European Archives of Oto-Rhino-Laryngology
https://doi.org/10.1007/s00405-023-08326-w

HEAD AND NECK

Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings

Jerome R. Lechien[1,2,3,4,5] · Carlos-Miguel Chiesa-Estomba[1,6] · Robin Baudouin[1,2,3] · Stéphane Hans[1,2,3]

Received: 5 October 2023 / Accepted: 27 October 2023

Abstract

Objectives To evaluate the ChatGPT-4 performance in oncological board decisions.

Methods Twenty medical records of patients with head and neck cancer were evaluated by ChatGPT-4 for additional examinations, management, and therapeutic approaches. The ChatGPT-4 propositions were assessed with the Artificial Intelligence Performance Instrument. The stability of ChatGPT-4 was evaluated through regenerated answers at 1-day interval.

Results ChatGPT-4 provided adequate explanations for cTNM staging in 19 cases (95%). ChatGPT-4 proposed a significant higher number of additional examinations than practitioners (72 versus 103; $p = 0.001$). ChatGPT-4 indications of endoscopy–biopsy, HPV research, ultrasonography, and PET–CT were consistent with the oncological board decisions. The therapeutic propositions of ChatGPT-4 were accurate in 13 cases (65%). Most additional examination and primary treatment propositions were consistent throughout regenerated response process.

Conclusions ChatGPT-4 may be an adjunctive theoretical tool in oncological board simple decisions.

Keywords Otolaryngology · Head neck surgery · ChatGPT · Decision · Performance · Artificial intelligence · GPT

Introduction

The development of artificial intelligence-powered language model of chatbot is an emerging field in medicine and surgery. These new generations of chatbots may respond to simple-to-complicated questions in all fields of medicine and research, and, consequently are considered as theoretical adjunctive clinical and research tools [1, 2]. To date, the studies investigating the accuracy of Chatbot Generative Pre-trained Transformer (ChatGPT, OpenIA, CA, USA) in theoretical knowledges, medical school examinations, and clinical vignettes, reported encouraging results [3–5]. ChatGPT-4 may be accurate for providing theoretical information, but may be limited when facing to real clinical cases from the otolaryngology consultation [5, 6]. Currently, there is no clinical study investigating the ChatGPT-4 performance in the assessment of real head and neck oncological cases.

The aim of this study was to evaluate the performance of ChatGPT-4 in oncological board decisions in head and neck oncology.

Jerome R. Lechien and Carlos-Miguel Chiesa-Estomba have similarly contributed and are joined as co-first authors.

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

[1] Research Committee of Young-Otolaryngologists of the International Federations of Oto-Rhino-Laryngological Societies (YO-IFOS), Paris, France

[2] Department of Otolaryngology-Head Neck Surgery, Foch Hospital, UFR Simone Veil, University Paris Saclay, Paris, France

[3] Phonetics and Phonology Laboratory (UMR 7018 CNRS, Université Sorbonne Nouvelle/Paris 3), Paris, France

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

[5] Division of Laryngology and Broncho-Esophagology, Department of Otolaryngology—Head and Neck Surgery, EpiCURA Hospital, Baudour, Belgium

[6] Department of Otorhinolaryngology—Head and Neck Surgery, Hospital Universitario Donostia, San Sebastian, Spain

