European Archives of Oto-Rhino-Laryngology (2024) 281:2123–2136
https://doi.org/10.1007/s00405-024-08525-z

HEAD AND NECK

Exploring the landscape of AI-assisted decision-making in head and neck cancer treatment: a comparative analysis of NCCN guidelines and ChatGPT responses

Filippo Marchi[1,2] · Elisa Bellini[1,2] · Andrea Iandelli[1] · Claudio Sampieri[3,4,5] · Giorgio Peretti[1,2]

Received: 14 December 2023 / Accepted: 2 February 2024 / Published online: 29 February 2024

Abstract

Purpose Recent breakthroughs in natural language processing and machine learning, exemplified by ChatGPT, have spurred a paradigm shift in healthcare. Released by OpenAI in November 2022, ChatGPT rapidly gained global attention. Trained on massive text datasets, this large language model holds immense potential to revolutionize healthcare. However, existing literature often overlooks the need for rigorous validation and real-world applicability.

Methods This head-to-head comparative study assesses ChatGPT's capabilities in providing therapeutic recommendations for head and neck cancers. Simulating every NCCN Guidelines scenarios. ChatGPT is queried on primary treatments, adjuvant treatment, and follow-up, with responses compared to the NCCN Guidelines. Performance metrics, including sensitivity, specificity, and F1 score, are employed for assessment.

Results The study includes 68 hypothetical cases and 204 clinical scenarios. ChatGPT exhibits promising capabilities in addressing NCCN-related queries, achieving high sensitivity and overall accuracy across primary treatment, adjuvant treatment, and follow-up. The study's metrics showcase robustness in providing relevant suggestions. However, a few inaccuracies are noted, especially in primary treatment scenarios.

Conclusion Our study highlights the proficiency of ChatGPT in providing treatment suggestions. The model's alignment with the NCCN Guidelines sets the stage for a nuanced exploration of AI's evolving role in oncological decision support. However, challenges related to the interpretability of AI in clinical decision-making and the importance of clinicians understanding the underlying principles of AI models remain unexplored. As AI continues to advance, collaborative efforts between models and medical experts are deemed essential for unlocking new frontiers in personalized cancer care.

Keywords Machine learning · Artificial intelligence (AI) models · ChatGPT · Cancer care · National Comprehensive Cancer Network (NCCN) Guidelines · Head and neck cancers

Introduction

Recent advancements in natural language processing (NLP) and machine learning (ML) have opened new possibilities for leveraging artificial intelligence (AI) models. These models, such as ChatGPT (Generative Pretrained Transformer), a powerful large language model (LLM) developed by OpenAI, have rapidly gained global attention since their public release in November 2022, attracting an unprecedented 100 million users within just 2 months.

The integration of AI in healthcare has witnessed a progressive evolution, with ML and NLP playing pivotal roles. ML algorithms, initially applied for tasks such as image recognition and diagnostics [1] have matured into sophisticated models capable of processing vast amounts

Filippo Marchi and Elisa Bellini contributed equally to this manuscript and share first authorship.

✉ Elisa Bellini
e.e.elisabellini@gmail.com

1 Unit of Otorhinolaryngology-Head and Neck Surgery, IRCCS Ospedale Policlinico San Martino, Largo Rosanna Benzi, 10, 16132 Genoa, Italy

2 Department of Surgical Sciences and Integrated Diagnostics (DISC), University of Genoa, 16132 Genoa, Italy

3 Department of Experimental Medicine (DIMES), University of Genoa, Genoa, Italy

4 Department of Otolaryngology-Hospital Cliníc, Barcelona, Spain

5 Functional Unit of Head and Neck Tumors-Hospital Cliníc, Barcelona, Spain

of medical data [2–4]. NLP, on the other hand, has enabled AI systems to understand and generate human-like text, facilitating communication between machines and healthcare professionals.

The development of large LLMs, exemplified by GPT-3 and its successor ChatGPT, represents a noteworthy milestone. These models, built on architectures leveraging billions of parameters and trained on massive text datasets, hold substantial potential to revolutionize diverse fields, including healthcare. ChatGPT has demonstrated its prowess in healthcare by passing medical exams, elucidating treatment risks and benefits, educating patients on obstructive sleep apnea, and generating automated hospital discharge summaries [5–9]. Furthermore, it can simplify complex medical terminology, enhancing patients' understanding of their options. It is crucial to emphasize, however, that while ChatGPT can be a valuable resource, it does not substitute for professional medical advice, and decisions should always be discussed with healthcare professionals [10].

Understanding the evolution of AI, from foundational machine learning to the development of advanced language models, is essential in appreciating the capabilities and drawbacks, and thus, the potential impact of ChatGPT in healthcare.

Clinical guidelines serve as the gold standard in medical decision-making, developed through meticulous analysis of high-level evidence, expert consensus, and a thorough review of the current medical literature.

Established in 1995, the National Comprehensive Cancer Network (NCCN) has become a cornerstone in oncology, committing to enhancing the quality, effectiveness, and efficiency of cancer care. The NCCN guidelines are a result of a comprehensive program focused on developing clinical practice guidelines for the management of several tumors. Rooted in consensus-building, these guidelines undergo continuous evaluation of evidence and structured feedback, reflecting the dynamic nature of cancer care to incorporate the latest high-level of evidence-base and consensus of a multidisciplinary panels of experts in cancer care [10, 11].

In the ever-evolving landscape of oncological decision-making, the integration of artificial intelligence, exemplified by tools like ChatGPT, necessitates rigorous performance benchmarking against the established NCCN guidelines [11].

In this context, our study aims to explore the landscape of AI-assisted decision-making in head and neck oncology, specifically focusing on the concordance between ChatGPT responses on treatment suggestions, adjuvant treatment indications, and follow-up recommendations and the established guidelines. By comparing ChatGPT's responses with the NCCN guidelines [11], we seek to evaluate the model's performance and highlight its potential impact in the context of oncological decision-making.

