Received: 17 August 2023 | Revised: 6 December 2023 | Accepted: 28 December 2023

DOI: 10.1002/alr.23323

IFAR: International Forum of Allergy & Rhinology

RESEARCH NOTE

ChatGPT-4 performance in rhinology: A clinical case series

Thomas Radulesco MD, PhD, MS[1,2] | **Alberto Maria Saibene MD, MA**[1,3] | **Justin Michel MD, PhD, MS**[2] | **Luigi Angelo Vaira MD**[1,4,5] | **Jérôme R. Lechien MD, PhD, MS**[1,6,7,8]

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

[2]Aix Marseille University, APHM, CNRS, IUSTI, La Conception University Hospital, ENT-HNS Department, Marseille, France

[3]Otolaryngology Unit, ASST Santi Paolo E Carlo, Department of Health Sciences, Università Degli Studi Di Milano, Milan, Italy

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

[5]PhD School of Biomedical Sciences, Department of Biomedical Sciences, University of Sassari, Sassari, Italy

[6]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

[7]Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, Paris Saclay University, Phonetics and Phonology Laboratory (UMR 7018 CNRS, Université Sorbonne Nouvelle/Paris 3), Paris, France

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

Correspondence
Jérôme R. Lechien, 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.
Email: Jerome.lechien@umons.ac.be

KEYWORDS
artificial intelligence, ChatGPT-4, head neck surgery, otolaryngology, performance, rhinology

Key points

- Chatbot Generative Pre-trained Transformer (ChatGPT)-4 indicated more than twice additional examinations than practitioners in the management of clinical cases in rhinology.
- The consistency between ChatGPT-4 and practitioner in the indication of additional examinations may significantly vary from one examination to another.
- The ChatGPT-4 proposed a plausible and correct primary diagnosis in 62.5% cases, while pertinent and necessary additional examinations and therapeutic regimen were indicated in 7.5%–30.0% and 7.5%–32.5% of cases, respectively.
- The stability of ChatGPT-4 responses is moderate-to-high. The performance of ChatGPT-4 was not influenced by the human-reported level of difficulty of clinical cases.

Luigi Angelo Vaira and Jérôme R. Lechien contributed equally and may be joined as co-senior authors.

1 | INTRODUCTION

The Chatbot Generative Pre-trained Transformer (ChatGPT) uses algorithms to respond to simple-to-complicated questions.[1] ChatGPT could be theoretically used as an adjunctive clinical tool for helping practitioners in the management of diseases.[1,2] To date, few studies investigated the performance and accuracy of GPT in the management of real clinical cases in rhinology. The objective of this study was to investigate the performance of ChatGPT-4 in the management of rhinologic cases.

2 | METHODS

Patients were consecutively recruited from the Departments of Otolaryngology of CHU Saint-Pierre (Brussels, Belgium) from July 2 to August 16, 2023. The following data were collected from the consultation: history, complaints, comorbidities, medication, physical examination (tympanoscopy, nasofibroscopy, oral and neck examination), additional examinations, and treatments. Among them, the following information were manually entered a posteriori into the Application Programming Interface (API) (https://chat.openai.com/) by a senior practitioner: symptoms, history, comorbidities, current medication, previous medications related to the condition, clinical findings, and additional examinations (in case of previous additional examinations). ChatGPT-4 was interrogated for the patient management with the following questions: *What are your primary diagnosis and differential diagnoses?*; *What are your additional examinations to make the diagnosis?*; and *What are your treatment(s) for the primary diagnosis?* The rhinologist conclusions of the consultation were not entered into the API, including the additional examinations and treatments. The ChatGPT-4 responses were assessed by three independent experienced board-certified rhinologists who established the appropriate management (agreement) using national and international consensus statements.[3–5] The complexity of clinical cases was evaluated at the end of the consultation by the main practitioner with a modified version of the General Items off the Amsterdam Clinical Challenge Scale (ACCS) test.[6]