Published online: 22 November 2023

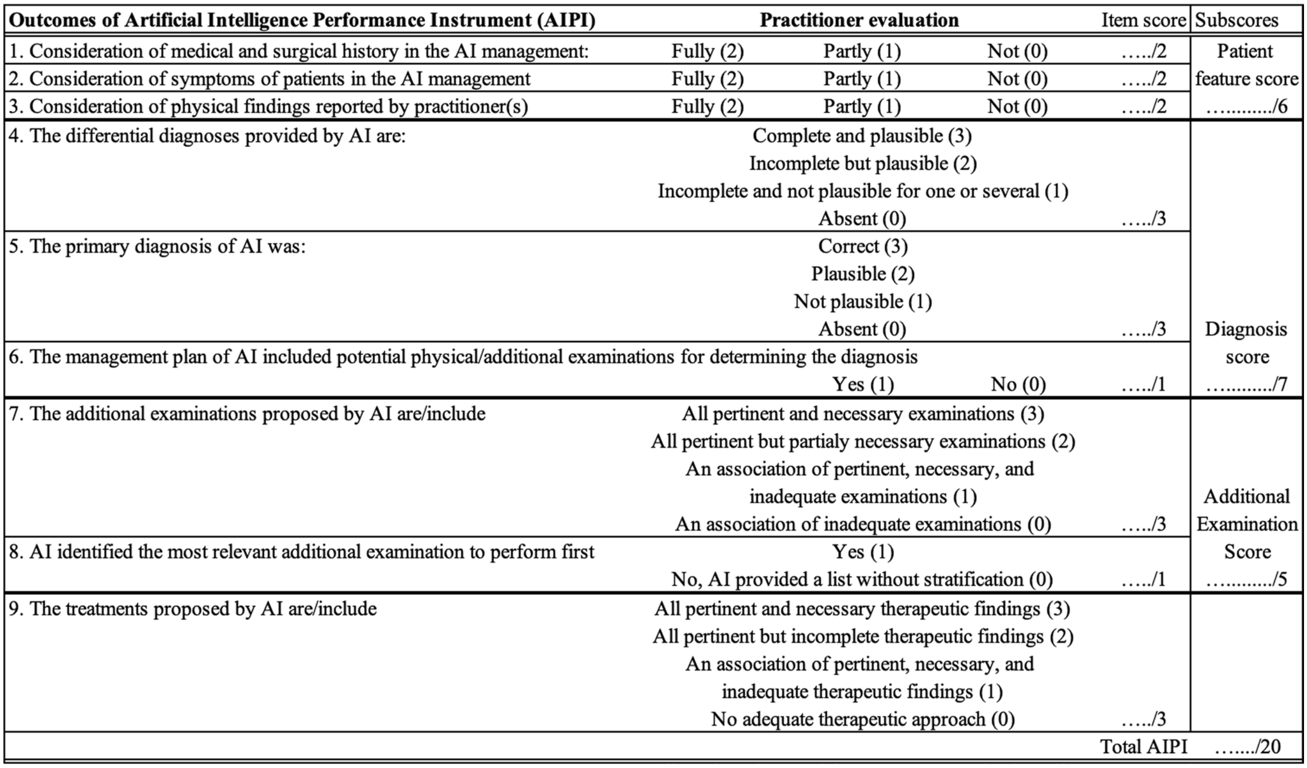
Methods

Patients and setting

The medical record data of 20 patients with head and neck cancer were consecutively collected from the Head and Neck Oncological Boards of the departments of Otolaryngology—Head and Neck Surgery of University Hospital of Brussels in August 2023. Patients were initially addressed to the consultation of the first and the senior authors of the study (J.R.L. and S.H.). The medical records of patients were completed according to clinical history, additional examinations, and pathological diagnosis. Incomplete clinical cases were excluded from the study. All patient cases were discussed in the oncological board and a decision was made according to the Guidelines of the French society of Otorhinolaryngology, and the European Head and Neck Society [7, 8]. The following data were collected: demographics, symptoms, clinical and endoscopic examination, additional examination findings, pathological diagnosis, primary and alternative therapeutic propositions.

Chatbot interrogation and outcomes

ChatGPT-4 was interrogated in two question steps for providing additional examinations, primary and alternative therapeutic strategies through the ChatGPT interface, which is accessible via the API (https://chat.openai.com). The questions were chosen according to the content of the medical records (first step—additional examinations), and the results of additional examinations (second step—treatment). Precisely, ChatGPT-4 was first interrogated for additional examinations (What are the requested additional examinations?), and for therapeutic strategies (According to the following additional examinations…, What are your primary and alternative therapeutic propositions?). The clinical case characteristics are available in Appendix 1. The ChatGPT-4 responses were regenerated at 1-day interval to assess the stability of responses over time. The ChatGPT-4 findings were collected in a database and judged according to the oncological board findings by a panel of two head and neck surgeons (J.RL. and S.H.). Both surgeons used the Artificial Intelligence Performance Instrument (AIPI) to rate the performance of ChatGPT-4. AIPI is a valid and reliable instrument in assessing the performance of chatbots in ear, nose and throat conditions [5] (Fig. 1).



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 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	
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 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 Score …………/5
8. AI identified the most relevant additional examination to perform first	Yes (1) No, AI provided a list without stratification (0)	…../1	
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)

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

Statistical analyses

Statistical analyses were performed through the Statistical Package for the Social Sciences for Windows (SPSS version 24,0; IBM Corp, Armonk, NY, USA). The total number of additional examinations proposed by ChatGPT-4 and practitioners was compared with Mann–Whitney *U* test. The therapeutic decisions of ChatGPT-4 and oncological board were coded in an excel database for a consistency analysis using Kendall tau. The stability of ChatGPT-4 response over time was assessed with kappa analysis. Coefficients were considered as low, moderate, and strong for $r_s < 0.30$, 0.30–0.60, and $r_s > 0.60$, respectively. A level of significance of $p < 0.05$ was used.

Results

Twenty patient medical records were collected and submitted to ChatGPT-4 for additional examinations and therapeutic strategies. The patient, pathological and oncological characteristics are available in Table 1. The mean age of patients was 58.7 ± 9.6 years. There were 8 females (40%). Laryngeal and oropharyngeal squamous cell carcinoma were the most prevalent malignancies accounting for 20% ($N = 4$) and 20% ($N = 4$) of cases, respectively. There were 12 (60%), 7 (35%), and 1 (5%), primary, recurrent, and secondary malignancies, respectively. Details of medical records are available in Appendix 1. Questions and ChatGPT-4 responses are available on request.