Materials and methods

In this head-to-head comparative study, we consulted ChatGPT version 3.5 to obtain the most accurate therapeutic recommendations for each combination of cancer site, stage and nodal involvement (according to the latest edition of the AJCC Staging Manual [12]), covering the landscape of head and neck cancer scenarios outlined by the NCCN guidelines [11] (i.e., from the NCCN Guidelines Version 2.2024 Table OR 1- Cancer of the Oral Cavity- Including Mucosal Lip-: T1-2,N0; T3N0; T1-3, N1-3; T4a,N0-3; T4b, N0-3). All the categories explored are listed in Table 1. These categories served as the basis for our comparative analysis. Then we compared the answer provided by the Chatbot with the Guidelines suggestions. The chatbot's answer was considered correct if responses were deemed fully aligned when they unequivocally matched the treatment suggestions, adjuvant treatment indications, and follow-up recommendations outlined in the NCCN Guidelines. Chatbot's responses were categorized as incorrect when they deviated from the NCCN recommendations one or more suggestions, either providing inaccurate information or suggestions not supported by the guidelines.

In Table 2 we reported, as examples, few questions we asked.

To address unintentional bias, a junior author (E.B.) consulted the chatbot, while two senior authors (F.M., A.I.) independently assessed responses. The process was repeated twice, with disagreements resolved by a senior author (G.P.). Patients with synchronous head and neck cancer, previous treatment(s) for head and neck cancer, and metastatic patients were excluded to focus specifically on treatment recommendations in scenarios without these complicating factors.

We utilized a range of performance metrics to comprehensively evaluate the concordance between ChatGPT responses and the NCCN Guidelines. Sensitivity emerged as a critical metric, given its role in identifying true positives and ensuring reliable recognition of appropriate treatments, a crucial aspect in the clinical context of accurate treatment suggestions. Sensitivity was calculated by considering true positives (TP) and false negatives (FN) (where ChatGPT failed to identify correct responses). Specificity was deemed inapplicable since, assuming NCCN Guidelines are the "gold standard, and thus always correct, true negatives (TN) were absent. Accuracy served as an overall measure, considering both TP and TN, providing an overview of the model's performance across diverse scenarios. Precision gauged the proportion of correctly identified answers (TP) among all cases predicted as correct by ChatGPT, offering insights

Table 1 Raw data of ChatGPT-generated suggestions and NCCN guidelines for head and neck cancer, including primary treatment, adjuvant treatment, and follow-up

Tumor site and stage		ChatGPT primary treatment	Concordance with NCCN Guidelines	ChatGPT adjuvant treatment	Concordance with NCCN Guidelines	ChatGPT follow-up	Concordance with NCCN Guidelines
Oral cavity							
	T1-2,N0	Correct	+	Correct	+	Correct	+
	T1-2,N0	Correct	+	Correct	+	Correct	+
	T3,N0	Correct	+	Correct	+	Correct	+
	T1-3,N1–3	Correct	+	Correct	+	Correct	+
	T4, N0-3	Correct	+	Correct	+	Correct	+
Oroph p16-							
	T1-2, N0	Correct	+	Not correct	−	Correct	+
	T1-2,N1	Correct	+	Correct	+	Correct	+
	T3-4a, N0-1	Correct	+	Correct	+	Correct	+
	T3-4a, N2-3	Correct	+	Correct	+	Correct	+
Oroph p16 +							
	T1–2, N0	Not correct	-	Correct	+	Correct	+
	T0–2,N1 (single node ≤ 3 cm)	Correct	+	Correct	+	Correct	+
	T0–2,N1 (single node > 3 cm, or 2 or more ipsilateral nodes ≤ 6 cm	Correct	+	Correct	+	Correct	+
	T0-2,N2	Correct	+	Correct	+	Correct	+
	T3, N0	Correct	+	Correct	+	Correct	+
	T0-3, N3	Correct	+	Correct	+	Correct	+
	T4, N0-3	Correct	+	Correct	+	Correct	+
Hypopharynx							
	Most T1,N0, selected T2,N0 (amenable to larynx- preserving [conservation] surgery)	Not correct	−	Correct	+	Correct	+
	T2–3,N0-3 (if requiring [amenable to] pharyngectomy with partial or total laryngectomy)	Correct	+	Correct	+	Correct	+
	Response after induction chemotherapy for T2-3, N0-3 or T1,N +	Correct	+	Correct	+	Correct	+
	T4a, N0-3	Correct	+	Correct	+	Correct	+

Table 1 (continued)

Tumor site and stage		ChatGPT primary treatment	Concordance with NCCN Guidelines	ChatGPT adjuvant treatment	Concordance with NCCN Guidelines	ChatGPT follow-up	Concordance with NCCN Guidelines
Nasopharynx							
	T1,N0,M0	Correct	+	Correct	+	Correct	+
	T2,N0,M0	Correct	+	Correct	+	Correct	+
	T0 (EBV +)-2,N1,M0	Correct	+	Correct	+	Correct	+
	T3,N0,M0	Correct	+	Correct	+	Correct	+
	T3-4,N1 3,M0	Not correct	−	Not correct	+	Correct	+
	Any T,N2 3,M0	Not correct	−	Correct	+	Correct	+
	M1 Oligometastatic disease	Correct	+	Correct	+	Correct	+
	M1 Widely metastatic and Good PS (0–2)	Correct	+	Correct	+	Correct	+
	M1 Widely metastatic and Poor PS (3–4)	Correct	+	Correct	+	Correct	+
Glottic Larynx							
	Carcinoma in situ	Correct	+	Correct	+	Correct	+
	Amenable to larynx-preserving (conservation) surgery (T1-T2,N0 or select T3,N0)	Correct	+	Not correct	−	Correct	−
	T3 requiring (amenable to) total laryngectomy (N0-1)	Not correct	−	Correct	+	Not correct	−
	T3 requiring (amenable to) total laryngectomy (N2-3)	Not correct	−	Correct	+	Not correct	−
	Response after induction chemotherapy	Correct	+	Correct	+	Not correct	−
	T4a,N0-3	Correct	+	Correct	−	Correct	+
	Selected T4a patients who decline surgery	Correct	+	Correct	+	Correct	+

Table 1 (continued)

