



Assessment of decision-making with locally run and web-based large language models versus human board recommendations in otorhinolaryngology, head and neck surgery

Christoph Raphael Buhr^{1,2} · Benjamin Philipp Ernst³ · Andrew Blaikie² · Harry Smith⁴ · Tom Kelsey⁴ · Christoph Matthias¹ · Maximilian Fleischmann⁵ · Florian Jungmann⁶ · Jürgen Alt⁷ · Christian Brandts⁸ · Peer W. Kämmerer⁹ · Sebastian Foersch¹⁰ · Sebastian Kuhn¹¹ · Jonas Eckrich¹

Received: 26 September 2024 / Accepted: 9 December 2024 / Published online: 10 January 2025
© The Author(s) 2024

Abstract

Introduction Tumor boards are a cornerstone of modern cancer treatment. Given their advanced capabilities, the role of Large Language Models (LLMs) in generating tumor board decisions for otorhinolaryngology (ORL) head and neck surgery is gaining increasing attention. However, concerns over data protection and the use of confidential patient information in web-based LLMs have restricted their widespread adoption and hindered the exploration of their full potential. In this first study of its kind we compared standard human multidisciplinary tumor board recommendations (MDT) against a web-based LLM (ChatGPT-4o) and a locally run LLM (Llama 3) addressing data protection concerns.

Material and methods Twenty-five simulated tumor board cases were presented to an MDT composed of specialists from otorhinolaryngology, craniomaxillofacial surgery, medical oncology, radiology, radiation oncology, and pathology. This multidisciplinary team provided a comprehensive analysis of the cases. The same cases were input into ChatGPT-4o and Llama 3 using structured prompts, and the concordance between the LLMs' and MDT's recommendations was assessed. Four MDT members evaluated the LLMs' recommendations in terms of medical adequacy (using a six-point Likert scale) and whether the information provided could have influenced the MDT's original recommendations.

Results ChatGPT-4o showed 84% concordance (21 out of 25 cases) and Llama 3 demonstrated 92% concordance (23 out of 25 cases) with the MDT in distinguishing between curative and palliative treatment strategies. In 64% of cases (16/25) ChatGPT-4o and in 60% of cases (15/25) Llama, identified all first-line therapy options considered by the MDT, though with varying priority. ChatGPT-4o presented all the MDT's first-line therapies in 52% of cases (13/25), while Llama 3 offered a homologous treatment strategy in 48% of cases (12/25). Additionally, both models proposed at least one of the MDT's first-line therapies as their top recommendation in 28% of cases (7/25). The ratings for medical adequacy yielded a mean score of 4.7 (IQR: 4–6) for ChatGPT-4o and 4.3 (IQR: 3–5) for Llama 3. In 17% of the assessments (33/200), MDT members indicated that the LLM recommendations could potentially enhance the MDT's decisions.

Discussion This study demonstrates the capability of both LLMs to provide viable therapeutic recommendations in ORL head and neck surgery. Llama 3, operating locally, bypasses many data protection issues and shows promise as a clinical tool to support MDT decisions. However at present, LLMs should augment rather than replace human decision-making.

Keywords Large language models · LLM · Artificial intelligence · AI · ChatGPT · Llama · Otorhinolaryngology · ORL · Head and neck · Digital health · Chatbot · Language model

Introduction

Multidisciplinary (human) tumor boards (MDTs) are a vital component of contemporary cancer treatment, allowing for collaborative decision-making among healthcare specialists from various disciplines. These discussions ensure that patients receive individually tailored treatment

plans based on expert opinions [1]. Nevertheless, the pre- and post-processing of MDT is time consuming and may delay treatment. Moreover, geographical factors can also bias decisions as local resources differ [2].

Large Language Models (LLMs) have recently demonstrated the capacity to quickly and efficiently process complex medical data, offering accurate interpretations [3–6]. These models leverage transformer-based architectures, which allow for parallel processing of large amounts of data, making them powerful tools in addressing the limitations of traditional MDTs [7, 8]. LLMs hold the potential to either support or supplement MDT-based tumor management decisions by providing real-time, impartial recommendations based on the latest available medical literature.

Despite these advantages, the use of web-based LLMs in clinical decision-making has been hampered by concerns around data protection and patient confidentiality [9]. This is especially true in medical fields, where sensitive patient data must be handled with the utmost care [10–12]. Locally operated LLMs present a promising alternative, as they function offline, eliminating the need for data transmission over the internet and mitigating the risk of data leakage [13].

