Validity and Reliability of An Instrument Evaluating the Performance of Intelligent Chatbot: The Artificial Intelligence Performance Instrument (AIPI).

Objective: 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 (α=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.

Conclusion: 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.

Key words: 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 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 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 (Figure 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.

Setting and Clinical Cases

Fifty clinical cases of outpatients consulting in the Departments of Otolaryngology-Head & 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 (Figure 2). The guidelines consisted of the scientific position paper/recommendations of the European and American Societies in Otolaryngology-Head & Neck Surgery.

The data of the consultation were presented to ChatGPT without mentioning the human differential diagnoses, 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; Figure 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; Figure 2).

Results:

Forty-five patients completed the consultation (Figure 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 (α=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 was 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% to 42% and 27% to 29% of cases, respectively (Table 5). Judges reported that additional examinations proposed by ChatGPT were associated with pertinent, necessary, and inadequate examinations in 62% to 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 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 (α>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% to 69%, and 56% to 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 (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 2,576 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 improve the research quality. Ear, nose, and throat symptoms and findings concern 10 to 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.

Acknowledgments: None.

References:

1.Pernencar C, Saboia I, Dias JC. How Far Can Conversational Agents Contribute to IBD Patient Health Care-A Review of the Literature. *Front Public Health*. 2022; 10:862432. doi: 10.3389/fpubh.2022.862432.

2. Wahlster W. Understanding computational dialogue understanding. *Philos Trans A Math Phys Eng Sci*. 2023; 381(2251):20220049. doi: 10.1098/rsta.2022.0049.

3.Hill-Yardin EL, Hutchinson MR, Laycock R, Spencer SJ. A Chat(GPT) about the future of scientific publishing. *Brain Behav Immun*. 2023; 110:152-154. doi: 10.1016/j.bbi.2023.02.022.

4. Choi JH, Hickman KE, Monahan A, Schwarcz D. ChatGPT Goes to Law School ? *Minnesota Legal Studies Research Paper No. 23-03* ; 2023.

5. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M.The Pros and Cons of Using ChatGPT in Medical Education: A Scoping Review. Stud Health Technol Inform. 2023; 305:644-647. doi: 10.3233/SHTI230580.

6. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

7. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. ChatGPT Performance in Laryngology and Head & Neck Surgery: A Clinical Case-Series. *Eur Arch Otorhinolaryngol*. 2023.

8. Rekman J, Hamstra SJ, Dudek N, Wood T, Seabrook C, Gofton W. A New Instrument for Assessing Resident Competence in Surgical Clinic: The Ottawa Clinic Assessment Tool. *J Surg Educ*. 2016; 73(4):575-82. doi: 10.1016/j.jsurg.2016.02.003.

9. Task Force for the Development of Student Clinical Performance Instruments. The development and testing of APTA Clinical Performance Instruments. American Physical Therapy Association. *Phys Ther*; 2002; 82(4):329-53.

10. Chen YY, Chiu YC, Chu TS, Hsu HY, Chen HL, Wu CC, Huang TS. 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*. 2022 ; 121(5):943-949. doi: 10.1016/j.jfma.2021.07.005.

11. Jubraj B, Patel S, Naseem I, Copp S, Karagkounis D. The acute care assessment tool: 'pharmacy ACAT. *Clin Teach* 2017;14: 184e8.

12. Braun LT, Lenzer B, Fischer MR, Schmidmaier R. Complexity of clinical cases in simulated learning environments: proposalfor a scoring system. *GMS J Med Educ*. 2019; 36(6):Doc80. doi: 10.3205/zma001288.

13. Gercama AJ, de Haan M, van der Vleuten CPM. 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. 2000;* 34(7):519–524.

14. Lee V, Brain K, Martin J. Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications: A Systematic Literature Review. *Acad Med*. 2017; 92(6):880-887. doi: 10.1097/ACM.0000000000001537.

15. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA*. 2009; 302(12):1316-26. doi: 10.1001/jama.2009.1365.

16. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, Cotofana S, Alfertshofer M. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *Eur Arch Otorhinolaryngol.* 2023. doi: 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. Exploring the potential of Chat-GPT as a supportive

tool for sialendoscopy clinical decision making and patient information support. *Eur Arch Otorhinolaryngol*. 2023. doi: 10.1007/s00405-023-08104-8

18. Hayois L, Dunsmore A. Common and serious ENT presentations in primary care. *InnovAiT*. 2023;16(2):79-86. doi:10.1177/17557380221140131

19. Hannaford PC, Simpson JA, Bisset AF, Davis A, McKerrow W, Mills R. The prevalence of ear, nose and throat problems in the community: results from a national cross-sectional postal survey in Scotland. *Fam Pract.* 2005; 22(3):227-33. doi: 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. 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*. 2009;17(2):145-8.