Tumor site and stage		ChatGPT primary treatment	Concordance with NCCN Guidelines	ChatGPT adjuvant treatment	Concordance with NCCN Guidelines	ChatGPT follow-up	Concordance with NCCN Guidelines
Supraglottic Larynx							
	Amenable to larynx-preserving (conservation) surgery (T1-T2,N0)	Correct	+	Correct	+	Correct	+
	Amenable to larynx-preserving (conservation) surgery (select T3,N0)	Not correct	–	Correct	+	Correct	+
	T3 requiring (amenable to) total laryngectomy (N0-1)	Correct	+	Correct	+	Correct	+
	Amenable to larynx-preserving (conservation) surgery (T1–2,N + and selected T3,N1)	Not correct	-	Correct	+	Correct	+
	T3 requiring (amenable to) total laryngectomy (N2-3)	Correct	+	Correct	+	Correct	+
	Response after induction chemotherapy	Correct	+	Correct	+	Correct	+
	T4a,N0-3	Correct	+	Correct	+	Correct	+
	Selected T4a patients who decline surgery	Correct	+	Correct	+	Correct	+
Maxillary Sinus Tumors							
	T1–2,N0 (All histologies except adenoid cystic)	Correct	+	Correct	+	Correct	+
	T1–2,N0 Adenoid cystic	Correct	+	Correct	+	Correct	+
	T3–T4a,N0	Correct	+	Correct	+	Correct	+
	T1–T4a,N +	Correct	+	Correct	+	Correct	+
	T4b,N0–3	Correct	+	Correct	+	Correct	+

Table 1 (continued)

Tumor site and stage		ChatGPT primary treatment	Concordance with NCCN Guidelines	ChatGPT adjuvant treatment	Concordance with NCCN Guidelines	ChatGPT follow-up	Concordance with NCCN Guidelines
Unknown Primary							
	Primary Found	Correct	+	Correct	+	Correct	+
	Primary Not Found, AdenoK (Level I-III)	Correct	+	Correct	+	Correct	+
	Primary Not Found, AdenoK (Level IV-V)	Not correct	+	Correct	+	Correct	+
	Primary not found, Poorly differentiated or nonkeratinizing SCC or NOS or anaplastic (not thyroid) or SCC, N0	Correct	+	Correct	+	Correct	+
	Primary not found, Poorly differentiated or nonkeratinizing SCC or NOS or anaplastic (not thyroid) or SCC, N1	Correct	+	Correct	+	Correct	+
	Primary not found, Poorly differentiated or nonkeratinizing SCC or NOS or anaplastic (not thyroid) or SCC, N2-N3	Correct	+	Correct	+	Correct	+
	Post Neck Dissection, N1 ENE neg	Correct	+	Correct	+	Correct	+
	N2,N3 ENE neg	Correct	+	Correct	+	Correct	+
	ENE pos	Correct	+	Correct	+	Correct	+
Very Advanced (T4b SCC)							
	Newly diagnosed (M0) T4b,N0–3 or unresectable nodal disease or unfit for surgery	Correct	+	Correct	+	Correct	+
	Newly diagnosed M1 disease at initial presentation	Correct	+	Correct	+	Correct	+
	Metastatic (M1) disease at initial presentation	Correct	+	Correct	+	Correct	+
	Recurrent or persistent disease, locoregional recurrence, second primary, or persistent disease with prior RT	Correct	+	Correct	+	Correct	+
	Recurrent or persistent disease with distant metastases	Correct	+	Correct	+	Correct	+

Table 2 Examples of questions asked to the chatbot and the responses obtained for different scenarios along with their respective evaluations

Subiste	Question	ChatGPT	NCCN	Correct/Incorrect
Oral Cavity	Which is the treatment of choice for T1-T2, N0 oral cavity cancer?	Surgery as the primary modality involves removing the tumor along with a margin of healthy tissue to ensure complete removal, with or without selective neck dissection. Another option is exclusive radiotherapy	Surgery (preferred) or radiotherapy	Correct
	When is an adjuvant radiotherapy required?	If the surgical resection of the tumor reveals positive margins, pathological features of the tumor (such as perineural invasion or lymphovascular invasion), extracapsular extension, in case of involvement of multiple lymph nodes	Adverse pathologic features: extranodal extension, positive margins, close margins, pT3 or pT4 primary, pN2 or pN3 nodal disease, nodal disease in levels IV or V, perineural invasion, vascular invasion, and lymphatic invasion	Correct
Larynx	Which is the treatment of choice of T3 (N2-3) glottic cancer requiring and that are amenable to total laryngectomy?	The treatment of choice often involves a combination of surgery and adjuvant therapy	Concurrent systemic therapy or surgery or induction chemotherapy	Incorrect
	In T3 (N2-3) glottic cancer requiring and that are amenable to total laryngectomy when do you suggest adjuvant therpay?	Adjuvant radiotherapy is requierd in case of pathological features such as close or positive margins, perineural invasion, or lymphovascular invasion. In some cases, especially when there are higher-risk features, concurrent chemoradiation may be considered	Radiotherapy or consider systemic therapy/RT. Adjuvant RT is considered in case of extranodal extension, positive margins, close margins, pT4 primary, pN2 or pN3 nodal disease, perineural invasion, vascular invasion, lymphatic invasion, and subglottic extension	Incorrect
Oropharynx	Which is the treatment of choice in T0-3, N3 oropharynx cancer HPV related?	Definitive concurrent chemoradiation, surgery associated with selective or modified radical neck dissection, induction chemotherapy followed by systemic therapy/RT	Concurrent systemic therapy/RT, Resection of the primary and ipsilateral or bilateral neck dissection, induction chemotherapy followed by RT or systemic therapy/RT	Correct
	In which cases of T0-3, N3 oropharyngeal cancer HPV related do you suggest adjuvant treatment?	If the surgical resection of the tumor reveals positive margins, pathological features of the tumor (such as perineural invasion or lymphovascular invasion), extracapsular extension, in case of involvement of multiple lymph nodes	Adverse pathologic features: extranodal extension, positive margins, close margins, pT3 or pT4 primary, pN2 or pN3 nodal disease, nodal disease in levels IV or V, perineural invasion, vascular invasion, and lymphatic invasion	Correct
Nasopharynx	Which is the treatment of choice of T1, N0 nasopharynx cancer	The treatment of choice often involves definitive radiation therapy	Definitive RT to nasopharynx and elective RT to neck	Correct
	Which is the treatment of choice of T3-4, N1-3 nasopharynx cancer	Typically involves a combination of radiation therapy and chemotherapy, known as concurrent chemoradiation (CRT) or clinical trials	Clinical trials (preferred) or induction chemotherapy followed by systemic therapy/RT (preferred) or concurrent systemic therapy/RT followed by adjuvant chemotherapy (category 2B) or concurrent systemic therapy/R	Correct