Previous studies have explored the potential of LLMs in assisting MDTs [14, 15]. Early evaluations of models like ChatGPT in ORL, head and neck surgery have shown some limitations, including a lack of personalized treatment plans and inconsistency in recommendations [16–18]. These studies were generally based on expert reviews of LLM outputs rather than direct comparisons to MDT recommendations, and they did not address whether LLM-generated information could influence MDT decisions. Furthermore, the reliance on web-based LLMs in prior research has raised concerns about data security, limiting the real-world clinical applicability of these findings.

In this study, we present the first comparison between recommendations made by a standard human MDT in ORL, head and neck surgery and those generated by two LLMs: a web-based model (ChatGPT-4o) and a locally operated model (Llama 3). By addressing the data security concerns of web-based systems, this study seeks to evaluate whether locally run LLMs can offer a secure tool to augment and improve cancer management decision-making in ORL, head, and neck surgery.

Material and methods

The study was performed as follows (Fig. 1). Twenty-five cases resembling patients with a malignant disease from ORL, head and neck surgery were designed (Table 1).

Profiles were correlated with realistic clinical scenarios from the author's clinical experience and reviewed by four experienced senior consultants. Each case was prepared in the same semantic fashion including age, sex, ECOG stadium, medical history, oncological history, drug consumption (cigarettes, alcohol), tumor entity, localization, first diagnosis, TNM-stadium, histological type, p16-, PD-L1 status, grading and treatment history to date. Each case was then presented in a multidisciplinary MDT. The implemented MDT included an otorhinolaryngologist, maxillofacial surgeon, radiation oncologist, medical oncologist, radiologist and a pathologist. MDT members were chosen from different affiliations in order to reduce geographical bias. In a virtually held MDT consultation each case was discussed and therapeutic regimen was recommended.

Subsequently, each individual tumor case was passed to two different LLMs. Different prompting strategies inspired by Griewing et al. were tried, to exclude biased output [15]. The finally implemented prompt also stated the patient's