Additional examinations

ChatGPT-4 provided adequate explanations for the cTNM stage information in 19 (95%) medical record cases according to the 8th edition of the AJCC/UICC TNM staging system [9]. In one case (patient number 3), ChatGPT-4 defined cT3 as locally advanced tumor size without additional information.

The practitioners and the oncological board indicated 72 additional examinations in patients, corresponding to a mean of 3.60 ± 0.94 per patient. ChatGPT-4 proposed 103 additional examinations (mean 5.15 ± 1.31), which was significantly higher than practitioners ($p = 0.001$). The consistency between oncological board and ChatGPT-4 in the indication of additional examinations is described in Table 2. There were significant strong consistencies between human and ChatGPT-4 for the indications of upper aerodigestive tract endoscopy, and research of human papilloma virus (HPV) infection, while the consistency analysis reported moderate consistencies for the indication of PET–CT, neck ultrasonography, and biopsy.

The mean AIPI score of ChatGPT-4 for additional examination management of ChatGPT-4 is 2.95 ± 0.83 (Appendix 2). According to AIPI, ChatGPT-4 proposed pertinent and necessary examinations in 25% of cases, whereas the additional examinations were judged as pertinent but not all necessary in 55% of cases (Table 3). Among the ChatGPT-4 inadequate propositions, practitioners reported that ChatGPT-4 systematically indicated fine-needle aspiration biopsy when patient had neck node at the clinical examination, and proposed biopsy in all lesions. Thus, ChatGPT-4 did not propose a resection–biopsy, which was indicated by the oncological board for cT1N0M0 vocal fold lesions according to the Guidelines of the European Laryngological Society [10]. Moreover, ChatGPT-4 overall proposed neck CT and MRI for all cancer localization. In 4 cases, ChatGPT-4 indicated a chest X-ray for the detection of lung metastasis.

Therapeutic strategies

The therapeutic options proposed by the oncological board and ChatGPT-4 are summarized in Table 2. There were moderate-to-strong significant consistencies between oncological board and ChatGPT-4 propositions for the following primary therapeutic options: surgery, palliative chemotherapy, and chemotherapy (induction) followed by chemoradiotherapy. There were no significant consistencies between ChatGPT-4 and the oncological board for alternative options (Table 2). Note that oncological board did not propose chemoradiotherapy or immunotherapy as primary therapeutic option, and surgery, or chemotherapy (induction) followed by chemoradiotherapy as alternative treatments in the present cohort. ChatGPT-4 never proposed immunotherapy as primary option, and surgery followed by postoperative radiotherapy as alternative option.

The AIPI score of ChatGPT-4 for the therapeutic management is available in Table 3. The ChaGPT-4 primary therapeutic management was considered as adequate (pertinent and optimal/suboptimal) regarding the oncological board decisions in 13 cases (65%; Table 3). Among the 7 inadequate management strategies, 4 (67%) concerned laryngeal malignancies. Precisely, practitioners reported the following inadequate therapeutic management of ChatGPT-4: proposition of chemoradiotherapy in a patient with many comorbidities contraindicating the chemotherapy ($N = 1$); proposition of total laryngectomy followed by radiotherapy in a patient with a history of primary laryngeal radiotherapy, while oncological board recommended partial laryngectomy ($N = 1$); proposition of postoperative radiotherapy in a patient with a history of primary radiotherapy, while oncological board indicated salvage surgery