Table 2 (continued)

Subsite	Question	ChatGPT	NCCN	Correct/Incorrect
Hypopharynx	What type of treatment do you suggest in T4a, N0–3 hypopharynx cancer?	Total laryngopharyngectomy with selective neck dissection, induction chemotherapy, systemic therapy, targeted therapies	Total laryngopharyngectomy + ipsilateral or bilateral neck dissection ± hemi- or total thyroidectomy, after ipsilateral or bilateral paratracheal lymph node dissection, induction chemotherapy (category 3), concurrent systemic therapy/RT, clinical trial	Correct
	When do you suggest follow up?	If the surgical resection of the tumor reveals positive margins, pathological features of the tumor (such as perineural invasion or lymphovascular invasion), extracapsular extension, in case of involvement of multiple lymph nodes	Adverse pathologic features: extranodal extension, positive margins, close margins, pT3 or pT4 primary, pN2 or pN3 nodal disease, nodal disease in levels IV or V, perineural invasion, vascular invasion, and lymphatic invasion	Correct

$$\textbf{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

$$\textbf{Accuracy} = \frac{\text{True Positives+True Negatives}}{\text{Total Cases}}$$

$$\textbf{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

$$\textbf{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision+Sensitivity}}$$

Fig. 1 Formulas to calculate perfomance metrics

into the model's ability to avoid false positives (FP). The F1 score, balancing precision and recall, provided a comprehensive measure of the model's performance, crucial in situations where FP and FN carry different weights. All the formulae are reported in Fig. 1. The confidence interval (95% CI) has been calculated for each metric across the three scenarios. These selected metrics collectively offer a nuanced evaluation of ChatGPT's performance in providing recommendations. The diagnostic performance of ChatGPT was assessed across three distinct sets of questions (Primary treatment, Adjuvant treatment, and Follow-up) using Receiver Operating Characteristic (ROC) curves and Area under the ROC Curve (AUC) calculations. The AUC values, along with their 95% confidence intervals, were computed to quantify the discriminatory ability of the test in each population. To statistically compare the diagnostic performance among populations, a bootstrap analysis was employed, generating 1000 resampled datasets for each population. Pairwise comparisons of AUCs were conducted using bootstrap-derived confidence intervals, with significance determined if the interval did not include zero. Statistical analyses were performed using the R software for statistical computing (R version 4.3.2).

Table 3 Performance metrics for ChatGPT in comparison to NCCN guidelines, including sensitivity, specificity (not applicable), overall accuracy, and F1 score for primary treatment (Treat) and the evaluation of Confidence Interval (CI)

	NCCN Treat	95% CI
ChatGPT sensitivity	100	0.97–1.00
ChatGPT specificity	Not applicable	Not applicable
Chat GPT accuracy	85.3%	0.78–0.92
F1 score	92.1%	Not applicable

Table 4 Performance metrics for ChatGPT in comparison to NCCN guidelines, including sensitivity, specificity (not applicable), overall accuracy, and F1 score for adjuvant treatment (Adj) categories and the evaluation of Confidence Interval (CI)

	NCCN Adj	95% CI
ChatGPT sensitivity	95.59%	0.98–1.00
ChatGPT specificity	Not applicable	Not applicable
Chat GPT accuracy	95.59%	0.76–0.95
F1 score	95.59%	Not applicable

Results

Performance evaluation of Chat GPT in NCCN-related scenarios

In our study, ChatGPT exhibited promising capabilities in addressing NCCN-related queries across diverse scenarios. Our investigation involved 68 hypothetical clinical cases for primary treatment, adjuvant treatment, and follow-up, covering all stages and tumor sites outlined in the NCCN Guidelines. A summary of the raw data collected is presented in Table 1.

For "Primary treatment", ChatGPT provided accurate suggestions in 58 cases and inaccuracies in 10 cases, resulting in TP = 58, TN = 0 (as NCCN is always correct), FP = 10, and FN = 0. This yielded metrics of 100% sensitivity (0.97–100 95% CI), 85.3% accuracy (0.78–0.92 95% CI), and an F1 Score of 0.92 (Table 3).

For "Adjuvant treatment", ChatGPT's suggestions were accurate in 65 cases and inaccurate in 3 cases, leading to 65 TP, 0 TN (as NCCN is always correct), 3 FP, and 0 FN. The resulting metrics were 100% sensitivity (0.98–100 95% CI), 95.59% accuracy (0.76–0.95 95% CI), and an F1 Score of 0.96 (Table 4).

Regarding "Follow-up Indication," ChatGPT's suggestions were accurate in 64 cases and inaccurate in 4 cases, resulting in 64 TP, 0 TN (as NCCN is always correct), 4 FP, and 0 FN. This yielded 100% sensitivity (0.88–100 95% CI), 94.12% accuracy (0.88–100 95% CI), and an F1 Score of 0.94 (Table 5).

Table 5 Performance metrics for ChatGPT in comparison to NCCN guidelines, including sensitivity, specificity (not applicable), overall accuracy, and F1 score for the follow-up (FU) categories and the evaluation of Confidence Interval (CI)

	NCCN FU	95% CI
ChatGPT sensitivity	94.12	0.88–1.00
ChatGPT specificity	Not applicable	Not applicable
Chat GPT accuracy	94.12	0.88–1.00
F1 score	94.12	Not applicable