Fig. 1 Workflow of the study

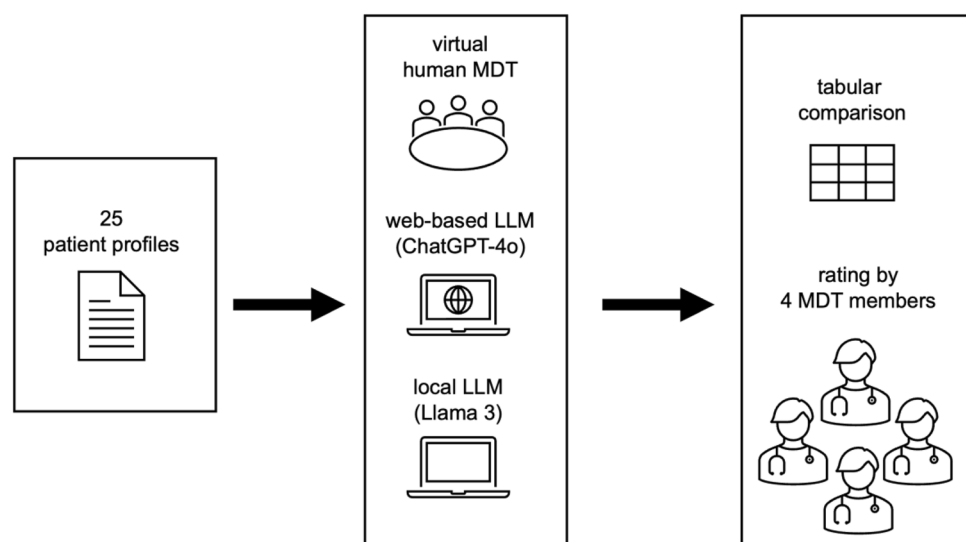


Table 1 Constructed cases

Larynx (L)								
	L1	L2	L3	L4	L5	L6	L7	L8
Age	63	70	65	50	68	66	83	37
Sex	Male	Male	Male	Male	Male	Female	Male	Male
ECOG-score	1	3	1	0	3	1	4	1
Relevant illnesses	Coronary heart disease, stent (age 60), chronic obstructive pulmonary disease, Gold III	Dry alcoholic, liver cirrhosis, Child–Pugh A	Coronary heart disease, myocardial infarction (age 58)	Depression	Arterial hypertension	–	Coronary heart disease; peripheral artery disease (iliac stent insertion 10 years ago); hypertension, atrial fibrillation, heart failure NYHA III, chronic renal failure, oral anticoagulation with Xarelto + ASA	–
Oncological history	–	–	–	–	T3 N2c M0 transoral tumor resection of an oropharyngeal carcinoma, neck dissection on both sides, adjuvant radiation therapy 10 years ago	–	Prostate cancer 2020, transurethral resection of the prostate	–
Smoking	60 pack years	40 pack years	35 pack years	37 pack years	20 pack years	–	50 pack years (stopped 20 years ago)	55 pack years
Drinking	–	Dry	–	–	Active	–	Occasionally	–
Tumor entity	Laryngeal cancer	Laryngeal cancer	Laryngeal cancer	Laryngeal cancer	Laryngeal cancer	–	Laryngeal cancer	Laryngeal cancer
Localization	Anterior commissure	Supra glottic	Anterior commissure/vestibular fold	Glottic vocal fold left	–	–	–	–
First diagnosis	2 month ago	This month	This month	This month	This month	This month	This month	This month
TNM first diagnosis	cT1b cN0 cM0	cT4a cN1 cM0	rcT3 cN1 cM0	cT1a N0 M0	cT4cN2bcM0	cT2 cN0 cM0	cT3 cN1 cM0	cT2 cN0 cM0
TNM	–	–	–	–	–	–	–	–
Histological subtype	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma	Chondrosarcoma	Squamous-cell carcinoma	Squamous-cell carcinoma

Table 1 (continued)

Larynx (L)								
	L1	L2	L3	L4	L5	L6	L7	L8
Previous therapy and course	Microsurgery	Emergency tracheostomy, biopsy	Laser resection T2N0M0 (5 months ago) Control panendoscopy (1 week ago)	Microsurgery	Panendoscopy 1. Vocal fold on the right 2. Pocket fold on the right 3. Arytenoid cartilage on the right 4. Thyroid cartilage on the right, each with squamous-cell carcinoma	Panendoscopy + biopsy (Arytenoid cartilage left)	Panendoscopy 1st vocal fold right, 2. Right vocal fold, left vocal fold, left vocal fold, Arytenoid cartilage on both sides (all carcinoma positive)	Panendoscopy with biopsy
p16	Negative	Negative	Negative	Positive	Negative	Negative	Negative	Positive
PD-L1	Not determined	TPS 0, CPS 0, IC 0	Not determined	Not determined	Not determined	Not determined	Not determined	Not determined
Grading	Moderately	Poorly	Moderately	Well	Poorly	Moderately	Moderately	Moderately
Larynx (L)								
	L9	L10	N1	N2	N3	M1	M2	M3
Age	66	55	71	73	85	60	39	63
Sex	Female	Female	Female	Male	Male	Female	Male	Male
ECOG-score	2	1	1	0	1	0	0	1
Relevant illnesses	Arterial hypertension; diabetes mellitus type II, chronic obstructive pulmonary disease; gold III	Depression, fibromyalgia	Arterial hypertension	–	Arterial hypertension	Arterial hypertension	–	Arterial hypertension, diabetes mellitus type II, polynuropathy, benign prostatic hyperplasia
Oncological history	–	–	–	–	–	–	–	–
Smoking	80 pack years	–	–	–	–	40 pack years	–	55 pack years
Drinking	–	–	–	–	–	Dry	–	Active
Tumor entity	Laryngeal cancer	Laryngeal cancer	Nasal cancer	Nasal cancer	Nasal cancer	Tongue cancer	Tongue cancer	Oral Cancer
Localization	–	–	Septum left	Right ethmoid bone	Right maxillary sinus	Left tongue edge	Left tongue edge	Mouth bottom left (crossing the midline)
First diagnosis	6 years ago	This month	1 year ago	This year	3 years ago	This year	This month	This month
TNM first diagnosis	cT1a cN0 cM0	Carcinoma in situ left	cT2N0M0	cT4a cN0 cM0	–	cT1cN0cM0	T1N0	cT4a cN2c cM0
TNM	rcT3 cN2b cM1 (pulmonary singular focus)	–	–	–	rcT1cN0cMx	–	pT3, pN3b(23/69, ECE+), L1, V0, Pn1, R0	cT4a cN2c cM0

Table 1 (continued)