Table 1 Patient features

N	G	Age	Cancer history	Additional examinations	Localization	Staging	Primary therapeutic options	Alternative options
1	F	72	–	MRI, PET–CT, Endoscopy, Biopsies	OSCC	cT4aN2bM0	RT	Surgery, flap, post-RT
2	M	45	Glottic cT1 (RT)	CT, PET–CT, Endoscopy, Biopsies	LSCC	cT3N0M0	CHEP	CTh, immunotherapy
3	M	75	Glottic cT1 (RT)	CT, PET–CT, Endoscopy, Biopsies	LSCC	cT3N0M0	Total laryngectomy	CTh, immunotherapy
4	M	55	Glottic cT1 (TLM)	CT, PET–CT	LSCC	cT1aN0M0	TLM or RT	–
5	M	56	–	CT, PET–CT, Endoscopy, Biopsies	LSCC	cT2N1M0	TORS partial laryngectomy and Neck dissection or RT	–
6	F	56	Medullar cancer (Surgery)	MRI, PET–CT, Calcitonin	Thyroid/neck	N1M0	TORS dissection or radioiodine	CTh
7	M	75	OSCC cT2N2 (RT)	MRI, PET–CT, Endoscopy, Biopsies	OSCC	cT4aN2M1	CTh	Immunotherapy
8	F	70	–	MRI, PET–CT, Endoscopy, Biopsies	OSCC	cT2N0M0	TORS and neck dissection, or RT	
9	M	70	Supraglottic cT3 (RT)	CT, PET–CT, Endoscopy, Biopsies	LSCC	cT2N0M0	Partial laryngectomy	CTh, immunotherapy
10	M	50	–	MRI, PET–CT, Endoscopy, FNAB	Unknown	cTxN1M0	TORS tonsillectomies, neck dissection or RT	CTh
11	F	49	–	MRI, PET–CT, Endoscopy, Biopsies, HPV	OSCC	cT2N1M0	TORS oropharyngectomy, neck dissection or RT	CTh
12	M	62	–	MRI, FNAB, HPV	Parotid	cT2N0M0	Parotidectomy and post-RT or RT	CRT
13	F	55	Parotid carcinoma (RT)	MRI, PET–CT, Chest CT, FNAB	Parotid	cT1N0M0	Parotidectomy, flap, reinnervation	CTh, immunotherapy
14	M	55	–	US, FNAB	Thyroid	cT1N0M0	Hemi-thyroidectomy	Iodine
15	M	55	–	MRI, FNAB	Sublingual	cT1N0M0	Surgery	RT
16	F	51	–	MRI, PET–CT, Endoscopy, Biopsies	Oral SCC	cT1N0M0	Partial glossectomy, sentinal node or RT	CTh
17	M	53	–	MRI, CT, PET–CT, Biopsies, HPV	Oral SCC	cT2N2aM0	Partial glossectomy, neck dissection or RT	CTh
18	F	45	–	MRI, CT, PET–CT, Biopsies, Hearing test	UCNT	cT4N2cM0	CTh (induction) and CRT	CRT
19	F	58	HSCC cT4N2 (CRT)	MRI, CT, PET–CT, Biopsies	HLSCC	pT2N0M0	Salvage pharyngolaryngectomy, neck dissection	CTh, immunotherapy
20	M	67	–	MRI, CT, PET–CT, Biopsies	Ethmoid	cT2N0M0	Surgery and post-RT	CRT

CHEP crico-hyodo-epiglotto-pexy; *CT* computed tomography; *CTh* chemotherapy; *FNAB* fine-needle aspiration biopsy; *H/L/OSCC* hypopharyngeal/laryngeal/oropharyngeal squamous cell carcinoma; *HPV* human papilloma virus; *MRI* magnetic resonance imaging; *PET–CT* positron emission tomography–computed tomodensitometry; *Post-RT* postoperative RT; *RT* radiotherapy; *TLM* transoral laryngeal microsurgery; *TORS* transoral robotic surgery; *UCNT* undifferentiated carcinoma nasopharyngeal type; *US* ultrasonography

($N = 1$); proposition of laryngeal radiotherapy (re-irradiation) in a patient with a history of non-response to primary laryngeal radiotherapy ($N = 1$); pre-operative radiotherapy in a patient with a cT2N2M0 oral SCC ($N = 1$); and radiotherapy in a patient with a failure of radiotherapy, while oncological board indicated salvage surgery ($N = 1$).

Stability of ChatGPT-4

The medical record findings were re-entered, and responses were regenerated at day 1 to analyze the stability of ChatGPT-4 over time. The consistency analysis between first and second ChatGPT-4 responses is available in Table 4.

Table 2 Consistency analyses for additional examinations and treatments

	Oncological			
	Board	ChatGPT-4	kappa	*p* value
Main additional examinations				
Neck MRI	14 (70)	18 (90)	0.167	NS
Neck CT	9 (45)	19 (95)	0.068	NS
PET–CT	16 (80)	14 (70)	0.583	0.004
Neck ultrasonography	1 (5)	3 (15)	0.459	0.015
Biopsy	12 (60)	14 (70)	0.565	0.010
Upper aerodigestive tract endoscopy	9 (45)	8 (40)	0.694	0.002
Specific biology/HPV detection (IHC)	4 (20)	7 (35)	0.634	0.001
Fine-needle aspiration biopsy	5 (25)	8 (40)	0.001	NS
Chest CT	1 (5)	5 (25)	0.091	NS
Primary treatments				
Surgery	15 (75)	15 (75)	0.467	0.037
Surgery and post-operative radiotherapy	2 (10)	4 (20)	0.154	NS
Radiotherapy	9 (45)	5 (25)	0.368	NS
Chemotherapy	1 (5)	2 (10)	0.643	0.002
Neoadjuvant chemotherapy and chemoradiotherapy	1 (5)	1 (5)	1.000	0.001
Targeted (radioiodine, tyrosine kinase blockers)	1 (5)	1 (5)	0.053	NS
Alternative treatments				
Radiotherapy	3 (15)	8 (40)	0.047	NS
Chemoradiotherapy	3 (15)	8 (40)	0.047	NS
Chemotherapy	11 (55)	10 (50)	0.100	NS
Immunotherapy	6 (30)	4 (20)	0.053	NS
Targeted (radioiodine, tyrosine kinase blockers)	1 (5)	10 (50)	0.100	NS

CT computed tomography; *IHC* immunohistological staining; *MRI* magnetic resonance imaging; *NS* non-significant

Most additional examinations and primary therapeutic options reported moderate-to-high consistency. ChatGPT-4 proposed 18 (90%) versus 19 (95%) MRI at first and second rounds, respectively, while neck CT was indicated in 19 (95%) and 10 (95%) cases at first and second rounds, respectively. PET–CT was indicated for 15 patients (75%) at the first round of responses and for 17 patients (85%) at the second round.