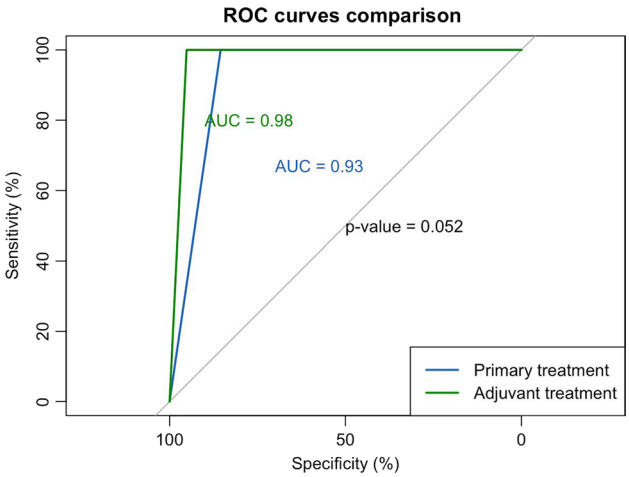

Fig. 2 ROC curves: comparison between primary treatment and adjuvant treatment

Combining data from all three scenarios, the overall sensitivity was 100%, accuracy stood at 92%, and the F1 score was 0.93. The overall precision reached 91.7%

The ROC curves comparison is illustrated in Fig. 2, where the area underneath the curves (AUC) was 0.98 for the adjuvant treatment and 0.93 for the primary treatment, with a corresponding *p*-value of 0.052.

The AUC for the adjuvant treatment was 0.98 and 0.93 for the follow-up (*p*-value of 1.000) (Fig. 3).

In Fig. 4, the AUC is 0.98 for the primary treatment and 0.93 for the follow-up, with a p-value of 0.028. The absence of specificity is acknowledged as it is not applicable in this context. The F1 score emphasizes the model's balanced performance across various NCCN stages.

Discussion

The integration of AI models, particularly LLMs like ChatGPT, into healthcare has witnessed a transformative shift in recent years [10, 13–16]. While AI already constitutes a valuable integrated tool in clinical practice across various specialties for the diagnosis and screening of

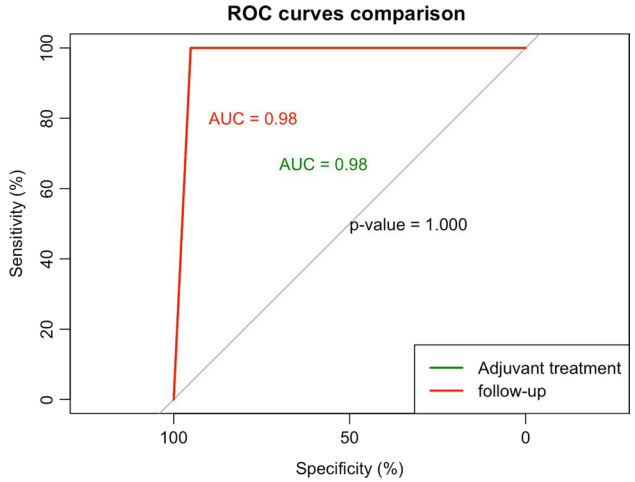

Fig. 3 ROC curves: comparison between adjuvant treatment and follow-up

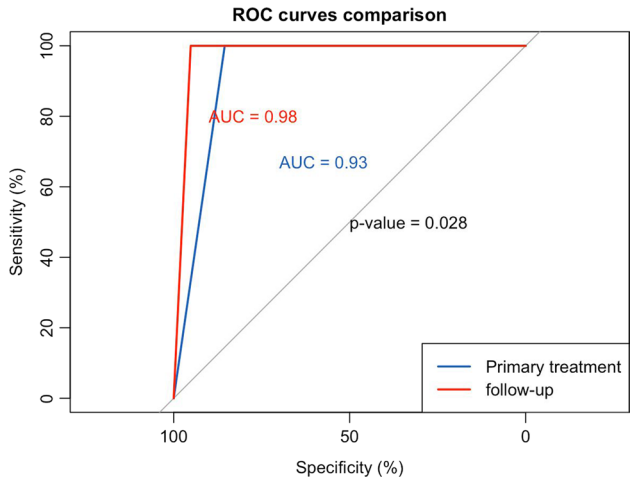

Fig. 4 ROC curves: comparison between primary treatment and follow-up

several tumors, its integration into therapeutic decision-making remains a distant prospect at present [17–20]. ChatGPT's success in tasks such as passing medical licensing exams and generating automated hospital discharge summaries underscores its potential utility in complex medical domains [8, 9, 30].

However, LLMs demonstrated significant potential in providing treatment recommendations aligned with those offered by various specialists [21, 22] and was also capable to provide patients guidance to interpret symptoms and dietary recommendations [23]. Recent strides in NLP and ML have ushered in a new era of AI applications, with ChatGPT at the forefront [7, 28, 29].

ChatGPT in cancer care: a potential paradigm shift

This study delves into the application of ChatGPT in the context of head and neck cancer treatment, scrutinizing its ability to align with the National Comprehensive Cancer Network (NCCN) Guidelines [11]. The results not only shed light on the model's performance but also open avenues for discussions on the evolving role of AI in oncological decision-making process and support. There is a dearth of prior research investigating this particular subject; however, in the case of neoplasms affecting other anatomical sites, the findings are heterogeneous [24, 25]. The results of employing ChatGPT to assist or simulate tumor board deliberations, or to offer patients guidance regarding appropriate treatments for specific tumors have exhibited notable variability (concordance ranging from 58 to 82%). These divergent results partially mirror the variability in treatment decisions, influenced not only by histologically and tumor stage but also by the genetic or somatic characteristics of the patient [26, 27].

ChatGPT emerges as a potential tool in assisting cancer patients, particularly in the intricate realm of treatment selection. Its ability to provide information, answer queries, and simplify medical terminology could enhance patient understanding and engagement [31, 32]. However, it is crucial to emphasize that ChatGPT complements, rather than replaces, professional medical advice [33–36]. Collaborative decision-making involving healthcare professionals remains paramount [37–39].

NCCN guidelines as the gold standard

The NCCN Guidelines [11], considered the "gold standard" in oncological care, provide a comprehensive framework for decision-making. Standardizing recommendations is crucial for minimizing disparities in clinical responses, especially in the context of head and neck cancers, where treatment protocols may vary based on country-specific conditions [40, 41]. Still, we cannot underline more the existing controversies since the integration of LLMs in the context of head and neck oncology, coupled with reliance on the National Comprehensive Cancer Network (NCCN) Guidelines, introduces specific risks that merit thoughtful consideration. One of the major concerns derives from the interpretability in clinical decision-making process.