2.1 | ChatGPT-4 performance and stability

Artificial intelligence performance instrument (AIPI) was used for the assessment of ChatGPT-4 performance.[2] AIPI is a validated and reliable instrument for the assessment of performance and accuracy of artificial intelligence (AI) chatbots.[2] AIPI includes nine items assessing patient history, symptoms, physical examination, diagnosis, additional examinations, management plan, and treatments (Appendix 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 four following subscores associating common items: patient feature score (/6), diagnosis score (/7), additional examination score (/5), and treatment score (/3). AIPI provides a comprehensive approach to otolaryngological cases, and reported high intra- and interrater reliabilities.[2] ChatGPT-4 results were regenerated five times (5-day interval) to evaluate stability of responses through a consistency analysis.

The study was approved by the IRB of CHU Saint-Pierre (no. BE0762023230708). Patient consented to participate.

2.2 | Statistical analyses

Statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (v.22.0; IBM Corp.). The additional examination indications by ChatGPT-4 and practitioner were compared with Mann–Whitney *U*-test, and consistency analysis (*kappa* analysis). The judge consistency (interrater reliability) was assessed for AIPI scores with Kendall tau. The stability of ChatGPT-4 regenerated responses was assessed with Fliess kappa analysis.

3 | RESULTS

The data of 40 patients with rhinologic conditions were presented to ChatGPT-4 (25 females; Appendix 2). The mean age of patients was 49.8 ± 17.1 years. The mean ACCS was 14.7 ± 5.5. Sixty-eight and 132 additional examinations were indicated by rhinologists (mean = 1.7 ± 1.1 per patient) and ChatGPT-4 (mean = 3.28 ± 0.78), respectively (p = 0.001, Table 1). The indication of allergy testing, psychophysical olfactory assessment, audiometry/tympanometry, positron emission tomography–computerized tomography, rhinomanometry, dental check-up, and biopsy were significantly consistent between ChatGPT-4 and rhinologists (Table 1). The total number of additional examinations indicated by ChatGPT-4 was poorly but significantly associated with the ACCS examination sub-score (r_s = 0.384; p = 0.014), meaning that practitioner difficulty to perform the

TABLE 1 Additional examination consistency.

	Main additional examinations			
	OTO	ChatGPT	Kappa	*p*-Value
Sinus tomodensitometry	22	30	0.158	NS
Rhinomanometry	9	2	0.307	0.007
Psychophysical olfactory testing	8	9	0.627	0.001
Audiometry	5	4	0.875	0.001
Tympanometry	5	3	0.724	0.001
PET-CT	2	1	0.655	0.001
Skin prick test/RAST	2	13	0.197	0.037
Biopsy	2	8	0.348	0.004
Sinus magnetic resonance imaging	1	11	0.127	NS
Bacterial sample and culture	1	11	0.127	NS
Neck tomodensitometry	1	5	0.043	NS
Dental assessment	1	2	0.655	0.001
Total number (Mann–Whitney *p*-value)	68	132	–	0.001

Note: A total of 68 and 132 examinations were indicated by OTO and ChatGPT. Only the additional examinations commonly indicated by OTO and ChatGPT-4 at least once time are reported in this table. ChatGPT-4 never indicated some additional examinations, including pH-impedance monitoring, polysomnography, and lung assessment.

Abbreviations: ChatGPT, Chatbot Generative Pre-trained Transformer; NS, non-significant; OTO, otolaryngologists; PET-CT, positron emission tomography–computerized tomography; RAST, Radioallergosorbent test.

clinical examination was associated with more additional examinations indicated by ChatGPT-4.

3.1 | ChatGPT-4 performance

The judges reported adequate interrater reliability in the ChatGPT-4 performance assessment (Appendix 3). According to AIPI, ChatGPT-4 was particularly performant in the proposition of primary diagnosis (Table 2 and Appendix 4). The primary diagnosis was correct or plausible in 62.5% of cases (Table 2). ChatGPT-4 proposed pertinent and necessary additional examinations in 6.3%, while the list of proposed additional examinations including pertinent, necessary, and inadequate examinations in 93.7% of cases. The therapeutic propositions were pertinent and necessary in 6.7% of cases. Judges noted that ChatGPT-4 proposed magnetic resonance imaging in a patient with a pacemaker; sinus tomodensitometry in patients who already had sinus tomodensitometry; and pH monitoring in patients who underwent impedance pH monitoring. ChatGPT-4 recommended biopsies for all nasal lesions despite of a clear diagnosis at the imaging (e.g., sinus osteoma, sphenoid mucocele).

