,
			heartburn, dysphonia	reflux, (Restech)	Hypersensitivity and	*Nasopharyngeal*	Reflux	PPI/alginate, nasal saline
			(> 12 months)		inflammation	*reflux*		Irrigation and corticoids
7	F	24	Globus, throat clearing,	None	Tongue tonsil	HEMII-pH testing	LPR	Diet, stress reduction,
			Abdominal pain, postnasal		Hypertrophy, laryngo-	*Negative allergy test*		PPI/alginate
			Drip/sticky mucus (2 years)		Pharyngeal inflammation			
8	F	40	Dysphonia, globus,	Suspected LPR	Vocal fold erythema	Voice quality	Suspected LPR	Diet, stress reduction,
			throat pain (6 months)		Laryngeal inflammation	assessment		PPI/alginate
9	F	53	Dysphonia, dysphagia,	Ehlers Danlos	Coated/tongue, tonsil	Voice quality	Suspected LPR	Diet, stress reduction,
			throat clearing,		Hypertrophy, laryngo-	Assessment		PPI/alginate
			throat mucus (> 1 year)		pharyngeal inflammation			
10	F	24	Dysphonia, dysphagia,	Tonsillectomy	Vocal cord nodules,	Voice quality	Vocal cord nodules	Diet, stress reduction,
			throat sticky mucus	Vocal cord	Laryngopharyngeal	assessment	Suspected chronic	PPI/alginate,
			(> 12 months)	nodules	inflammation		LPR	Speech therapy
11	F	65	Hypoacousia, dysphonia,	Recurrent chronic	Adenoid hypertrophy,	Audiometry,	Chronic otitis media,	Diet, stress reduction,
			dysphagia,	Otitis media	chronic otitis media,	Tympanometry, voice	LPR, Eustachian tube	PPI/alginate, nasal saline
			Sticky mucus (> 9 months)		laryngeal inflammation	Quality assessment	Dysfunction	Irrigation and corticoids
12	F	54	Dysphagia,	Breast cancer,	Inferior turbinate	Voice quality	Eustachian tube	Diet, stress reduction,
			globus, heartburn	COPD, hypo-	hypertrophy, laryngo-	assessment, audiometry,	Dysfunction,	PPI/alginate
			tinnitus (> 15 months)	thyroidism	pharyngeal inflammation	Tympanometry	suspected LPR	
13	M	67	Cough, throat pain,	Nonacid LPR	Coated tongue,	*HEMII-pH:*	LPR	Diet, stress reduction,
			postnasal drip,	(HEMII-pH)	tonsil erythema,	nonacid LPR		alginate only
			globus (7 months)		laryngeal inflammation			

Table 6 (continued)