Understanding how the model processes clinical nuances is vital for clinicians to trust and effectively incorporate AI-generated suggestions. Medical practitioners must cultivate a deep comprehension of LLMs architecture and operational principles, enabling judicious interpretations and fostering trust in AI-generated recommendations within the decision-making process [42]. Our study assumes the infallibility of the Guidelines, but it's important to acknowledge that guidelines themselves are subject to ongoing refinement. Those

provided by the NCCN, undergo continuous evaluation based on high-level emerging evidence and are updated at least annually, sometimes addendums are made within the same edition. Consequently, the lack of real-time adaptation in AI models may lead to potential discrepancies between model-generated suggestions and the latest evidence-based practices. For this reason, the responses currently provided by the chatbot may not be accurate in the future.

Performance evaluation of ChatGPT

Our study focused on the concordance between ChatGPT and the NCCN Guidelines in the specific context of head and neck cancer. The results demonstrated promising capabilities. Noteworthy is the model's high sensitivity, emphasizing its robustness in providing relevant suggestions across diverse scenarios. We observed a high overall accuracy of 92%, showcasing the model's effectiveness across diverse tumor sites and stages. However, we encountered 10 inaccuracies in primary treatment suggestions, as well as a few in adjuvant treatment and follow-up recommendations. These results might be explained by the fact that indications for adjuvant treatment are more standardized, even in various subsites of the head and neck, while the choice of primary treatment can vary significantly based on different anatomical characteristics of each subsite, nodal involvement, HPV/EBV status, and the evolving nature of cancer treatment protocols and approaches. In the realm of "Follow-up Indication", ChatGPT displayed reliable performance, accurately suggesting follow-up recommendations in 64 out of 68 cases. The model achieved an accuracy of 94.12%, yielding an F1 Score of 0.94. This indicates the model's proficiency in providing consistent and accurate guidance for post-treatment monitoring.

To the best of our knowledge, there is a lack of studies in the current literature addressing AI-assisted decision-making for treatment suggestions in head and neck oncology. The only study in head and neck oncology explored the chatbot capabilities in oropharyngeal cancer patients, is focused on education in diagnosis, treatment and follow-up. The authors reported a higher accuracy in "follow-up suggestion" rather than "diagnosis and treatment suggestions" [43]

While the current study illuminates the potential of ChatGPT in supporting cancer care decisions, certain considerations should be acknowledged. The absence of specificity in our evaluation, owing to the assumed infallibility of the NCCN Guidelines, underscores the need for fine distinction in interpretation. While specificity was considered inapplicable due to the assumption that NCCN Guidelines are always correct, we want to underscore that in certain real-world clinical scenarios the guidelines may not be infallible since they do not take into account several patient's features. Therefore, our study reveals knowledge gaps that hinder achieving true clinical decision support and the need for more nuanced understanding and integration of patient-specific factors into AI models. Additionally, we focused on hypothetical clinical cases necessitating of further exploration in real-world clinical settings. However, when we, as head and neck surgeons within a multidisciplinary tumor board, recommend what we consider to be the most appropriate treatment in accordance with guidelines, we take into account specific patient-related factors such as patient's age, performance status, medical comorbidities, current medications, life expectancy, quality of life, and any prior treatments [44, 45]. Most of these features are not taken into account by LLMs. Therefore, the generalizability of our findings to actual clinical practice should be approached with caution beyond the evaluated context. The efficacy of ChatGPT in dynamic healthcare settings requires careful consideration of potential contextual variations. Our goal is to demonstrate that ChatGPT, when queried appropriately, can provide accurate information for educating patients about their condition and for guiding non-trained physicians in head and neck oncology to make proper referrals. Moreover, even though we have considered clinical cases reported in NCCN guidelines and the consequent indications, the choice of treatment in everyday clinical practice is a multifactorial decision where, at least, two other considerations must be made: first of all, within the same T stage, certain extensions may preclude a specific therapeutic approach (e.g., endoscopic vs. open; transoral vs. transcervical), and this consequence, merged with the patient's will, can be a modifying factor in the final therapeutic choice. Secondly, multiple patient's comorbidities can rule out certain extensive surgical approaches or prevent the administration of combined medical therapy with radiotherapy in case the primary indication is not surgical. Consequently, the model's suggestions may lack the necessary granularity required for individualized patient care.

To enhance ChatGPT's accuracy in oncological decision-making, several considerations are crucial. Firstly, fine-tuning the model for domain-specific knowledge by incorporating diverse oncological datasets during training can improve its familiarity with intricate details of cancer treatment. Additionally, the integration of real patient data is recommended to introduce the nuances of actual clinical cases, contributing to more context-aware responses. Ensuring a dynamic integration mechanism for guideline updates allows ChatGPT to adapt to the evolving landscape of cancer care in real-time, aligning its recommendations with the latest evidence-based practices. Implementing a user feedback mechanism within ChatGPT enables continuous learning, allowing healthcare professionals to provide insights and improvements, particularly when deviations from guidelines are observed. Lastly, it is imperative to emphasize the importance of model transparency. Clinician education on AI principles is fundamental before the integration of such

models into routine practice. A deeper understanding of how ChatGPT processes information and generates responses is essential for healthcare professionals to trust and effectively incorporate its recommendations into their decision-making processes.

Nevertheless, we believe it is only a matter of time before technological advancements enable the integration of all these specific characteristics into guidelines, allowing for a personalized and specific recommendation for each head and neck cancer patient. This could be the most significant progress we may witness in the coming years. What we may not yet integrate, which currently holds great significance in the doctor-head and neck cancer patient-interaction, is the patient's preference and desire, stemming from their will to heal or not, to maintain the highest possible quality of life, perhaps at the expense of prognosis, or vice versa [30, 46, 47]. This process will require much more time and a different architecture capable of interpreting patients' expectations and desires.

Our study has few limitations and it is crucial to acknowledge these aspects as they influence the interpretation and application of our findings. First and foremost, the use of hypothetical scenarios in our evaluation may not fully capture the complexity of real-world clinical decision-making. These scenarios, while meticulously designed to encompass a broad spectrum of cases outlined in the NCCN Guidelines, inherently lack the nuances presented by individual patient characteristics and unique clinical contexts.