Discussion

The ongoing development of chatbots or software using artificial intelligence is changing our practice in medicine and surgery. To date, less than thirty studies were conducted in otolaryngology—head and neck surgery about the usefulness, accuracy, and performance of ChatGPT [3–6].

The primary findings of the present study supported that ChatGPT-4 commonly proposes a higher number of additional examinations for the oncological check-up compared to practitioners of the oncological board. In most cases (55%), the ChatGPT-4 propositions associated adequate and unnecessary examinations. This observation was similarly observed in recent studies, where authors reported that ChatGPT-4 proposed a list of potential additional examinations without selecting the most adequate for the clinical situation [1, 6, 11]. Radulesco et al. reported that ChatGPT-4 proposed a significant higher number of additional examinations than practitioners for establishing the diagnosis of nasal and ear disorders. As observed in our study, authors observed significant agreement between otolaryngologists and ChatGPT-4 for the indications of only some common examinations [11]. The findings of the present investigation and those of the literature support that ChatGPT-4 functions as an electronic encyclopedia proposing an exhaustive list of additional examinations without selecting the most adequate examinations for the cancer type or localization. The systematic indications to perform neck CT and MRI in upper aerodigestive tract malignancies, a fine-needle aspiration biopsy when patient reported neck nodes, or the association of chest X-ray and CT for the lung check-up are three examples supporting this impersonalized approach.

Interestingly, the ChatGPT-4 therapeutic options for patient cancer were judged as adequate in 65% of cases, which is a better accuracy rate than other studies conducted in otolaryngology practice [5, 6, 11]. In a recent clinical

Table 3 Artificial intelligence performance instrument findings

AIPI outcomes	*N* (%)
Consideration of medical and surgical histories	
Fully considered	14 (70)
Partly considered	3 (15)
Not considered	3 (15)
Consideration of patient symptoms	
Fully considered	19 (95)
Partly considered	1 (5)
Not considered	0 (0)
Consideration of clinical and fibroscopic examinations	
Fully considered	16 (80)
Partly considered	4 (20)
Not considered	0 (0)
Relevant additional examination	
Pertinent and necessary	5 (25)
Pertinent and not all necessary	11 (55)
Pertinent, necessary, and inadequate	4 (20)
Only inadequate examinations	0 (0)
Primary therapeutic options	
Pertinent and optimal	11 (55)
Pertinent but suboptimal	2 (10)
Association of pertinent/necessary and inadequate	6 (30)
No adequate strategy	1 (5)

The performance of ChatGPT-4 was evaluated by two board certified head and neck surgeons

AIPI artificial intelligence performance instrument; *N* number

Table 4 Stability of ChatGPT-4 propositions

	kappa	*p* value
Main additional examinations		
Neck MRI	0.643	0.002
Neck CT	0.053	NS
PET–CT	0.583	0.004
Neck ultrasonography	0.494	0.010
Biopsy	0.737	0.001
Upper aerodigestive tract endoscopy	0.138	NS
Specific biology/HPV detection (IHC)	0.468	0.035
Fine-needle aspiration biopsy	0.600	0.006
Chest CT	0.067	NS
Primary treatments		
Surgery	0.571	0.010
Surgery and post-operative radiotherapy	0.231	NS
Radiotherapy	0.059	NS
Chemotherapy	0.643	0.002
Neoadjuvant chemotherapy and chemoradiotherapy	1.000	0.001
Targeted (radioiodine, tyrosine kinase blockers)	0.459	0.015
Chemoradiotherapy	0.231	NS

CT computed tomography; *IHC* immunohistochemistry; *MRI* magnetic resonance imaging; *NS* non-significant; *PET–CT* positron emission tomography–computed tomography

study, our group showed that the treatments proposed by ChatGPT in otolaryngology were judged as pertinent in 22% of cases [5]. In this otolaryngology general practice study, the level of difficulty of clinical cases was not predictive for the ChatGPT performance [5], while in the present study, ChatGPT-4 presented some difficulties to propose adequate therapeutic options for complicated laryngeal cancer cases, especially when patients had history of laryngeal radiotherapy. The findings of both studies suggested that the performance of ChatGPT remains unpredictable and may widely vary from one clinical case type to another.