3.2 | Stability of responses

ChatGPT-4 reported highest stabilities for therapeutic options and additional examinations (Appendix 5). Some differential diagnoses were not consistently proposed by ChatGPT-4 from the first to the fifth regenerated answers, for example, chronic rhinosinusitis without polyps, acute respiratory infection, and septal deviation/idiopathic turbinate hypertrophy.

4 | DISCUSSION

The primary finding of the present study suggested that ChatGPT-4 may propose correct or plausible primary diagnosis in more than half rhinologic cases, which supports the moderate performance of GPT-4 found in previous otolaryngology studies.[2,7,8] Similarly, the performance of GPT-4 was low-to-moderate in additional examinations propositions. Precisely, ChatGPT-4 proposed a list of potential additional examinations without selecting the most adequate for the clinical situation. Similar observations were found in laryngology,[8] where ChatGPT-4 indicated a higher number of additional examinations compared to practitioners for establishing a diagnosis.[8] The accuracy of ChatGPT-4 for the treatment strategy was disappointing with only 16.7% of necessary and complete strategies according to judges, which corroborate the finding of Ayoub et al. who observed that ChatGPT-4 scored worse that Google Search when providing medical recommendations.[9]

According to the stability of ChatGPT-4, our results support that the GPT-4 stability may vary from one disease to another. In a recent study, Perlis reported some inconsistencies between regenerated ChatGPT-4 answers in the management of depression, especially in the indication of some medications and not others.[10] Although the diagnosis approaches of rhinologic diseases and depression are significantly different, the Perlis's findings and our observations may support a certain amount of instability in the GPT-4 output. The lack of consideration of some key points of the medical history of patient, for example, recent imaging, led us to believe that ChatGPT-4 appeared to identify and apply a number of heuristics commonly applied in rhinologic clinical practice. In other words, ChatGPT-4 functions as a performant encyclopedia proposing potential diagnoses, additional examinations, and treatments without considering some important history findings of patients.

TABLE 2 Performance of Chatbot Generative Pre-trained Transformer (ChatGPT)-4.

AIPI outcomes	Judge 1, *N* (%)	Judge 2, *N* (%)	Judge 3, *N* (%)	Total, mean
Primary diagnosis				
Correct	22 (55.0)	19 (47.5)	20 (50.0)	20.3 (50.8)
Plausible	3 (7.5)	6 (15.0)	6 (15.0)	5.0 (12.5)
Not plausible	14 (35.0)	15 (37.5)	12 (30.0)	13.7 (34.3)
Absent	1 (2.5)	0 (0.0)	2 (5.0)	1.0 (2.5)
Relevant additional examination				
Pertinent and necessary	12 (30.0)	3 (7.5)	4 (10.0)	6.3 (15.8)
Pertinent and not all necessary	9 (22.5)	4 (10.0)	8 (20.0)	7.0 (17.5)
Pertinent, necessary, and inadequate	17 (42.5)	24 (60.0)	26 (65.0)	23.3 (58.3)
Only inadequate examinations	2 (5.0)	9 (22.5)	2 (5.0)	4.3 (10.8)
Treatment regimen				
Pertinent and necessary	13 (32.5)	3 (7.5)	4 (10.0)	6.7 (16.7)
Pertinent and incomplete	12 (30.0)	2 (5.0)	10 (25.0)	8.0 (20.0)
Association of pertinent/necessary and inadequate	13 (32.5)	29 (72.5)	18 (45.0)	20.0 (50.0)
No adequate strategy	2 (5.0)	6 (15.0)	8 (20.0)	5.3 (13.3)

Note: The table describes the ChatGPT-4 performance in primary diagnosis, additional examination propositions, and treatments. The GPT-4 performance was assessed by three judges who evaluated the primary diagnosis, proposition of additional examination, and treatment regimen responses by GPT-4 for each case. The table data consisted of number (%) of AIPI items. The judge assessments were performed in a blind manner regarding the other judge responses.
Abbreviation: AIPI, artificial intelligence performance instrument.