Looking ahead, additional rigorous validation are warranted. While our study provides valuable insights into ChatGPT's performance, further validation using real patient data and prospective studies will offer a more comprehensive understanding of ChatGPT's applicability and limitations in the dynamic landscape of oncological decision-making. To integrate AI models into clinical decision-making, it is essential to propose and undertake structured pathways, along with the validation of our findings. In the future, a prospective study assessing the alignment between a patient's informed choice, ChatGPT recommendations, and the tumor board's recommendations in various clinical cases could represent an initial step toward the adoption of ChatGPT as an additional tool in clinical practice. Subsequent evaluations could then be conducted to assess the extent of congruence among these three facets.

Conclusions

To the best of our knowledge, this study is the first to highlight ChatGPT's potential in assisting with head and neck cancer treatment decisions. Acknowledging its current limitations, future endeavors should prioritize refining the model and fostering a collaborative approach involving clinicians educated in AI principles. Its high sensitivity and overall accuracy, aligning with NCCN Guidelines, demonstrate its promise as a complementary resource for healthcare professionals. However, further validation and integration into real-world clinical settings are necessary before considering widespread adoption. As AI evolves, collaborative efforts between AI models and medical experts are crucial to unlocking new frontiers in personalized and standardized cancer care decision support, previously thought unattainable.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Informed consent No patients were enlisted in the study presented, thereby obviating the necessity for any informed consent.

Research involving human participants and/or animals No animal studies are presented in this manuscript. No potentially identifiable human images or data are presented in this study.

References

1. Gayathri Devi K, Radhakrishnan R (2015) Automatic segmentation of colon in 3D CT images and removal of opacified fluid using cascade feed forward neural network. Comput Math Methods Med. https://doi.org/10.1155/2015/670739
2. Vadhwana B, Tarazi M, Patel V (2023) The role of artificial intelligence in prospective real-time histological prediction of colorectal lesions during colonoscopy: a systematic review and meta-analysis. Diagnostics. https://doi.org/10.3390/diagnostics13203267
3. Chung CW, Chou SC, Hsiao TH, Zhang GJ, Chung YF (2024) Machine learning approaches to identify systemic lupus erythematosus in anti-nuclear antibody-positive patients using genomic data and electronic health records. BioData Min. https://doi.org/10.1186/s13040-023-00352-y
4. Sampieri C, Azam MA, Ioppi A, Baldini C, Moccia S, Kim D et al (2024) Real-time laryngeal cancer boundaries delineation on white light and narrow-band imaging laryngoscopy with deep learning. Laryngoscope. https://doi.org/10.1002/lary.31255
5. Barbour AB, Barbour TA (2023) A radiation oncology board exam of ChatGPT. Cureus 15:1–5. https://doi.org/10.7759/cureus.44541
6. Meskó B, Topol EJ (2023) The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 6:120. https://doi.org/10.1038/s41746-023-00873-0
7. Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel, Switzerland). https://doi.org/10.3390/healthcare11060887
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C et al (2023) Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Heal 2:e0000198. https://doi.org/10.1371/journal.pdig.0000198