N	G	Age	Symptoms	History/medication	Clinical examination	Otolaryngologist consultation findings		
						Additional examinations	Diagnosis	Treatment
14	M	53	Dysphonia, cough,	Septoplasty,	Postnasal drip	*Nasopharyngeal*	LPR	Diet, stress reduction,
			sticky mucus,	Nonacid naso-	Laryngopharyngeal	*pH testing: non-acid*		alginate only
			throat clearing (24 months)	pharyngeal reflux	inflammation	*nasopharyngeal reflux*		
15	F	62	Dry mouth, sticky	Recurrent	Sticky mucus,	Biology: positive	Resistant LPR	Diet, stress reduction,
			mucus, cough, globus	Suspected LPR	tongue tonsil edema	for Chlamydia	to PPI, infectious	alginate, antibiotics
			follow-up (> 6 months)	Aspecific laryngitis	Laryngeal inflammation	Pneumonia	laryngitis	(clarithromycin)
16	M	27	Globus, dysphonia, sticky	Hearth insufficiency	Left septal deviation	*Normal sinus CT*	Recurrent/	Diet, stress reduction,
			mucus, left nasal obstruction,	Ineffective	Laryngopharyngeal	*Nonacid naso-*	chronic nonacid	alginate only
			halitosis (> 19 months)	PPI-therapy	inflammation	*pharyngeal reflux*	LPR	
17	F	53	Chronic hoarseness,	Tobacco	Bilateral Reinke edema	Voice quality	Reinke edema	Stop tobacco,
			throat clearing, globus,	overuse (30 PY)	(grade III), laryngo-	assessment		In-office laser surgery,
			sticky mucus (> 4 years)		pharyngeal inflammation			speech therapy
18	M	51	Dysphonia, suspicion	Crohn, COVID-19	Left vocal fold polyp	Voice quality	Left vocal fold	In-office laser polyp
			of vocal fold paralysis,	Suspected LPR	Laryngopharyngeal	assessment	polyp	surgery, speech therapy,
			globus, throat clearing (6 months)		inflammation		Suspected LPR	diet/stress, alginate
19	F	61	Right parotid tumor,	Gastritis	Right parotid mass	Neck MRI and CT	Parotid lympho-	Imaging and cytology
			progressive growth	HIV, pacemaker		Cytology (US)	epithelial cyst	
			(6 months)					
20	F	32	Sudden dysphonia	Voice professional	Right vocal cord	Voice quality	Vocal cord	In-office laser
			after crying (1-w)		hemorrhage	assessment	hemorrhage	cauterization
21	M	56	Right neck mass,	Alcohol/tobacco	Right piriform sinus	Neck CT, PetCT,	Hypopharyngeal	Oncological board
			weight loss (10 kg)	overuses	exophytic mass	biopsy, biology and	primary carcinoma	discussion
			dysphagia (6 months)	(30 years)		nutrition check-up		
22	F	36	20 kg loss after a diet,	None	Glottal insufficiency	Voice quality	Glottis insufficiency	Speech therapy,
			dysphonia, voice			assessment		vocal cord
			fatigue (3 months)					augmentation
23	F	32	Dysphonia post-	Thyroidectomy	Right vocal cord	Voice quality	Vocal cord	Medialization,
			thyroidectomy (1 month)	for goiter	paralysis	assessment	paralysis	speech therapy
24	M	56	Recurrent laryngeal cancer	Alcohol/tobacco	Persistent carcinoma	PetCT and biopsy:	Laryngeal carcinoma	Salvage laryngectomy

Table 6 (continued)

N	G	Age	Symptoms	History/medication	Clinical examination	Otolaryngologist consultation findings		
						Additional examinations	Diagnosis	Treatment
			after primary chemoradiation	overuses	5 months after the treatment	resistant carcinoma	resistant to	
			(cT3 carcinoma)				chemoradiation	
25	F	66	cT3 supraglottic cancer,	Radiotherapy for	Epiglottis carcinoma	Neck CT, PetCT	Second supraglottic	Salvage surgery
			Weight loss (6 kg),	supraglottic cancer		*Biopsy: carcinoma*	carcinoma	
			Dysphagia	(10 years), hypertension				
26	F	49	Aspirations, cough,	None	Coated tongue, normal	Videofluoroscopy	Suspected LPR	Diet, stress reduction,
			globus, throat,		FEES, laryngeal			PPI/alginate
			sticky mucus (9 months)		inflammation			
27	F	50	Chronic cough, negative	None	Laryngopharyngeal	*HEMII-pH testing:*	Laryngeal	Amitriptyline, GABA
			pH testing, normal		hypersensitivity	negative	hypersensitivity	pentin, or superior
			pulmonary examinations					laryngeal nerve infiltration
28	F	36	Dysphonia, voice	Asthma, inhaled	Vocal fold dryness,	Voice quality	Laryngitis post-	Stop inhaled corticoids/
			fatigue (6 months)	corticosteroids	sticky mucus	assessment	inhaled	change drugs
				(9 months)			corticosteroids	
29	M	66	Bilateral vocal cord paralysis	Thyroid cancer	Bilateral vocal cord	Neck CT scan	Bilateral vocal cord	Bilateral CO2
			postthyroidectomy, tracheotomy,	Thyroidectomy	paralysis in adduction		paralysis	anterior crico-
			Wish for decannulation	Tracheotomy				arytenoidectomy
30	M	70	Bilateral odynophagia,	None	Bilateral stylohyoid	Neck CT scan	Eagle syndrome	Transoral robotic
			otalgia (6 months)		calcified ligaments			styloidectomy
31	F	66	Recurrent dysphagia,	Resistant LPR	Telangiectasia	Manometry, GI,	CREST syndrome	Vasodilators,
			globus, weight loss,	to PPI, alginate,	(fingers), laryngeal	biology (immun),	Esophageal	immunosuppressant
			telangiectasia (3 years)	magaldrate	inflammation	biopsy	scleroderma	
32	F	34	Dysphonia, arthralgia,	None	Orange nodules	Voice quality	Bamboo nodes	Corticoids,
			voice professional		on vocal cord	assessment, biology	Rheumatoid	speech therapy
			(> 12 mo)			(autoimmun), biopsy	polyarthritis	
33	M	40	Progressive dyspnea when	None	Left laryngeal ventricle	Neck CT	Laryngocele	Surgery
			playing trumpet, neck mass,		hypertrophy, left			
			dysphagia (9 months)		neck mass			
34	M	70	Dysphagia, globus,	Cervical arthro-	FEES: normal	Videofluoroscopy	Arthrodesis-related	Speech therapy
			throat pain (1 year)	desis (1 year), diabetes, hypertension		Neck CT	dysphagia (iatrogenic)	(swallowing)