	Larynx (L)		Nose (N)		Mouth (M)				
	L9	L10	N1	N2	N3	M1	M2	M3	
Histological subtype	Squamous-cell carcinoma	Cacinoma in situ	Squamous-cell carcinoma	Squamous-cell carcinoma after inverted papilloma	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma	
Previous therapy and course	pT1a cN0 cM0 vocal fold carcinoma, left chordectomy 6 years ago Control panendoscopy (evidence of carcinoma in situ and renewed laser resection alio loco 5 years ago), continued nicotine consumption, tumor follow-up irregularly observed Panendoscopy 1st vocal fold on the right, vocal fold on the left, pocket fold on the left, arytenoid hump on both sides (all carcinoma positive)	Panendoscopy, vocal fold stripping	Paranasal sinus revision on both sides Paranasal sinus mapping on both sides: tumor nasal septum left remaining biopsies tumor-free	Paranasal sinus revision on both sides Paranasal sinus mapping on both sides	Paranasal sinus surgery Caldwell Luc surgery and endonasal paranasal sinus surgery Revision paranasal sinus surgery: Squamous-cell carcinoma	Biopsy	Panendoscopy and biopsy transoral tumor resection, neck dissection on both sides, radial flap	Panendoscopy and biopsy	
p16	Negative	Not determined	Negative	Negative	Positive	Negative	Negative	Negative	
PD-L1	TPS 50, CPS 60, IC 10	Not determined	Not determined	Not determined	Not determined	Not determined	Not determined	TPS 90, CPS 100, IC 10	
Grading	Moderately	Not determined	Well	Moderately	Poorly	Moderately	Moderately	Moderately	
Pharynx (P)									
Cancer of unknown primary (CUP)									
	P1	P2	P3	P4	P5	CUP1	CMFS1	CMFS2	CMFS3
Age	60	63	81	45	72	40	64	60	50
Sex	Male	Female	Male	Male	Female	Female	Male	Male	Male
ECOG-score	3	0	1	0	1	0	1	1	2

Table 1 (continued)

	Pharynx (P)					Cancer of unknown primary (CUP)			Craniomaxillofacial surgery (CMFS)		
	P1	P2	P3	P4	P5	CUP1	CMFS1	CMFS2	CMFS3		
Relevant illnesses	Chronic obstructive pulmonary disease	–	Hypothyroidism, arterial hypertension, chronic obstructive pulmonary disease: gold II, coronary heart disease, peripheral arterial disease, left thigh (stent insertion)	–	Arterial hypertension, diabetes mellitus type II, adipositas per magna	–	Arterial hypertension, alcoholic, obesity	Arterial hypertension, Diabetes mellitus type II, adipositas	Arterial hypertension, adipositas, benign prostatic hyperplasia		
Oncological history	Chordectomy: squamous-cell carcinoma vocal fold left Panendoscopy with recurrence Laryngectomy, neck dissection on both sides, tracheotomy	–	pTA low grade urothel carcinoma (Transurethral resection of the urinary bladder)	–	pT2, pN0 Tonsil carcinoma right Tumor resection + neck dissection radial flap + right 2014 Ductal carcinoma in situ Surgery + sentinel (–)	–	–	–	–		
Smoking	48 pack years	–	50 pack years	–	60 pack years	–	30 pack years	40 pack years	20 pack years		
Drinking	–	–	Active	–	–	–	Active	Active	–		
Tumor entity	Tongue cancer left	Oropharynx cancer	Hypopharynx cancer	Tonsil cancer	–	Cancer of unknown primary	Cancer bottom	Mandibula cancer	Squamous-cell carcinoma		
Localization	Base of tongue left	Oropharynx: soft palate left with infiltration in nasopharynx	right	Tonsil left	Right tongue edge	Lymph node metastasis on the left	Mouth bottom right	Left mandible, tongue infiltration, midline crossing	Regio 16 17		
First diagnosis	This month	5 years ago	This month	This month	10 years ago	This month	2 years ago	This week	This month		
TNM first diagnosis	cT2 cN0 cM0 p16-	cT3N2 p16+	cT3 cN3 cM0	cTX cN1 cM0	pT2 pN0 cM0	cTX N2b cM0	cT1 cN1 M0	cT4 cN3b cM1 (pulmonary, bilateral foci do not match lung cancer, multiple foci)	cT1 cN0 cM0		

Table 1 (continued)

	Pharynx (P)					Cancer of unknown primary (CUP)	Craniomaxillofacial surgery (CMFS)		
	P1	P2	P3	P4	P5		CMFS1	CMFS2	CMFS3
TNM	–	rT4a, N1, L1, V0, Pn0, R1, M0	–	pT1pN1cM0	cT4 cN2 cM1 (pulmonary multiple), p16–	cTX N2b cM0	–	–	–
Histological subtype	Squamous-cell carcinoma	Squamous-cell carcinoma	–	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma	Squamous-cell carcinoma
Previous therapy and course	Panendoscopy and biopsy	Primary chemoradiotherapy 5 years ago	–	Lymph node extirpation 01/2024 Cancer of unknown primary, panendoscopy + tonsillectomy: squamous-cell carcinoma Tonsil 02/2024	–	Lymph node extirpation level II (for suspected lymphoma)	Biopsy one week ago	Biopsy one week ago	Leukoplakia for 2 years. Biopsy 2 weeks ago: squamous-cell carcinoma
p16	Negative	Positive	–	Positive	Negative	Positive	Not determined	Not determined	Negative
PD-L1	Not determined	TPS 0, CPS 0, IC 0	–	Not determined	Not determined	Not determined	Not determined	Not determined	Not determined
Grading	Moderately	Moderately	–	Moderately	Moderately	Moderately	Well	Poorly	Moderately