9. Cheong RCT, Unadkat S, Mcneillis V, Williamson A, Joseph J, Randhawa P et al (2023) Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. Eur Arch Oto Rhino Laryngol. https://doi.org/10.1007/s00405-023-08319-9
10 Marchi F, Sampieri C (2023) From data analysis to paper writing: How Artificial intelligence is changing the face of scientific literature. Oral Oncol 138:106312. https://doi.org/10.1016/j.oraloncology.2023.106312
11. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®)-Head and Neck Cancers 2024;1.
12. Lydiatt WM, Patel SG, Ridge JA, O'Sullivan B, Shah JP (2017) Staging head and neck cancers. AJCC Cancer Staging Manual. https://doi.org/10.1007/978-3-319-40618-3_5
13 Yue T, Wang Y, Zhang L, Gu C, Xue H, Wang W et al (2023) Deep learning for genomics: from early neural nets to modern large language models. Int J Mol Sci. https://doi.org/10.3390/ijms242115858
14. Sampieri C, Baldini C, Azam MA, Moccia S, Mattos LS, Vilaseca I et al (2023) Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: a guide for physicians and state-of-the-art review. Otolaryngol Neck Surg 169:811–829. https://doi.org/10.1002/ohn.343
15. Nielsen JPS, von Buchwald C, Grønhøj C (2023) Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol 143:779–782. https://doi.org/10.1080/00016489.2023.2254809
16. Yoshiyasu Y, Wu F, Dhanda AK, Gorelik D, Takashima M, Ahmed OG (2023) GPT-4 accuracy and completeness against International consensus statement on allergy and rhinology: rhinosinusitis. Int Forum Allergy Rhinol. https://doi.org/10.1002/alr.23201
17. Azam MA, Sampieri C, Ioppi A, Benzi P, Giordano GG, De Vecchi M et al (2022) Videomics of the upper aero-digestive tract cancer: deep learning applied to white light and narrow band imaging for automatic segmentation of endoscopic images. Front Oncol 12:900451. https://doi.org/10.3389/fonc.2022.900451
18. Azam MA, Sampieri C, Ioppi A, Africano S, Vallin A, Mocellin D et al (2022) Deep learning applied to white light and narrow band imaging videolaryngoscopy: toward real-time laryngeal cancer detection. Laryngoscope 132:1798–1806. https://doi.org/10.1002/lary.29960
19. Zhou S, Han S, Chen W, Bai X, Pan W, Han X et al (2023) Radiomics-based machine learning and deep learning to predict serosal involvement in gallbladder cancer. Abdom Radiol (New York). https://doi.org/10.1007/s00261-023-04029-2
20 Popovic D, Glisic T, Milosavljevic T, Panic N, Marjanovic-Haljilji M, Mijac D et al (2023) The importance of artificial intelligence in upper gastrointestinal endoscopy. Diagnostics (Basel, Switzerland). https://doi.org/10.3390/diagnostics13182862
21 Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J et al (2023) Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol. https://doi.org/10.1186/s10195-023-00740-4
22. Wilhelm TI, Roos J, Kaczmarczyk R (2023) Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res 25:1–13. https://doi.org/10.2196/49324
23. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X et al (2023) An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res 25:e51300. https://doi.org/10.2196/51300
24. Gabriel J, Gabriel A, Shafik L, Alanbuki A, Larner T (2023) Artificial intelligence in the urology multidisciplinary team meeting: can ChatGPT suggest European association of urology guideline-recommended prostate cancer treatments? BJU Int. https://doi.org/10.1111/bju.16240
25 Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J (2023) Challenging ChatGPT 3.5 in senology—an assessment of concordance with breast cancer tumor board decision making. J Pers Med. https://doi.org/10.3390/jpm13101502
26. Haemmerli J, Sveikata L, Nouri A, May A, Egervari K, Freyschlag C et al (2023) ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? BMJ Health Care Informatics 30:1–7. https://doi.org/10.1136/bmjhci-2023-100775
27. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K et al (2023) Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 308:1831–1844. https://doi.org/10.1007/s00404-023-07130-5
28. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M et al (2023) Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 6:e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689
29. Choo JM, Ryu HS, Kim JS, Cheong JY, Baek S-J, Kwak JM et al (2023) Conversational artificial intelligence (chatGPT™) in the management of complex colorectal cancer patients: early experience. ANZ J Surg. https://doi.org/10.1111/ans.18749
30 Hueso M, Álvarez R, Marí D, Ribas-Ripoll V, Lekadir K, Vellido A (2023) Is generative artificial intelligence the next step toward a personalized hemodialysis? Rev Investig Clin. https://doi.org/10.24875/RIC.23000162
31. Ferreira AL, Chu B, Grant-Kels JM, Ogunleye T, Lipoff JB (2023) Evaluation of ChatGPT dermatology responses to common patient queries. JMIR Dermatol 6:e49280. https://doi.org/10.2196/49280
32. Braun E-M, Juhasz-Böss I, Solomayer E-F, Truhn D, Keller C, Heinrich V et al (2023) Will I soon be out of my job? Quality and guideline conformity of ChatGPT therapy suggestions to patient inquiries with gynecologic symptoms in a palliative setting. Arch Gynecol Obstet. https://doi.org/10.1007/s00404-023-07272-6
33. Sanchez-Ramos L, Lin L, Romero R (2023) Beware of references when using ChatGPT as a source of information to write scientific articles. Am J Obstet Gynecol 229:356–357. https://doi.org/10.1016/j.ajog.2023.04.004
34. Lecler A, Duron L, Soyer P (2023) Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging 104:269–274. https://doi.org/10.1016/j.diii.2023.02.003
35. Nune A, Iyengar KP, Manzo C, Barman B, Botchu R (2023) Chat generative pre-trained transformer (ChatGPT): potential implications for rheumatology practice. Rheumatol Int 43:1379–1380. https://doi.org/10.1007/s00296-023-05340-3
36. Dallari V, Sacchetto A, Saetti R, Calabrese L, Vittadello F, Gazzini L (2023) Is artificial intelligence ready to replace specialist doctors entirely? ENT specialists vs ChatGPT: 1–0, ball at the center. Eur Arch Oto Rhino Laryngol. https://doi.org/10.1007/s00405-023-08321-1
37. Chavez MR, Butler TS, Rekawek P, Heo H, Kinzler WL (2023) Chat generative pre-trained transformer: why we should embrace this technology. Am J Obstet Gynecol 228:706–711. https://doi.org/10.1016/j.ajog.2023.03.010
38. Ferres JML, Weeks WB, Chu LC, Rowe SP, Fishman EK (2023) Beyond chatting: the opportunities and challenges of ChatGPT in medicine and radiology. Diagn Interv Imaging 104:263–264. https://doi.org/10.1016/j.diii.2023.02.006
39. Dave T, Athaluri SA, Singh S (2023) ChatGPT in medicine: an overview of its applications, advantages, limitations, future

prospects, and ethical considerations. Front Artif Intell 6:1169595. https://doi.org/10.3389/frai.2023.1169595
40. Bullock MJ, Beitler JJ, Carlson DL, Fonseca I, Hunt JL, Katabi N et al (2019) Data set for the reporting of nodal excisions and neck dissection specimens for head and neck tumors: explanations and recommendations of the Guidelines From the International Collaboration on Cancer Reporting. Arch Pathol Lab Med 143:452–462. https://doi.org/10.5858/arpa.2018-0421-SA
41. Lewis JSJ, Adelstein DJ, Agaimy A, Carlson DL, Faquin WC, Helliwell T et al (2019) Data set for the reporting of carcinomas of the nasopharynx and oropharynx: explanations and recommendations of the guidelines from the international collaboration on cancer reporting. Arch Pathol Lab Med 143:447–451. https://doi.org/10.5858/arpa.2018-0405-SA
42. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M et al (2023) Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 6:e2336997. https://doi.org/10.1001/jamanetworkopen.2023.36997
43 Davis RJ, Ayo-Ajibola O, Lin ME, Swanson MS, Chambers TN, Kwon DI et al (2023) Evaluation of oropharyngeal cancer information from revolutionary artificial intelligence chatbot. Laryngoscope. https://doi.org/10.1002/lary.31191
44. Stone A, Liu J, Lin J, Schiff BA, Ow TJ, Mehta V et al (2023) Value of adherence to posttreatment follow-up guidelines for head and neck squamous cell carcinoma. Laryngoscope. https://doi.org/10.1002/lary.30909
45. Miller MC, Shuman AG (2016) Survivorship in head and neck cancer: a primer. JAMA Otolaryngol Head Neck Surg 142:1002–1008. https://doi.org/10.1001/jamaoto.2016.1615
46. Jabbour J, Dhillon HM, Shepherd HL, Sundaresan P, Milross C, Clark JR (2018) The relationship between role preferences in decision-making and level of psychological distress in patients with head and neck cancer. Patient Educ Couns 101:1736–1740. https://doi.org/10.1016/j.pec.2018.05.023
47. Brom L, Hopmans W, Pasman HRW, Timmermans DRM, Widdershoven GAM, Onwuteaka-Philipsen BD (2014) Congruence between patients' preferred and perceived participation in medical decision-making: a review of the literature. BMC Med Inform Decis Mak 14:25. https://doi.org/10.1186/1472-6947-14-25