Table 6 (continued)

N	G	Age	Symptoms	History/medication	Clinical examination	Otolaryngologist consultation findings		
						Additional examinations	Diagnosis	Treatment
35	F	36	Dysphonia, throat pain	Vocal cord nodule	Lack of vibration	Voice quality	Vocal fold scars	Speech therapy,
			Voice professional	surgery (12 mo)	of vocal cord	assessment		resection of scars,
			(12 mo)					PRP injection
36	F	41	Sudden dysphonia	Diabetes, burnout	Normal cough,	Voice quality	Psychogenic	Speech therapy,
			(12 months)		aphonia, NFN	assessment	dysphonia	psychotherapy
37	F	30	Recurrent throat pain,	Tonsil abscess	Grade III tonsils	-	Recurrent tonsil	Tonsillectomy
			fever and lymphadenopathy,	(2 times) treated			infections	
			chronic dysphagia (5 years)	with antibiotics				
38	M	20	Left tonsil ulceration	Oral sexual	Left tonsil ulceration	Biology (sexual	Syphilis	Antibiotics
			(3 months)	practice		diseases), biopsy		
						and culture		
39	F	38	Dysphonia, dysphagia,	Thyroidectomy	Normal vocal cord	HEMII-pH testing	Suspected LPR	Diet, stress reduction,
			cough, globus, sticky	Diabetes, arthrosis	mobility, laryngeal	Voice quality		PPI/alginate
			mucus (4 years)		inflammation	assessment		
40	F	45	Singer with difficulty	Thyroidectomy	Normal vocal cord	Voice quality	Superior laryngeal	Speech therapy
			to reach high-pitch	(12 month), hip	mobility, hyposensitivity	assessment	nerve injury	
			sounds (6 month)	prosthesis (2 years)	right tongue base		during surgery	
41	M	20	Left deafness (1 m)	None	Left cerumen earwax	Audiometry	Ear cerumen block	Removal earwax
42	M	75	Progressive bilateral		Normal	Audiometry	Presbycusis	Hearing aids
			deafness (2 years)					
43	F	45	Acute nasal obstruction	Septoplasty (3 days)	Nasal septal hematoma	Sinus CT	Nasal septal	Surgical drainage
			Nasal pain	Hypertension			Hematoma	
44	F	34	Postnasal drip, sneezing	Type 1 diabetes	Inflammatory nasal turbine	Skin prick test	Allergic rhinitis	Antihistamines,
			(April, yearly)		mucosa, sneezing	Sinus CT		Nasal corticosteroids
45	F	30	Dizziness, duration: 2 s,	Hypertension,	Normal	-	Benign paroxysmal	Vestibular rehabilitation,
			nausea (occasionally)	Cholesterolemia			positional vertigo	maneuvers