The knowledge about the performance of ChatGPT in the management of real clinical cases is important for practitioners and patients regarding the availability of ChatGPT for patients (prior to the consultation) or young practitioners.

The main strengths of the present study are the originality and consideration of real clinical cases. The moderate-to-high interrater reliability between judges (consistency) is the primary limitation. The judges come from three different countries and, consequently, some disagreements may be related to differences between national guidelines underlying their daily practice. Finally, the inability of GPT-4 to analyze endoscopic video limits its accuracy because many diagnoses are based on image features in rhinology.

5 | CONCLUSION

ChatGPT-4 performance is moderate in rhinologic practice. ChatGPT-4 appears to be more efficient in primary and differential diagnosis establishment than in the proposition of necessary additional examinations and accurate therapeutic regimen. However, the accuracy for diagnostic purposes remains low, which makes controversial the use of ChatGPT-4 in clinical practice.

ACKNOWLEDGMENTS

None.

CONFLICT OF INTEREST STATEMENT

The authors declare they have no conflicts of interest.

FUNDING INFORMATION

The authors received no specific funding for this work.

INFORMED CONSENT

Patients consented to participate.

ORCID

Thomas Radulesco MD, PhD, MS https://orcid.org/0000-0002-5939-5372
Alberto Maria Saibene MD, MA https://orcid.org/0000-0003-1457-6871
Luigi Angelo Vaira MD https://orcid.org/0000-0002-7789-145X

Jérôme R. Lechien MD, PhD, MS https://orcid.org/0000-0002-0845-0845

REFERENCES

1. Yoshiyasu Y, Wu F, Dhanda AK, Gorelik D, Takashima M, Ahmed OG. GPT-4 accuracy and completeness against International Consensus Statement on Allergy and Rhinology: rhinosinusitis. *Int Forum Allergy Rhinol*. 2023;13(12):2231-2234. doi:10.1002/alr.23201
2. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the artificial intelligence performance instrument (AIPI). *Eur Arch Otorhinolaryngol*. 2023. doi:10.1007/s00405-023-08219-y
3. Patel ZM, Holbrook EH, Turner JH, et al. International consensus statement on allergy and rhinology: olfaction. *Int Forum Allergy Rhinol*. 2022;12(4):327-680. doi:10.1002/alr.22929
4. Wise SK, Damask C, Roland LT, et al. International consensus statement on allergy and rhinology: allergic rhinitis—2023. *Int Forum Allergy Rhinol*. 2023;13(4):293-859. doi:10.1002/alr.23090.5
5. Orlandi RR, Kingdom TT, Smith TL, et al. International consensus statement on allergy and rhinology: rhinosinusitis 2021. *Int Forum Allergy Rhinol*. 2021;11(3):213-739. doi:10.1002/alr.22741
6. 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.
7. Chiesa-Estomba CM, Lechien JR, Vaira LA, et al. 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
8. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. Chat-GPT performance in laryngology and head and neck surgery: a clinical case-series. *Eur Arch Otorhinolaryngol*. 2024;281(1):319-333. doi:10.1007/s00405-023-08282-5
9. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of ChatGPT versus google search for medical knowledge acquisition. *Otolaryngol Head Neck Surg*. 2023. doi:10.1002/ohn.465
10. Perlis RH. Research Letter: application of GPT-4 to select next-step antidepressant treatment in major depression. *medRxiv*. 2023. doi:10.1101/2023.04.14.23288595

How to cite this article: Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR. ChatGPT-4 performance in rhinology: A clinical case series. *Int Forum Allergy Rhinol*. 2024;14:1123–1130. https://doi.org/10.1002/alr.23323

APPENDIX 1: Artificial intelligence performance instrument (AIPI).