The accuracy of the ChatGPT-4 response may, moreover, vary from one response to another. Indeed, our analysis reported moderate consistency between re-generated responses in the indications of PET–CT, neck ultrasonography, or HPV detection, while there was no significant consistency for the endoscopy, or Chest CT indications throughout regenerated response process. Similarly, the therapeutic options may vary from one round of responses to another. Perlis also investigated the stability of ChatGPT-4 throughout regenerated answers in the management of depression in psychiatry [12]. This author reported some inconsistencies between re-generated answers. Precisely, he observed that ChatGPT-4 did not consider the history of patient, and, for example, recommended selective serotonin reuptake inhibitors after a trial failure based on selective serotonin reuptake inhibitors [12].

The lack of accuracy and stability of ChatGPT-4 limits the spread of such artificial intelligence-powered language model in clinical practice according to the risk of providing inadequate clinical information. Large language models are non-deterministic and, as demonstrated in our study and others [11], their outputs may vary with each run that is curtailed by fine-tuning-specific hyperparameters [13]. The ChatGPT-4 advanced adjustments of hyperparameters are currently not fully available, which may limit the understanding of the system's responses and propositions. This point and the low number of clinical cases are the primary limitation of this preliminary study. The main strengths of the present study are the originality and consideration of real oncological cases. To the best of our knowledge, this study is the first investigation of the accuracy of ChatGPT-4 in head and neck oncology and surgery. Future large-database studies are needed to explore the accuracy of ChatGPT-4 and other artificial intelligence-powered language models, such as Llama 2.0, and to determine their respective performance in the management of laryngeal, hypopharyngeal, oropharyngeal, oral, thyroid, sinus and salivary gland malignancies.

Conclusion

ChatGPT-4 may become a promising adjunctive tool in head and neck oncology. To date, ChatGPT-4 appears to be more efficient for theoretical information, including the cTNM staging, the list of potential useful additional examinations, or therapeutic options, than for providing a personalized therapeutic management considering the patient history and past treatments. Future clinical studies are needed to assess the performance of ChatGPT-4 and future updated models in large database of real head and neck oncological cases.

Appendix 1: Cases

N	G	Age	Symptoms	History/medication	Clinical examination	Oncological board-practitioners		
						Additional examinations	Diagnosis	Oncological board treatments
1	F	72	Dysphagia, right otalgia, weight loss (>6kg in 3 months)	HT, DB2, CD, RTU Current ATC	Trismus, normal tongue mobility, Exophytic lesion of the lateral oropharyngeal wall (right) Right neck node	MRI, PET–CT, endoscopy Biopsies: SCC	cT4aN2bM0 OSCC	Primary: radiotherapy Alternative: surgery, free flap, postoperative RT
2	M	45	Dysphonia (2 months)	Radiotherapy for cT1N0 of the glottic region (4y), Past TC	Anterior commissure lesion, decreased movement of vocal folds	Neck CT, PET–CT Endoscopy, biopsy: LSCC	cT3N0M0 LSCC Thyroid cartilage invasion	Primary: CHEP only Alternative: CTh/immunotherapy
3	M	75	Dysphonia, neck pain, dysphagia, weight loss (> 5kg in 3 months)	CP, HT, DB2, current ATC, RT for cT1N0 glottic LSCC (4 years)	Fixed right hemilarynx No exophytic lesion	Neck CT, PET–CT, endoscopy Biopsies: SCC	cT3N0M0 LSCC	Primary: total laryngectomy Alternative: CTh/immunotherapy.
4	M	55	Dysphonia (2 months)	HT, current TC TLM for cT1a glottis LSCC (6 months)	Exophytic lesion of the right vocal cord Normal laryngeal mobility	Neck CT and PET–CT No biopsy regarding morphological lesion	cT1aN0M0 LSCC (recurrence)	Primary: TLM or RT Alternative: –
5	M	56	Dysphagia (6 months)	Current TC	Exophytic lesion of right ary—epiglottic fold and epiglottis Normal vocal cord exam	Neck CT, PET–CT Endoscopy, biopsy: LSCC	cT2N1M0 Supraglottic LSCC	Primary: TORS supraglottic laryngectomy and neck dissections or RT Alternative: –
6	F	56	Throat pain, globus (6 months)	Thyroidectomy (1 year) for medullar cancer, bilateral neck dissection	Right oropharyngeal wall mass	Neck MRI, PET–CT Calcitonin biology	Neck recurrence of medullar thyroid cancer (parapharyngeal space)	Primary: TORS node surgery, or targeted therapy Alternative: CTh
7	M	75	Throat pain, dysphagia, weight loss (>7kg–3 months)	HT, current ATC RT for cT2N2 OSCC (5 years)	Left oropharyngeal wall exophytic lesion, ipsilateral neck nodes	Neck MRI, PET–CT, endoscopy Biopsies: SCC	cT4aN2M1 OSCC spinal bone metastases	Primary: chemotherapy Alternative: immunotherapy bone radiation
8	F	70	Dysphagia, throat pain (3 months)	HT, no ATC	Right base of tongue ulcerative lesion	Neck MRI, PET–CT, endoscopy Biopsies: SCC	cT2N0M0 OSCC	Primary: TORS and ipsi-lateral neck dissection or RT Alternative: CTh