21. Millstein J, Agarwal A.What can doctors and patients do with ChatGPT? | Expert Opinion. Philadelphia Inquirer. 2023.

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)

Table 1 footnotes: Abbreviations: SD=standard deviation.

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

Table 2 footnotes: Abbreviations: AIPI= Artificial Intelligence Performance Instrument; CT=control; OTO=otolaryngologists.

Table 3: Test-retest reliability.

AIPI Outcomes	rs	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

Table 3 footnotes: Abbreviations: 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

Table 4 footnotes: The interrater reliability analysis was carried out with Kendall tau. Abbreviations: NS=non significant.

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)

Table 5 footnotes: -.

Figure 1: The Artificial Intelligence Performance Instrument.

Figure 1 footnotes: AIPI score ranges from 0 (inadequate management) to 20 (adequate management).

Figure 2: Chart flow.

Figure 2 footnotes: Abbreviations: OCAT: Ottawa Clinic Assessment Tool: OTO=otolaryngologist.

Appendix 1: 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-mo)	Asthma	Submandibular mass	Neck US, MRI and biology	Salivary lithiasis	NSAID, pilocarpine, sialadenoscopy
2	M	65	Hearing loss Throat clearing, globus (6-mo)	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-mo)	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, throat clearing (7-mo)	COVID-19	Dry eyes, coated tongue, Laryngopharyngeal inflammation	Psychophysical evaluations	Postviral OD Suspected LPR	Olfactory cleft PRP injection,diet, stress reduction, PPI/alginate
5	M	39	Recurrent throat clearing, postnasal drip, sticky mucus (>3-year)	Nasopharyngeal reflux (Restech)	Mulberry turbinate, & hypertrophy Laryngeal inflammation	*Normal sinus CT* *Nasopharyngeal Reflux*	Recurrent/ chronic LPR	Drug change: Magaldrate to alginate, continue diet and stress reduction.
6	M	75	Nasal Congestion, heartburn, dysphonia (>12-mo)	Nasopharyngeal reflux, (Restech)	Laryngopharyngeal hypersensitivity & inflammation.	*Normal sinus CT* *Nasopharyngeal reflux*	Nasopharyngeal reflux	Diet, stress reduction, PPI/alginate, nasal saline irrigation & corticoids.
7	F	24	Globus, throat clearing, Abdominal pain, postnasal drip/sticky mucus (2-y)	None	Tongue tonsil hypertrophy, laryngopharyngeal inflammation	HEMII-pH testing *Negative allergy test*	LPR	Diet, stress reduction, PPI/alginate
8	F	40	Dysphonia, globus, throat pain (6-mo)	Suspected LPR	Vocal fold erythema Laryngeal inflammation	Voice quality assessment	Suspected LPR	Diet, stress reduction, PPI/alginate
9	F	53	Dysphonia, dysphagia, throat clearing,	Ehlers Danlos	Coated/tongue, tonsil hypertrophy, laryngo-	Voice quality assessment	Suspected LPR	Diet, stress reduction, PPI/alginate

			throat mucus (>1-y)		pharyngeal inflammation			
10	F	24	Dysphonia, dysphagia, throat sticky mucus (>12-mo)	Tonsillectomy Vocal cord nodules	Vocal cord nodules, Laryngopharyngeal inflammation	Voice quality assessment	Vocal cord nodules Suspected chronic LPR	Diet, stress reduction, PPI/alginate, Speech therapy
11	F	65	Hypoacousia, dysphonia, dysphagia, Sticky mucus (>9-mo)	Recurrent chronic otitis media	Adenoid hypertrophy, chronic otitis media, laryngeal inflammation	Audiometry, Tympanometry, voice quality assessment	Chronic otitis media, LPR, Eustachian tube dysfunction	Diet, stress reduction, PPI/alginate, nasal saline irrigation & corticoids
12	F	54	Dysphagia, globus, heartburn tinnitus (>15-mo)	Breast cancer, COPD, hypothyroidism	Inferior turbinate hypertrophy, laryngopharyngeal inflammation	Voice quality assessment, audiometry, Tympanometry	Eustachian tube dysfunction, suspected LPR	Diet, stress reduction, PPI/alginate
13	M	67	Cough, throat pain, postnasal drip, globus (7-mo)	Nonacid LPR (HEMII-pH)	Coated tongue, tonsil erythema, laryngeal inflammation	*HEMII-pH:* nonacid LPR	LPR	Diet, stress reduction, alginate only
14	M	53	Dysphonia, cough, sticky mucus, throat clearing (24-mo)	Septoplasty, Nonacid nasopharyngeal reflux	Postnasal drip Laryngopharyngeal inflammation	*Nasopharyngeal pH testing: nonacid nasopharyngeal reflux*	LPR	Diet, stress reduction, alginate only
15	F	62	Dry mouth, sticky mucus, cough, globus follow-up(>6-mo)	Recurrent suspected LPR Aspecific laryngitis	Sticky mucus, tongue tonsil edema Laryngeal inflammation	Biology: positive for Chlamydia Pneumonia	Resistant LPR to PPI, infectious laryngitis	Diet, stress reduction, alginate, antibiotics (clarithromycin)
16	M	27	Globus, dysphonia, sticky mucus, left nasal obstruction, halitosis (>19-mo)	Hearth insufficiency Ineffective PPI-therapy	Left septal deviation Laryngopharyngeal inflammation	*Normal sinus CT* *Nonacid nasopharyngeal reflux*	Recurrent/ chronic nonacid LPR	Diet, stress reduction, alginate only
17	F	53	Chronic hoarseness, throat clearing, globus, sticky mucus (>4-y)	Tobacco overuse (30 PY)	Bilateral Reinke edema (grade III), laryngopharyngeal inflammation	Voice quality assessment	Reinke edema	Stop tobacco, In-office laser surgery, speech therapy
18	M	51	Dysphonia, suspicion of vocal fold paralysis,	Crohn, COVID-19 Suspected LPR	Left vocal fold polyp Laryngopharyngeal	Voice quality assessment	Left vocal fold polyp	In-office laser polyp surgery, speech therapy,