The bold font indicate important information

details including age, sex, ECOG stadium, medical history, oncological history, drug consumption (cigarettes, alcohol), tumor entity, localization, first diagnosis, TNM-stadium, histological type, p16-, PD-L1 status, grading and therapy to date. If applicable, the prompt also mentioned the surgeon's opinion on resection of the tumor, before mentioning that

the patient is presented in an MDT and referring to surgical-, radio-, systemic- and palliative therapy options and asking for a treatment plan for each entity if applicable. Moreover, the prompt required a clear ranking of suitable treatments. Finally, the LLMs were forced to limit the answer to 200 words (Fig. 2). After designing the prompt, patient profiles

Fig. 2 Exemplary prompt for patient (L1) translated by deepL.com (Cologne, Germany)

How should the following patient with a head and neck tumor be treated based on the latest and most relevant international literature?

Age: 63 years

Gender: male

ECOG: 0

Pre-existing conditions: Coronary heart disease, stent (at 60 years of age), chronic obstructive pulmonary disease Gold III

Oncological history: -

Smoking: 60 pack years

Alcohol: -

Tumor Entity: Laryngeal carcinoma

Localization: Anterior commissure

Initial diagnosis: 2 months ago

TNM stage at initial diagnosis: cT1b cN0 cM0

Histologic type: Squamous cell carcinoma

p16: negative

PD-L1: not determined

Grading: moderate

Previous therapy and course:

Microlaryngoscopy and sampling already performed

The surgeon considers surgical resection to be possible.

The case should be presented in an interdisciplinary tumor board.

Please prepare a step-by-step recommendation that takes into account further surgical therapy, radiotherapy, systemic therapy or palliative therapy.

If surgical therapy is an option, please draw up a treatment plan.

If radiotherapy is an option, please draw up a treatment plan.

If systemic therapy is an option, please draw up a treatment plan.

If only palliative therapy is available, please draw up a treatment plan.

Please recommend a clear order of priority as to which therapy is indicated as first-line therapy, whether there are equivalent or subordinate therapy concepts.

Please limit your answer to 200 words.

were passed to the latest web-based version of OpenAI's ChatGPT-4o and the locally running Llama 3 LLM without further training. The local LLM was run on a standard MacBook Pro 2023, M3 Pro, 18 GB RAM (Apple, California, USA).

Recommendations of the human tumor board, web-based ChatGPT-4o and locally Llama 3 were accumulated in an Excel sheet to allow a side-by-side comparison and analyzing concordance. Moreover, 4 members of the MDT rated recommendations provided by ChatGPT-4o and Llama 3 for medical adequacy on a 6-point Likert-scale (0 = very poor, 6 = excellent). Furthermore, the 4 raters stated if, in their opinion, the information provided by the LLM might have influenced the recommendation of the MDT (yes/no/abstention). Moreover, there was space for raters to express free text comments.

Statistical analysis of Likert-scale ordinal variables (qualitative-ordinal) was performed with GraphPad Prism for MacOS Version 10.3.0 (GraphPad Software, La Jolla, CA, USA). Normality distribution was tested with the D'Agostino and Pearson test. Group comparisons were performed using the Mann Whitney Test. Ethics approval was not required as no real patient data was processed.

Results

An overview of the therapy recommendations as stated by the MDT, ChatGPT-4o and Llama 3 as well as their recommendation on curative or palliative therapy regimen and agreement in first-line therapy regimen for each specific case are shown in Table 2.

Table 3 summarizes the comparison between the MDT, ChatGPT-4o and Llama 3. The human tumor board suggested a curative therapy regimen for 88% (22/25), ChatGPT-4o for 72% (18/25) and Llama 3 88% (22/25) of the cases. Thus, a palliative strategy was stated for 12% (3/25) of the cases by the human tumor board and Llama 3, while ChatGPT-4o suggested a palliative scenario 28% (7/25). ChatGPT-4o and the human tumor board showed a concordance for this aspect of 84% (21/25). In all cases of discordance ChatGPT-4o favored a palliative therapy whereas the human board suggested a curative therapy regimen. Llama 3 showed a concordance of 92% (23/25) with the human board.