The additional examinations in italics consisted of results of examination at the consultation time

COPD chronic obstructive pulmonary disease, *COVID-19* coronavirus disease 2019, *CRS(w)NP* chronic rhinosinusitis (without) nasal polyposis, *CT* computed tomography, *EMG* electromyography, *FEES* fiberoptic endoscopic evaluation of swallowing, *FESS* functional endoscopic sinus surgery, *EED* external ear duct, *ETD* Eustachian tube dysfunction, *GERD* gastroesophageal reflux disease, *GI* gastrointestinal endoscopy, *LPRD* laryngopharyngeal reflux disease, *MRI* magnetic resonance imaging, *NFN* normal nasofibroscopy, *NSAID* non-steroidal anti-inflammatory drug, *OD* olfactory dysfunction, *OSAS* obstructive sleep apnea syndrome, *PPI* proton pump inhibitors, *PRP* platelet-rich plasma, *PY* pack/year, *US* ultrasonography

Appendix 2

See Table 7

Table 7 ChatGPT findings regarding clinical cases

N	Otolaryngologist	ChatGPT		
	Diagnosis	Additional examination	Differential diagnosis	Treatment regarding the clinical presentation
1	Salivary lithiasis	Neck US, MRI, Biology, Prick skin tests	Adenitis, Abscess, Adenopathy, Parotitis, Thyroiditis	Hot compress, pilocarpine, Surgery
2	Bilateral EED stenosis acute suspected LPR	Ear CT pH metry, GI, throat bacteriology	EED stenosis, Chronic otitis media, Presbycusis, ETD, GERD, CRS, allergic rhinitis, LPR, chronic tonsilitis	Canaloplasty, balloon dilatation, saline irrigation, Skin flap, PPIs, H2 blockers, Diet/Stress management
3	Chronic otitis media, recurrent suspected LPR	Audiometry, Tympanometry pH metry, GI, throat bacteriology	Chronic otitis media (effusion/suppurative), cholesteatoma, tympanosclerosis	Nasal corticoids or transtympanic tube, Allergy checkup
4	Postviral OD Suspected LPR	Psychophysical evaluations, sinus X-Ray, Neck CT, Biology (B12 level, CRP)	Postviral OD, CRSNP, CRSwNP, Neurodegenerative disease, Autoimmun disease	PPIs, H2 blockers, Diet/Stress management
5	Recurrent/ chronic LPR	pH metry, GI, neck CT, sinus/throat cultures	GERD, unspecified CRS, allergic rhinitis, LPR, Chronic tonsilitis	PPI and H2 blockers Diet and stress management
6	Nasopharyngeal reflux	RAST (IgE), nasal cytology (eosinophiles), sinus CT	LPR, allergic rhinitis, CRSNP	speech therapy PPI and H2 blockers
7	LPR	Sinus CT, sinus X-Ray, biology (CRP)	Acute sinusitis, GERD, upper aerodigestive tract infection, COPD, viral bronchitis	PPI and H2 blockers Diet and stress management
8	Suspected LPR	Sinus CT, X-Ray, Biology (CRP)	Acute sinusitis, GERD, laryngopharyngeal infection, COPD, viral bronchitis	PPI and H2 blockers Diet and stress management
9	Suspected LPR	Neck CT, prick test, Genetics for Ehlers Danlos swallowing study, GI	LPR, chronic pharyngitis, vocal cord dysfunction, eosinophilic esophagitis, oral candidosis, tongue tonsillitis	Speech therapy, PPI and H2 blockers Diet and stress management
10	Vocal cord nodules Suspected chronic LPR	Neck CT, voice quality assessment, allergy testing	Vocal cord nodules, vocal cord polyps, GERD, LPR, chronic laryngitis	Corticoids, voice rest, PPI and H2 blockers Diet and stress management
11	Chronic otitis media, LPR, Eustachian tube dysfunction	Throat swab culture, audiometry, swallowing study, GI	Laryngitis, chronic pharyngitis, otitis media, GERD,	Speech therapy, PPI and H2 blockers, diet and stress management, GERD-surgery
12	Eustachian tube dysfunction, suspected LPR	Neck CT, thyroid function tests	LPR, unspecified CRS, hypothyroidism, bronchitis, cancer related symptoms (unspecified ?)	PPI and H2 blockers, diet and stress management, nasal decongestants, antibiotics and corticoids
13	LPR	Neck CT, throat swab, allergy testing	Chronic tonsilitis, chronic pharyngitis, GERD, LPR	PPI and H2 blockers Diet and stress management
14	LPR	Neck CT, throat cultures, allergy tests	Acute laryngitis, chronic laryngitis, GERD, COPD, vocal cord nodules	PPI and H2 blockers Diet and stress management
15	Resistant LPR to PPI, infectious	Neck CT, salivary gland function test,	CRSwNP, chronic tonsilitis, GERD, OSAS, Sjögren syndrome	PPI and H2 blockers, stress reduction, diet, antibiotics