			globus, throat clearing (6-mo)		inflammation		Suspected LPR	diet/stress, alginate
19	F	61	Right parotid tumor, progressive growth (6-mo)	Gastritis HIV, pacemaker	Right parotid mass	Neck MRI and CT Cytology (US)	Parotid lympho-epithelial cyst	Imaging and cytology
20	F	32	Sudden dysphonia after crying (1-w)	Voice professional	Right vocal cord hemorrhage	Voice quality assessment	Vocal cord hemorrhage	In-office laser cauterization
21	M	56	Right neck mass, weight loss (10 kg) dysphagia (6-mo)	Alcohol/tobacco overuses (30 years)	Right piriform sinus exophytic mass	Neck CT, PetCT, biopsy, biology & nutrition check-up	Hypopharyngeal primary carcinoma	Oncological board discussion
22	F	36	20 kg loss after a diet, dysphonia, voice fatigue (3-mo)	None	Glottal insufficiency	Voice quality assessment	Glottis insufficiency	Speech therapy, vocal cord augmentation
23	F	32	Dysphonia post-thyroidectomy (1 mo)	Thyroidectomy for goiter	Right vocal cord paralysis	Voice quality assessment	Vocal cord paralysis	Medialization, speech therapy
24	M	56	Recurrent laryngeal cancer after primary chemoradiation (cT3 carcinoma)	Alcohol/tobacco overuses	Persistent carcinoma 5-mo after the treatment	PetCT and biopsy: resistant carcinoma	Laryngeal carcinoma resistant to chemoradiation	Salvage laryngectomy
25	F	66	cT3 supraglottic cancer, Weight loss (6 kg), Dysphagia	Radiotherapy for supraglottic cancer (10-y), hypertension	Epiglottis carcinoma	Neck CT, PetCT *Biopsy: carcinoma*	Second supraglottic carcinoma	Salvage surgery
26	F	49	Aspirations, cough, globus, throat, sticky mucus (9-mo)	None	Coated tongue, normal FEES, laryngeal inflammation	Videofluoroscopy	Suspected LPR	Diet, stress reduction, PPI/alginate
27	F	50	Chronic cough, negative pH testing, normal pulmonary examinations	None	Laryngopharyngeal hypersensitivity	*HEMII-pH testing:* negative	Laryngeal hypersensitivity	Amitriptyline, GABA pentin, or superior laryngeal nerve infiltration
28	F	36	Dysphonia, voice fatigue (6-mo)	Asthma, inhaled corticosteroids	Vocal fold dryness, sticky mucus	Voice quality assessment	Laryngitis post-inhaled	Stop inhaled corticoids/ change drugs