N	G	Age	Symptoms	History/medication	Clinical examination	Oncological board-practitioners		
						Additional examinations	Diagnosis	Oncological board treatments
9	M	70	Dysphonia (4 months)	Current TC Radiotherapy for cT3 Supraglottic cancer	Exophytic lesion of the left vocal cord and anterior commissure of the larynx	Neck CT, PET–CT, endoscopy	cT2N0M0 Supraglottic LSCC (recurrence)	Primary: partial laryngectomy Alternative: CTh/immunotherapy.
10	M	50	Neck mass (6 months)	None	Right neck node 2cm Endoscopy: normal	Neck MRI, PET–CT, endoscopy Fine-needle aspiration biopsy	cTxN1M0 SCC	Primary: TORS tonsillectomy Neck dissection, no post-RT or RT Alternative: CTh
11	F	49	Throat pain (4 months)	None	Ulceration lesion in left tonsil and left neck adenopathy	Neck MRI, PET–CT Biopsies: SCC HPV detection	cT2N1M0 OSCC cT2N0M0	Primary: TORS tonsillectomy neck dissection, no post-RT or RT Alternative: CTh
12	M	62	Left parotid gland nodules	HT, left superficial parotidectomy for pleomorphic adenoma (10 years)	Left parotid node, no facial palsy, no adenopathy	Neck MRI Fine-needle aspiration biopsy HPV detection	Parotid carcinoma	Primary: parotidectomy neck dissection, postoperative RT or RT Alternative: CRT
13	F	55	Right facial nerve paralysis, nodules (3 weeks)	Parotidectomy and RT for a right adeno-carcinoma (4 years)	Right facial nerve paralysis Parotid region nodules No lymph adenopathy	Neck MRI, PET–CT Chest CT Fine-needle aspiration biopsy	cT1N0M0 Recurrence of Adenocarcinoma	Primary: parotidectomy free flap, reinnervation Alternative: CTh/immunotherapy.
14	M	55	Right EU-Tirads 5 thyroid nodule (8 months)	None	Normal, endoscopy normal No vocal fold paralysis	Fine-needle aspiration biopsy Ultrasonography	cT1N0M0 papillary carcinoma	Primary: lobectomy Alternative: iodine
15	M	55	Right sublingual gland nodule (6 months)	HT, current TC	Right sublingual nodule, Examination: normal	Neck MRI Fine-needle aspiration	cT1N0M0 muco-epidermoid carcinoma	Primary: sublingual surgery, low grade cancer, no postoperative RT Alternative: RT
16	F	51	Oral cavity pain, and tongue ulceration (6 months)	Current TC	Right tongue ulceration of 1cm (latero-posterior)	Neck/oral MRI, PET–CT Endoscopy and biopsy	cT1N0M0 Oral SCC	Primary: partial glossectomy Sentinel node dissection or RT Alternative: CTh
17	M	53	Oral cavity pain, ulceration of inferior tongue part and oral cavity floor (7 months)	HT, current ATC	Left tongue and oral cavity floor lesion, left neck nodes	Neck MRI, CT, PET–CT Biopsies: SCC HPV detection	cT2N2aM0 Oral SCC	Primary: partial glossectomy, neck dissection, FAMM, or RT Alternative: CTh
18	F	45	Epistaxis, diplopia, Right deafness, neck nodes (5 months)	None	Right exophytic nasopharyngeal lesion and right chronic otitis Multiple cervical nodes	MRI, Neck CT, PET–CT Biopsy Audiometry, tympanometry	cT4N2cM0 UCNT	Primary: induction CTh and CRT Alternative: CRT

N	G	Age	Symptoms	History/medication	Clinical examination	Oncological board-practitioners		
						Additional examinations	Diagnosis	Oncological board treatments
19	F	58	Aphagia, weight loss throat pain (6 months)	Past TC, CRT for a cT4N2M0 HSCC (2 years)	Laryngopharyngeal edema and saliva, no neck node	Neck CT, MRI, PET–CT Biopsies: SCC	pT2N0M0 HLSCC (recurrence)	Primary: salvage pharyngo-laryngectomy and neck dissections Alternative: CTh/immunotherapy.
20	M	67	Unilateral epistaxis and obstruction, diplopia	HT	Exophytic lesion in right nasal cavity, no neck node	Nasal CT, MRI, PET–CT Biopsy	cT2N0M0 Ethmoid Intestinal-type Adenocarcinoma	Primary: endoscopic nasal surgery, postoperative RT Alternative: CRT