Table 7 (continued)

N	Otolaryngologist	ChatGPT		
	Diagnosis	Additional examination	Differential diagnosis	Treatment regarding the clinical presentation
	laryngitis	Allergy, polysomnography		
16	Recurrent/ chronic nonacid LPR	Esophageal manometry, pH Metry, allergy	LPR, GERD, CRSwNP, chronic tonsilitis, postnasal drip	PPI and H2 blockers, stress reduction, diet, nasal saline irrigation and corticosteroids
17	Reinke edema	Vocal cord biopsy, stroboscopy, lung testing (spirometry)	Reinke edema, vocal cord dysfunction, vocal cord polyps, bilateral vocal cord nodules	Vocal hygiene, speech therapy, smoking cessation, surgery
18	Left vocal fold polyp Suspected LPR	Neck CT Laryngeal biopsy	LPR, CRSwNP, Allergic rhinitis, vocal fold polyp, vocal fold nodules	Polypectomy/resection of mass, corticoids, speech therapy
19	Parotid lympho-epithelial cyst	Neck US, CT, cytology, biology (CRP)	Pleiomorphic or Whartin tumor, malignancy, lymph node, metastasis	Surgery
20	Vocal cord hemorrhage	None	Vocal cord hemorrhage	Voice rest, anti-inflammatory drugs, vocal cord abuse reduction
21	Hypopharyngeal primary carcinoma	Biopsy, neck CT or MRI	Laryngeal or pharyngeal carcinoma, metastasis, benign mass, granulomatosis, pseudotumor	Biopsy, neck CT or MRI
22	Glottis insufficiency	None	Muscle atrophy, vocal cord paresis, psychogenic dysphonia, vocal fold lesions	Biology (autoimmun diseases), speech therapy, psychological support
23	Vocal cord paralysis	EMG, Neck CT and MRI	Recurrent laryngeal nerve injury	Speech therapy, medialization, thyroplasty, reinnervation
24	Laryngeal carcinoma resistant to chemoradiation	Neck CT or MRI, oncological board assessment	Persistent carcinoma	Salvage surgery, immunotherapy, clinical trials, palliative
25	Second supraglottic carcinoma	Neck CT or MRI, biopsy	Supraglottic carcinoma	Radiation, chemotherapy or chemoradiotherapy, surgery, immunotherapy
26	Suspected LPR	pH testing, EMG	LPR, chronic laryngitis, laryngeal hypersensitivity, postnasal drip, chronic tonsilitis	Diet, stress reduction, mucolytics, nasal corticoids, antihistamine, PPIs, H2 blockers, laryngeal desensitization (breath)
27	Laryngeal hypersensitivity	pH testing, manometry, allergy, inhaled broncho-dilators, methacholine test	Cough variant asthma, GERD, LPR	PPIs, inhaled corticoids, speech therapy
28	Laryngitis post-inhaled corticosteroids	Lung assessment	Vocal cord dysfunction, Reinke edema, vocal cord nodules, vocal cord polyps, muscle tension dysphonia	Continue asthma treatment
29	Bilateral vocal cord paralysis	–	Bilateral vocal cord paralysis (adduction) and permanent tracheotomy	Vocal cord lateralization, speech therapy, electrolarynx,
30	Eagle syndrome	Neck CT or MRI	Tonsilloliths, tonsillar hypertrophy, chronic pharyngitis, GERD	PPIs, H2 blockers, diet/Stress management, saltwater gargles, good oral hygiene/hydration, tonsillectomy
31	CREST syndrome scleroderma	GI, biology	Scleroderma, GERD	HE blockers, prokinetic, esophageal dilatation
32	Bamboo nodes	–	Reinke edema, vocal cord polyps	Speech therapy, vocal cord surgery