Outcomes of AIPI	Practitioner evaluation			Item score	Sub-scores
1. Consideration of medical and surgical history in the AI management:	Fully (2)	Partly (1)	Not (0)	…../2	Patient feature score …./6
2. Consideration of symptoms of patients in the AI management	Fully (2)	Partly (1)	Not (0)	…../2	
3. Consideration of physical findings reported by practitioner(s)	Fully (2)	Partly (1)	Not (0)	…../2	
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	Diagnosis score …./7
5. The primary diagnosis of AI was:	Correct (3) Plausible (2) Not plausible (1) Absent (0)			…../3	

Continued.

Outcomes of AIPI	Practitioner evaluation	Continued.	Item score	Sub-scores
6. The management plan of AI included potential physical/additional examinations for determining the diagnosis	Yes (1)	No (0)	…../1	
7. The additional examinations proposed by AI	All pertinent and necessary examinations (3)		…../3	Additional examination score …..…../4
	All pertinent but partially necessary examinations (2)			
	An association of pertinent, necessary, and inadequate examinations (1)			
	An association of inadequate examinations (0)			
8. AI identified the most relevant additional examination to perform first	Yes (1)		…../1	
	No, AI provided a list without stratification (0)			
9. The treatments proposed by AI are/include	All pertinent and necessary therapeutic findings (3)		…../3	Treatment score …..…../3
	All pertinent but incomplete therapeutic findings (2)			
	An association of pertinent, necessary, and inadequate therapeutic findings (1)			
	No adequate therapeutic approach (0)			
				Total AIPI score:……/20

Note: AIPI is a nine-item tool that assesses the performance of an artificial intelligence (AI) regarding its consideration of history, symptoms, physical examination, diagnosis, additional examinations of a patient disease and the related management plan, and treatments. 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 four following sub-scores associating common items: patient feature score (/6), diagnosis score (/7), additional examination score (/4), and treatment score (/3).

APPENDIX 2: Patient symptoms.

Primary diagnoses	
Nasopharyngeal reflux	9
Postviral olfactory disorder	8
Acute/recurrent CRS	4
Eustachian tube dysfunction	3
Septal deviation/idiopathic turbinate hypertrophy	2
Nasopharyngeal cancer	2
Allergic rhinitis	2
CRSNP	2
CRSwNP	1
Chronic otitis media	1
Rhinitis medicamentosa	1
Tornwaldt/nasopharyngeal cyst	1
Sinus mucocele	1
Vasomotor rhinitis	1
Empty nose syndrome	1

Primary diagnoses	
Obstructive sleep apnea syndrome	1
Inverted papilloma	1
Fungal ball	1
Posttraumatic anosmia	1
Nasal foreign body	1
Sinus osteoma	1
Juvenile nasal angiofibroma	0
Septal abscess	0

Note: Five patients had several primary diagnoses.
Abbreviations: CRS, chronic rhinosinusitis; CRSNP, chronic rhinosinusitis with nasal polyposis; CRSwNP, chronic rhinosinusitis without nasal polyps.

APPENDIX 3: Interrater reliability.

AIPI outcomes	Judge 1 versus 2		Judge 1 versus 3		Judge 2 versus 3	
	Kendall	*p*-Value	Kendall	*p*-Value	Kendall	*p*-Value
Patient feature score	0.536	0.001	0.459	0.011	0.132	NS
Diagnostic score	0.344	0.007	0.740	0.001	0.565	0.002
Additional examination score	0.176	NS	0.621	0.001	0.493	0.017
Treatment	0.358	0.013	0.621	0.001	0.133	NS
AIPI total score	0.522	0.001	0.729	0.001	0.593	0.001

Abbreviations: AIPI, artificial intelligence performance instrument; NS, non-significant.

APPENDIX 4: Performance analysis of Chatbot Generative Pre-trained Transformer (ChatGPT)-4 in diagnostic, examinations, and treatment.