In 64% (16/25) of the instances ChatGPT-4o and in 60% of cases (15/25) Llama 3 mentioned all first-line recommendation provided by the human board as a possible therapy regimen, although sometimes in a different order. Whereas ChatGPT-4o stated at least one of the first-line recommendations in another 12% (3/25) of the instances, Llama 3 did so in 8% (2/25) more instances. While ChatGPT-4o stated all first-line therapy regimen as stated by

the human board in 52% (13/25), Llama 3 showed this high concordance in 48% (12/25). Both, ChatGPT-4o and Llama 3 stated at least one of the human board's first-line therapy regimen as first-line regimen in 20% (5/25) of the cases each.

The rating of 4 MDT revealed a mean score of 4.7 (IQR 4–6) for ChatGPT-4o and a significant inferior ($p < 0.001$) mean score of 4.3 (IQR 3–5) for Llama 3 on medical adequacy (Fig. 3). For one patient, two raters concluded that the additional information from the LLM might have influenced the MDT decision. In this case, the patient had ECOG stage 3 and the MDT had suggested a curative treatment regimen. Overall, the raters stated that the information from the LLM could have influenced the decision of the tumor board in 17% (33/200) of the cases. However, this statement varied beyond raters. One rater stated that the information from the LLMs may have influenced the decision in for 32% (8/25) of the cases for ChatGPT-4o and Llama 3, respectively. In contrast, another rater saw no possible influence on the MDT decision in any case.

Whereas the human tumor board usually provided detailed information about the surgery, the LLMs often recommend surgical therapy without specification or the extent of surgery. Although web-based ChatGPT-4o and locally run Llama 3 received exactly the same prompts, ChatGPT-4o responded in German whereas Llama 3 answered in English language. While ChatGPT-4o met the word limit every single time (mean 150 words), Llama 3 provided too long answers and met the word limit just once (mean 246 words). Once Llama 3 pretended to "accept" the word limit by providing "Word count: 199" at the end of the response, in this case the response was 248 word long.

Discussion

Previous monocentric studies evaluated the feasibility of LLMs in providing recommendations for tumor treatment regimes in ORL, head and neck surgery [16–18]. However, most of these studies were dedicated to older versions and mainly focused solely on ChatGPT, with one exception evaluating Claude 3.

Our study however is the first one evaluating the web-based ChatGPT-4o and, in order to address data protection queries, the locally run LLM (Llama 3). In contrast to prior studies, recommendations of the LLMs were directly compared to human tumor board decisions and subsequently rated by experts. More specific, four MDT members evaluated the LLMs' recommendations for medical adequacy using a 6-point Likert scale and assessed whether the provided information could have influenced the MDT's final decision.

Table 2 Therapy recommendations of MDT, ChatGPT-4o and Llama 3

Patient	Surgery feasible (according to surgeon)	MDT	ChatGPT	Llama	Curative vs palliative therapy plan			First-line therapy of MDT mentioned		First-line therapy of MDT prioritized	
					MDT	ChatGPT	Llama	ChatGPT	Llama	ChatGPT	Llama
L1	yes	surgery (partial resection) vs prim. RT	1. surgery (adj. RT) 2. RT - no system therapy or pall. therapy	1. surgery (adj. (C)RT) 2. RT - no system therapy - pall. Therapy as last resort	cur	cur	cur	yes	yes	partly	partly
L2	yes	1. surgery +adj. (C)RT (+CT: R1, >3LN, ECE) 2. LE, def. prim. (C)RT (platin) if bad GC -> taxan evaluation	1. pall. RT 2. pall. IT + RT vs CT	1. LE + ND 2. RCT 3. pall. RT	cur	pall	cur	no	yes	no	yes
L3	yes	LE + ND bs, TT + (adj.) vs pRCT (platin) incl. TT (PEG)	1. surgery (part./total LE + ND)+ adj RCT 2.prim. RCT	1. surgery + (adj. RT) 2. RT (CT) - pall. therapy as last resort	cur	cur	cur	yes	partly	partly	partly
L4	yes	panendoscopy, CT staging: 1. Surgery 2. prim. RT HPV-Status	1. surgery (lokal / part LE) 2. prim. RT	1. surgery (transoral /open) + (adj. R(C)T vs immunotherapy) - pall. therapy as last resort	cur	cur	cur	yes	yes	yes	yes
L5	yes	cur. therapy dep. GC: 1. Surgery: LE, ND +(adj.) 2. prim. RCT 3. IT, PDL1-Status	1. pall. IT, (pall. RT) 2. surgery GC improves - no prim. RT	1. pall. care + symptom control 2. CT 3. pall. RT 4. surgery (LE + ND)	cur	pall	pall	no	no	no	no
L6	yes	part. LE (open/laser dep view), LE if R0 not feasible	1. surgery (adj. RT) - pall. therapy as last resort	1. surgery 2. RT adj. or salvage - pall. Therapy as last resort	cur	cur	cur	yes	yes	yes	yes
L7	yes	ECOG 4 -> pall plan: best supportive Care (TT) + (pall. RT)	1. (TT) + pall. RT 2. pall care	1. pall. RT 2. CT + (IT) 3. Surgery 4. Pall. care	pall	pall	pall	yes	yes	yes	yes
L8	yes	1. surgery (laser) (if R0 feasible) 2. prim. RT no smoking HPV-Status	1. surgery: part. LE, ND (adj. RT) 2.prim. RT	1. surgery + adj. RT (IT) vs pall. care	cur	cur	cur	yes	yes	yes	yes
L9	n.a.	1. Bx lung (VATS incl resection) -> MDT: NSCLC vs metastasis -> cur RCT vs pall. IT	1. pall. IT and/or CT 2. pall. RT 3. surgery only tracheal stent) 4. pall. care	1. prim. RCT 2. prim. CT 3. surgery: local control 4. pall. care	cur	pall	cur	no	no	no	no
L10	yes	If R0 -> follow up vs surgery (laser)	1. surgery + follow-up (3 month interval for 2y) 2. adj. RT	1. surgery (adj. RT) 2. prim. RT - (CT or IT) - pall. Therapy as last resort	cur	cur	cur	yes	yes	yes	yes
N1	yes	surgery (open SRP /part. ablatio nasi) vs prim. RT	1. surgery (adj. RT) 2. prim. RT	surgery (adj. RT) vs prim. RT; (IT or CT)	cur	cur	cur	yes	yes	partly	yes
N2	yes	prim. CRT (Platin) vs radical surgery	1. surgery (adj. R(C)T) 2. pall. IT	1 surgery + (adj.CT and/or RT) 2. RCT 3. prim. CT 4. pall. therapy as last resort	cur	cur	cur	partly	yes	partly	partly