Table 7 (continued)

N	Otolaryngologist	ChatGPT		
	Diagnosis	Additional examination	Differential diagnosis	Treatment regarding the clinical presentation
	Rheumatoid arthritis			(Removal of lesion)
33	Laryngocele	Neck CT or MRI, biopsy	Laryngeal papillomatosis, laryngeal cancer, laryngocele, lymphadenopathy	Surgery, radiation, voice therapy
34	Arthrodesis-related dysphagia (iatrogenic)	Esophageal manometry, videofluoroscopy, endoscopic US, biopsy	Achalasia, esophageal stricture, esophageal spasm, esophageal cancer	Calcium channel blockers, nitrates, or botulinum toxin injections, esophageal dilatation, surgery or chemo-radiation
35	Vocal fold scars	Neck CT or MRI, EMG	Vocal fold scar, granuloma, Reinke edema, vocal cord paresis,	Speech therapy, corticoids, surgery of the lesion
36	Psychogenic dysphonia	EMG and psychological evaluation	Functional dysphonia, muscle tension dysphonia, conversion disorder, vocal fold paresis or spasmodic dystonia	Speech therapy, psychotherapy, muscle relaxants
37	Recurrent tonsil infections	Throat culture, Neck CT or MRI, PSG	Chronic tonsillitis, tonsillar hypertrophy, tonsilloliths	Antibiotics, tonsillectomy
38	Syphilis	Throat culture, Neck CT or MRI, biopsy	Bacterial or viral tonsillitis, tonsil abscess, HPV infection, cancer	Antibiotics, analgesics
39	Suspected LPR	pH study, GI, allergy evaluation	LPR, chronic laryngitis, allergic rhinitis	PPIs, H2 blockers, diet/ Stress management, speech therapy
40	Superior laryngeal nerve injury during surgery	EMG, Neck CT and MRI, neurological consultation	Superior laryngeal nerve dysfunction, hypoglossal dysfunction, vocal cord muscle atrophy	Speech therapy, nerve reconstruction
41	Ear cerumen block	Audiometry Tympanometry	Cerumen earwax	Removal
42	Presbycusis	Audiometry Tympanometry	Presbycusis, sensorineural hearing loss	Hearing aids, Assistive listening devices, lip reading and speech therapy
43	Nasal septal Hematoma	–	Postoperative edema	Nasal decongestants, irrigation, corticoids
44	Allergic rhinitis	Allergy testing, rhino-manometry, nasal smear	Allergic rhinitis, non-allergic rhinitis	Avoiding triggers, antihistamines, nasal corticoids, saline irrigation, immuno-therapy
45	Benign paroxysmal positional vertigo	Audiometry, electro-nystagmography	Benign paroxysmal positional vertigo	Vestibular rehabilitation, maneuvers

COPD chronic obstructive pulmonary disease, *COVID-19* coronavirus disease 2019, *CRS(w)NP* chronic rhinosinusitis (without) nasal polyposis, *CT* computed tomography, *EMG* electromyography, *FEES* fiberoptic endoscopic evaluation of swallowing, *FESS* functional endoscopic sinus surgery, *EED* external ear duct, *ETD* Eustachian tube dysfunction, *GERD* gastroesophageal reflux disease, *GI* gastrointestinal endoscopy, *LPRD* laryngopharyngeal reflux disease, *MRI* magnetic resonance imaging, *NFN* normal nasofibroscopy, *NSAID* non-steroidal anti-inflammatory drug, *OD* olfactory dysfunction, *OSAS* obstructive sleep apnea syndrome, *PPI* proton pump inhibitors, *PRP* platelet-rich plasma, *PY* pack/year, *US* ultrasonography