AIPI outcomes, mean (SD)	Judge 1	Judge 2	Judge 3	Mean
1. Medical and surgical history (/2)	1.31 (0.77)	1.26 (0.69)	1.38 (0.67)	1.33 (0.54)
2. Symptoms (/2)	1.40 (0.74)	1.38 (0.71)	1.43 (0.60)	1.40 (0.58)
3. Physical examinations (/2)	1.20 (0.85)	1.00 (0.93)	1.29 (0.64)	1.18 (0.70)
Patient feature score (/6)	3.88 (2.19)	3.58 (1.68)	4.10 (1.64)	3.87 (1.62)
4. Differential diagnoses (/3)	1.78 (0.95)	1.02 (0.80)	1.76 (0.83)	1.45 (0.64)
5. Primary diagnosis (/3)	2.10 (1.06)	2.10 (0.93)	1.91 (1.18)	2.08 (0.95)
6. Management plan (/1)	0.65 (0.48)	0.98 (0.1)	0.62 (0.50)	0.79 (0.26)
Diagnostic score (/7)	4.53 (2.14)	4.13 (1.20)	4.29 (2.00)	4.33 (1.53)
7. Additional examinations (/3)	1.33 (0.92)	1.03 (0.58)	1.33 (0.58)	1.21 (0.60)
8. Most relevant additional examination (/1)	0.50 (0.64)	0.03 (1.16)	0.33 (0.48)	0.28 (0.36)
Additional examination score (/4)	1.83 (1.26)	1.04 (0.64)	1.67 (1.86)	1.49 (0.78)
9. Treatment (/3)	1.90 (0.93)	1.05 (0.71)	1.67 (0.66)	1.49 (0.67)
10. AIPI total score (/20)	12.03 (5.37)	9.83 (3.20)	11.71 (3.89)	11.15 (3.89)

Note: The AIPI score ranges from 0 to 20 and is subdivided into the four following sub-scores: patient feature score (/6), diagnosis score (/7), additional examination score (/4), and treatment score (/3). The three judges reported adequate interrater reliability (Kendall tau).

Abbreviations: AIPI, artificial intelligence performance instrument; SD, standard deviation.

APPENDIX 5: Stability and consistency of Chatbot Generative Pre-trained Transformer (ChatGPT)-4 answers.

	Fleiss kappa	*p*-Value
Differential diagnosis		
Allergic rhinitis	0.439	0.002
Non-allergic rhinitis	1.000	0.001
Eustachian tube dysfunction	0.432	0.002
CRSNP	0.702	0.001
CRSwNP	0.053	NS
Acute respiratory infection	0.250	NS
GERD/LPR	0.432	0.002
Septal deviation/idiopathic turbinate hypertrophy	0.167	NS
Olfactory dysfunction	1.000	0.001
Nasopharyngeal/nasal cancer	0.702	0.001
Additional examinations		
Sinus tomodensitometry	1.000	0.001
Psychophysical olfactory testing	0.053	NS
pH study	0.432	0.002
Audiometry	1.000	0.001
Tympanometry	1.000	0.001
Skin prick test/RAST	0.333	0.018
Bacterial sample and culture	0.219	NS
Sinus MRI	0.826	0.001
Therapeutic approaches		
Saline solution	1.000	0.001
Nasal corticosteroids	0.594	0.001
Antihistamines	1.000	0.001
Decongestant	1.000	0.001
Surgery	0.500	0.001
Olfactory training	1.000	0.001
Antireflux medication (PPI/H2 blockers)	1.000	0.001

Note: The consistency analysis was carried out with Fleiss kappa analysis.

Abbreviations: CRSNP, chronic rhinosinusitis with nasal polyposis; CRSwNP, chronic rhinosinusitis without nasal polyps; GERD, gastroesophageal reflux disease; LPR, laryngopharyngeal reflux disease; MRI, magnetic resonance imaging; PPI, proton pump inhibitors; RAST, Radioallergosorbent test.