Our findings reveal high concordance between the LLMs and the MDT, particularly in the critical distinction between curative and palliative therapy strategies. Llama 3, the locally run model, exhibited a 92% (23/25) concordance

rate with the MDT, while ChatGPT-4o achieved 84% (21/25). ChatGPT-4o mentioned all first-line recommendation of the MDT in 64% (16/25) of the patients, Llama in 60% (15/25). While ChatGPT-4o stated the same first-line

Table 2 (continued)

Patient	Surgery feasible (according to surgeon)	MDT	ChatGPT	Llama	Curative vs palliative therapy plan			First-line therapy of MDT mentioned		First-line therapy of MDT prioritized	
					MDT	ChatGPT	Llama	ChatGPT	Llama	ChatGPT	Llama
N3	yes	1. surgery (same entry) 2. RT	1. surgery (adj. RT) 2. consider IT in case of progress	1. surgery (open+transoral) + (adj. RT) 2. RT 3. syst therapie 4. pall. therapy as last resort	cur	cur	cur	yes	yes	yes	yes
M1	yes	panendoscopy: surgery + incl. ipsilat. ND (after surgery -> MDT regarding adj.)	1. surgery (ND) + (adj. RT)	1. surgery vs CRT (CT neoadj/adj) - pall. therapy as last resort	cur	cur	cur	yes	yes	yes	partly
M2	no	adj RCT (platin), re-staging 3 month after therapy	1. CRT 2. IT (after PDL1 determination)	1 surgery (part. glossektomie, ND left, radial flap, adj. RT) - p16 + PDL1 determination - (IT) - pall. therapy as last resort	cur	cur	cur	yes	no	yes	no
M3	no	prim. CRT + (clinical trial)	1. CRT 2. RT + IT 3. Pall. therapy	1. neoadj. CRT + surgery (if feasible) 2. CRT 3. RT 4. pall care + CT (platin)	cur	cur	cur	yes	no	yes	no
P1	yes	surgery (laser, transoral) vs pCRT (if no CRT history)	1. surgery (part. paryngoto) + adj R(C)T 2. pall CT or IT	1.surgery (lat. pharyngotomie) + (ND) +(adj (C)RT) 2. CRT 3. IT 4. pall. therapy as last resort	cur	cur	cur	partly	partly	partly	partly
P2	no	1.re-CRT 2. pall. therapy (platin doublette)	1. IT (+ CT) 2. (C)RT 3. pall. therapy	1. system CT + IT 2. pall. therapy	cur	cur	cur	partly	no	no	no
P3	n.a.	Panendoscopy (if squamous-cell carcinoma: surgery incl LE, part. pharyngotomy, ND bs, flap vs prim. RCT)	1. pall. RT vs pall. IT + CT 2. RT	Endoscopy, Bloodtest, CT head & neck 1. prim. RCT 2. IT 3. pall. care	cur	pall	cur	no	yes	no	yes
P4	n.a.	ND ipsilateral + (re-resection if R0 + <3LN, discuss -> adj. RT vs follow-up) HPV Status (clinical trial)	1. prim CRT 2. RT 3. CT + IT	1. Surgery: (tonsillectomy +lateral pharyngotomy) + ND (Level II-III) + adj. (C)RT	cur	cur	cur	no	yes	no	yes
P5	no	TT dep. GC 1. PDL1 Status -> CT vs IT 2. consider RT after 3 cycles	1. pall. IT + pall. CT 2. pall. RT	1. pall RCT / (prim. CRT) 2. CT	pall	pall	pall	yes	no	yes	no
CUP1	n.a.	PET-CT + CUP-Panendo + ND ipsilateral/bs (depending on PET-CT); depending on histology -> (C)RT HPV-status (clinical trial)	1. prim. (C)RT (surgery ND, later IT)	1. surgery 2. (C)RT 3. CT or IT 4. Pall care	cur	cur	cur	no	no	no	no
CMFS1	yes	Panendo: 1. surgery, ND bs 2. prim. CRT	1. surgery: ipsilateral ND + ((C)RT) 2. pall. IT	1. surgery: adj. RT 2. prim. CT + (IT)	cur	cur	cur	yes	yes	yes	yes
CMFS2	no	TT+IT vs CT + IT depending on PDL1 Status MDT after 3 cycles -> re-evaluation CRT	1. pall. IT + CT (dep. PDL-status mono IT) 2. pall. RT - pall. support	1. prim. (C)RT 2. (CT (platin doublette)) 3. Pall. Care	pall	pall	cur	yes	no	yes	no
CMFS3	yes	Panendo, surgery, clinical follow-up	1. surgery +(ND) + (adj. (C)RT)	1. surgery + ND + adj. CRT 2. prim. CRT 3. (pall. care)	cur	cur	cur	yes	yes	yes	yes