Appendix 3

See Table 8

Table 8 Convergent validity details

AIPI outcomes	OCAT outcomes			Total score
	Differential diagnoses	Management plan	Therapeutic approach	
1. Medical and Surgical History	0.223 (NS)	0.126 (NS)	0.016 (NS)	0.046 (NS)
2. Symptoms	0.052 (NS)	0.174 (NS)	0.024 (NS)	0.055 (NS)
3. Physical examinations	0.444 (0.004)	0.403 (0.010)	0.320 (0.044)	0.465 (0.002)
Patient feature score	0.376 (0.017)	0.061 (NS)	0.125 (NS)	0.498 (0.001)
4. Differential diagnoses	0.449 (0.004)	0.065 (NS)	0.223 (NS)	0.299 (NS)
5. Primary diagnosis	0.519 (0.001)	0.018 (NS)	0.105 (NS)	0.251 (NS)
6. Management plan	0.280 (NS)	0.109 (NS)	0.003 (NS)	0.145 (NS)
Diagnostic score	0.569 (0.001)	0.113 (NS)	0.172 (NS)	0.128 (NS)
7. Additional examinations	0.093 (NS)	0.010 (NS)	0.130 (NS)	0.100 (NS)
8. Most relevant additional examination	0.052 (NS)	0.027 (NS)	0.052 (NS)	0.035 (NS)
Additional examination score	0.270 (NS)	0.023 (NS)	0.141 (NS)	0.151 (NS)
9. Treatment	0.150 (NS)	0.328 (0.044)	0.244 (NS)	0.292 (NS)
10. AIPI total score	0.495 (0.001)	0.101 (NS)	0.204 (NS)	0.319 (0.045)

The Pearson coefficient is provided with the *p* value

AIPI Artificial Intelligence Performance Instrument, *NS* non-significant, *OCAT* Ottawa Clinical Assessment Tool

Author contributions JRL: design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. AM: design, acquisition of data, data analysis and interpretation, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. IG: design, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. LAV: design, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. SH: design, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. CMC-E: design, acquisition of data, data analysis and interpretation, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding None.

Data availability Data are available on request to the first author after authorization of the University.

Declarations

Conflict of interest The authors have no conflicts of interest.

Ethics declarations The guest editor was not involved with the peer review process of this article.

Informed consent Patients consented to participate to the study.

References

1. Pernencar C, Saboia I, Dias JC (2022) How far can conversational agents contribute to IBD patient health care-a review of the literature. Front Public Health 10:862432. https://doi.org/10.3389/fpubh.2022.862432
2. Wahlster W (2023) Understanding computational dialogue understanding. Philos Trans A Math Phys Eng Sci 381(2251):20220049. https://doi.org/10.1098/rsta.2022.0049
3. Hill-Yardin EL, Hutchinson MR, Laycock R, Spencer SJ (2023) A Chat(GPT) about the future of scientific publishing. Brain Behav Immun 110:152–154. https://doi.org/10.1016/j.bbi.2023.02.022