A/TC alcohol/tobacco consumption; *CD* coronary disease; *CHEP* crico-hyodo-epiglotto-pexy; *CP* chronic pancreatitis; *CT* computed tomography; *CTh* chemotherapy; *DB2* diabetes type 2; *FAMM* facial artery musculomucosal; *FNAB* fine-needle aspiration biopsy; *H/L/OSCC* hypopharyngeal/laryngeal/oropharyngeal squamous cell carcinoma; *HT* hypertension; *MRI* magnetic resonance imaging; *PET–CT* positron emission tomography–computed tomodensitometry; *RT* radiotherapy; *RTU* respiratory tuberculosis; *TLM* transoral laryngeal microsurgery; *TORS* transoral robotic surgery; *UCNT* undifferentiated carcinoma nasopharyngeal type; *US* ultrasonography

Appendix 2: Artificial intelligence performance instrument scores of ChatGPT-4

AIPI outcomes	Mean (SD)
1. Medical and Surgical History (/2)	1.55 ± 0.76
2. Symptoms (/2)	1.95 ± 0.22
3. Physical examinations (/2)	1.80 ± 0.41
Patient feature score (/6)	5.30 ± 1.08
4. Differential diagnoses (/3)	3.00 ± 0.01
5. Primary diagnosis (/3)	3.00 ± 0.01
6. Management plan (/1)	0.80 ± 0.41
Diagnostic score (/7)	6.80 ± 0.41
7. Additional examinations (/3)	2.05 ± 0.69
8. Most relevant additional examination (/1)	0.90 ± 0.31
Additional examination score (/4)	2.95 ± 0.83
9. Treatment (/3)	2.30 ± 0.87
10. AIPI total score (/20)	17.35 ± 2.32

The performance of ChatGPT-4 was evaluated by two board certified head and neck surgeons. Note that for the primary and differential diagnosis outcomes, the score of ChatGPT-4 was considered as maximum, because no need to perform a differential diagnosis

AIPI artificial intelligence performance instrument; *SD* standard deviation

Author contributions Study concept and design: Lechien, Hans, Chiesa-Estomba. Acquisition, analysis, or interpretation of data: Lechien, Hans. Drafting of the manuscript: Lechien. Critical revision of the manuscript for important intellectual content: Hans, Chiesa-Estomba, Baudouin.

Funding None.

Data availability Data are available on request.

Declarations

Conflict of interest The authors have no conflict of interest.

Ethical declarations The author Jerome R. Lechien is also guest editor of the special issue on ‘ChatGPT and Artificial Intelligence in Otolaryngology—Head and Neck Surgery’. He was not involved with the peer review process of this article.

Informed consent Patients consented to the study.

References

1. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino A, Beltramini GA et al (2023) Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery : a multicenter collaborative analysis. Otolaryngol head neck surg. https://doi.org/10.1002/ohn.489
2. 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
3. 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
4. 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

5. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA (2023) Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the artificial intelligence performance instrument (AIPI). Eur Arch Otorhinolaryngol. https://doi.org/10.1007/s00405-023-08219-y
6. 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. https://doi.org/10.1007/s00405-023-08282-5
7. Cuny F, Babin E, Lacau-Saint-Guily J, Baujat B, Bensadoun R, Bozec A, Chevalier D, Choussy O, Deneuve S, Fakhry N, Guigay J, Makeieff M, Merol JC, Mouawad F, Pavillet J, Rebiere C, Righini C, Sostras MC, Tournaille M, Vergez S, SFORL work group (2015) French Society of ENT (SFORL) guidelines for care pathway organization in head and neck oncology (short version). Early management of head and neck cancer. Eur Ann Otorhinolaryngol Head Neck Dis 132(4):205–208. https://doi.org/10.1016/j.anorl.2015.06.007
8. Verdonck-de Leeuw I, Dawson C, Licitra L, Eriksen JG, Hosal S, Singer S, Laverty DP, Golusinski W, Machczynski P, Varges Gomes A, Girvalaki C, Simon C, Leemans CR (2022) European Head and Neck Society recommendations for head and neckcancer survivorship care. Oral Oncol 133:106047. https://doi.org/10.1016/j.oraloncology.2022.106047
9. Lydiatt WM et al (2017) Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. CA Cancer J Clin 67(2):122–137
10. Piazza C, Paderno A, Sjogren EV, Bradley PJ, Eckel HE, Mäkitie A, Matar N, Paleri V, Peretti G, Puxeddu R, Quer M, Remacle M, Vander Poorten V, Vilaseca I, Simo R (2021) Salvage carbon dioxide transoral laser microsurgery for laryngeal cancer after (chemo)radiotherapy: a European Laryngological Society consensusstatement. Eur Arch Otorhinolaryngol 278(11):4373–4381. https://doi.org/10.1007/s00405-021-06957-5
11. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR (2023) ChatGPT-4 Performance in Rhinology: A Clinical Case-Series. Revision. Int For Allerg Rhinol
12. Perlis RH (2023). MedRxiv. https://doi.org/10.1101/2023.04.14.23288595
13. Lechien JR, Briganti G, Vaira LA (2023) Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. Eur Arch Otorhinolaryngol

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.