				(9-mo)			corticosteroids	
29	M	66	Bilateral vocal cord paralysis postthyroidectomy, tracheotomy, Wish for decannulation	Thyroid cancer Thyroidectomy Tracheotomy	Bilateral vocal cord paralysis in adduction	Neck CT scan	Bilateral vocal cord paralysis	Bilateral CO2 anterior crico-arytenoidectomy
30	M	70	Bilateral odynophagia, otalgia (6-mo)	None	Bilateral stylo-hyoid calcified ligaments	Neck CT scan	Eagle syndrome	Transoral robotic styloidectomy
31	F	66	Recurrent dysphagia, globus, weight loss, telangiectasia (3-y)	Resistant LPR to PPI, alginate, magaldrate	Telangiectasia (fingers), laryngeal inflammation	Manometry, GI, biology (immun), biopsy	CREST syndrome Esophageal scleroderma	Vasodilators, immunosuppressant
32	F	34	Dysphonia, arthralgia, voice professional (>12 mo)	None	Orange nodules on vocal cord	Voice quality assessment, biology (autoimmun), biopsy	Bamboo nodes Rheumatoid polyarthritis	Corticoids, speech therapy
33	M	40	Progressive dyspnea when playing trumpet, neck mass, dysphagia (9-mo)	None	Left laryngeal ventricle hypertrophy, left neck mass	Neck CT	Laryngocele	Surgery
34	M	70	Dysphagia, globus, throat pain (1-y)	Cervical arthrodesis (1-y), diabetes, hypertension	FEES: normal	Videofluoroscopy Neck CT	Arthrodesis-related dysphagia (iatrogenic)	Speech therapy (swallowing)
35	F	36	Dysphonia, throat pain Voice professional (12 mo)	Vocal cord nodule surgery (12 mo)	Lack of vibration of vocal cord	Voice quality assessment	Vocal fold scars	Speech therapy, resection of scars, PRP injection
36	F	41	Sudden dysphonia (12-mo)	Diabetes, burnout	Normal cough, aphonia, NFN	Voice quality assessment	Psychogenic dysphonia	Speech therapy, psychotherapy
37	F	30	Recurrent throat pain, fever and lymphadenopathy, chronic dysphagia (5-y)	Tonsil abscess (2 times) treated with antibiotics	Grade III tonsils	-	Recurrent tonsil infections	Tonsillectomy
38	M	20	Left tonsil ulceration	Oral sexual	Left tonsil ulceration	Biology (sexual	Syphilis	Antibiotics

			(3-mo)	practice		diseases), biopsy & culture		
39	F	38	Dysphonia, dysphagia, cough, globus, sticky mucus (4-y)	Thyroidectomy Diabetes, arthrosis	Normal vocal cord mobility, laryngeal inflammation	HEMII-pH testing Voice quality assessment	Suspected LPR	Diet, stress reduction, PPI/alginate
40	F	45	Singer with difficulty to reach high-pitch sounds (6-mo)	Thyroidectomy (12-mo), hip prosthesis (2-y)	Normal vocal cord mobility, hyposensitivity right tongue base	Voice quality assessment	Superior laryngeal nerve injury during surgery	Speech therapy
41	M	20	Left deafness (1-m)	None	Left cerumen earwax	Audiometry	Ear cerumen block	Removal earwax
42	M	75	Progressive bilateral deafness (2-y)		Normal	Audiometry	Presbycusis	Hearing aids
43	F	45	Acute nasal obstruction Nasal pain	Septoplasty (3-d) Hypertension	Nasal septal hematoma	Sinus CT	Nasal septal Hematoma	Surgical drainage
44	F	34	Postnasal drip, sneezing (April, yearly)	Type 1 diabetes	Inflammatory nasal turbine mucosa, sneezing	Skin prick test Sinus CT	Allergic rhinitis	Antihistamines, Nasal corticosteroids
45	F	30	Dizziness, duration: 2s, nausea (occasionally)	Hypertension, Cholesterolemia	Normal	-	Benign paroxysmal positional vertigo	Vestibular rehabilitation, maneuvers

Appendix 1 footnotes: The additional examinations in italics consisted of results of examination at the consultation time. Abbreviations: 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: ChatGPT findings regarding clinical cases.

	Otolaryngologist		ChatGPT	
N	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,	LPR, chronic pharyngitis, vocal cord	Speech therapy, PPI and H2 blockers

		Genetics for Ehlers Danlos swallowing study, GI	dysfunction, eosinophilic esophagitis, oral candidosis, tongue tonsillitis	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 laryngitis	Neck CT, salivary gland function test, Allergy, polysomnography	CRSwNP, chronic tonsilitis, GERD, OSAS, Sjögren syndrome	PPI and H2 blockers, stress reduction, diet, antibiotics
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 Rheumatoid arthritis	-	Reinke edema, vocal cord polyps	Speech therapy, vocal cord surgery (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 chemoradiation
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, rhinomanometry, nasal smear	Allergic rhinitis, non-allergic rhinitis	Avoiding triggers, antihistamines, nasal corticoids, saline irrigation, immunotherapy
45	Benign paroxysmal positional vertigo	Audiometry, electronystagmography	Benign paroxysmal positional vertigo	Vestibular rehabilitation, maneuvers

Appendix 2 footnotes: Abbreviations: 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: Convergent validity details.

	OCAT outcomes			
AIPI outcomes	Differential Diagnoses	Management Plan	Therapeutic Approach	Total Score
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)

Appendix 3 footnotes: The Pearson coefficient is provided with the p-value. Abbreviations: AIPI=Artificial Intelligence Performance Instrument; NS=non-significant; OCAT=Ottawa Clinical Assessment Tool.