4. Choi JH, Hickman KE, Monahan A, Schwarcz D (2023) ChatGPT goes to law school? Minnesota legal studies research paper No. 23-03
5. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M (2023) The Pros and Cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 305:644–647. https://doi.org/10.3233/SHTI230580
6. https://futureoflife.org/open-letter/pause-giant-ai-experiments/
7. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM (2023) ChatGPT performance in laryngology and head & neck surgery: a clinical case-series. Eur Arch Otorhinolaryngol
8. Rekman J, Hamstra SJ, Dudek N, Wood T, Seabrook C, Gofton W (2016) A new instrument for assessing resident competence in surgical clinic: the Ottawa clinic assessment tool. J Surg Educ 73(4):575–582. https://doi.org/10.1016/j.jsurg.2016.02.003
9. Task Force for the Development of Student Clinical Performance Instruments, American Physical Therapy Association (2002) The development and testing of APTA clinical performance instruments. Phys Ther 82(4):329–353
10. Chen YY, Chiu YC, Chu TS, Hsu HY, Chen HL, Wu CC, Huang TS (2022) Is the rating result reliable? A new approach to respond to a medical trainee's concerns about the reliability of Mini-CEX assessment. J Formos Med Assoc 121(5):943–949. https://doi.org/10.1016/j.jfma.2021.07.005
11. Jubraj B, Patel S, Naseem I, Copp S, Karagkounis D (2017) The acute care assessment tool: pharmacy ACAT. Clin Teach 14:184e8
12. Braun LT, Lenzer B, Fischer MR, Schmidmaier R (2019) Complexity of clinical cases in simulated learning environments: proposalfor a scoring system. GMS J Med Educ 36(6):80. https://doi.org/10.3205/zma001288
13. Gercama AJ, de Haan M, van der Vleuten CPM (2000) Reliability of the Amsterdam clinical challenge scale (ACCS): a new instrument to assess the level of difficulty of patient cases in medical education. Med Educ 34(7):519–524
14. Lee V, Brain K, Martin J (2017) Factors influencing mini-CEX rater judgments and their practical implications: a systematic literature review. Acad Med 92(6):880–887. https://doi.org/10.1097/ACM.0000000000001537
15. Kogan JR, Holmboe ES, Hauer KE (2009) Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA 302(12):1316–1326. https://doi.org/10.1001/jama.2009.1365
16. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, Cotofana S, Alfertshofer M (2023) ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol. https://doi.org/10.1007/s00405-023-08051-4
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. Hayois L, Dunsmore A (2023) Common and serious ENT presentations in primary care. InnovAiT 16(2):79–86. https://doi.org/10.1177/17557380221140131
19. Hannaford PC, Simpson JA, Bisset AF, Davis A, McKerrow W, Mills R (2005) The prevalence of ear, nose and throat problems in the community: results from a national cross-sectional postal survey in Scotland. Fam Pract 22(3):227–233. https://doi.org/10.1093/fampra/cmi004
20. Vasileiou I, Giannopoulos A, Klonaris C, Vlasis K, Marinos S, Koutsonasios I, Katsargyris A, Konstantopoulos K, Karamoutsos C, Tsitsikas A, Marinos G (2009) The potential role of primary care in the management of common ear, nose or throat disorders presenting to the emergency department in Greece. Qual Prim Care 17(2):145–148
21. Millstein J, Agarwal A (2023) What can doctors and patients do with ChatGPT? | Expert Opinion. Philadelphia Inquirer

Authors and Affiliations

Jerome R. Lechien[1,2,3,4,5,11] · Antonino Maniaci[1,6] · Isabelle Gengler[1,7] · Stephane Hans[1,4] · Carlos M. Chiesa-Estomba[1,2,8] · Luigi A. Vaira[1,9,10]

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

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

[2] Young Confederation of the European Oto-Rhino-Laryngological Head and Neck Surgery Societies (Y-CEORLHNS), Dublin, Ireland

[3] 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

[4] Phonetics and Phonology Laboratory (UMR 7018 CNRS, Université Sorbonne Nouvelle/Paris 3), Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, School of Medicine, UFR Simone Veil, Université Versailles Saint-Quentin-en-Yvelines (Paris Saclay University), Paris, France

[5] Department of Otorhinolaryngology and Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium

[6] Department of Medical, Surgical Sciences and Advanced Technologies G.F. Ingrassia, ENT Section, University of Catania, 95123 Catania, Italy

[7] Department of Otolaryngology-Head and Neck Surgery, University of Cincinnati Medical Center, Cincinnati, OH, USA

[8] Department of Otorhinolaryngology – Head and Neck Surgery, Donostia University Hospital – Biodonostia Research Institute, St. Sebastian, Spain

[9] Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy

[10] Biomedical Science Department, Biomedical Science PhD School, University of Sassari, Sassari, Italy

[11] Faculty of Medicine, Department of Human Anatomy and Experimental Oncology, UMONS Research Institute for Health Sciences and Technology, Avenue du Champ de Mars, 6, B7000 Mons, Belgium