MDT (human) multidisciplinary tumor board recommendations; *prim* primary; *adj* adjuvant; *pall*. palliative; *cur* curative; *RT* radiation therapy; *CRT* chemoradiation therapy; *CT* chemotherapy; *IT* immunotherapy; *R* resection margin; *LN* lymph node; *ECE* extra capsular expansion; *GC* general condition; *LE* laryngectomy; *ND* neck dissection; *bs* both sides; *TT* tracheotomy; *PEG* percutaneous endoscopic gastrostomy; *CT* computerized tomography; *HPV* humane papillomavirus; *ECOG-Status* Eastern Cooperative Oncology Group-Status; *Bx* Biopsy; *VATS* video-assisted thoracic surgery; *SRP* septorhinoplastic

The color increases readability substantially

Table 3 Summarized comparison of therapy recommendation

Curative vs palliative therapy regimen					
MDT		ChatGPT-4o		Llama 3	
		Concordance with MDT		Concordance with MDT	
Curative therapy regi- men	22 patients (22/25 = 88%)	18 patients (18/25 = 72%)	84% (21/25)	22 patients (22/25 = 88%)	92% (23/25)
Palliative therapy regimen	3 patients (3/25 = 12%)	7 patients (7/25 = 28%)		3 patients (3/25 = 12%)	
First-line therapy of the MDT mentioned by the LLMs					
			ChatGPT-4o	Llama 3	
Stated all first-line therapy regimen of the MDT			16 patients (16/25 = 64%)	15 patients (15/25 = 60%)	
Stated some first-line therapy regimen of the MDT as first-line			5 patients (5/25 = 20%)	5 patients (5/25 = 20%)	
Stated all first-line therapy regimen stated as first-line like the MDT			13 patients (13/25 = 52%)	12 patients (12/25 = 48%)	
First-line therapy regimen stated					
		MDT	ChatGPT-4o	Llama 3	
First-line therapy regimen					
Surgery		12 patients (12/25 = 48%)	13 patients (13/25 = 52%)	16 patients (16/25 = 64%)	
Radiotherapy		–	3 patients (3/25 = 12%)	1 patient (1/25 = 4%)	
Chemoradiotherapy		3 patients (3/25 = 12%)	3 patients (3/25 = 12%)	2 patients (2/25 = 8%)	
(Chemo)radiotherapy		–	1 patient (1/25 = 4%)	1 patient (1/25 = 4%)	
Surgery vs. radiotherapy (equal)		2 patients (2/25 = 8%)	–	1 patient (1/25 = 4%)	
Surgery vs. chemoradiotherapy (equal)		3 patients (3/25 = 12%)	–	1 patient (1/25 = 4%)	
Immune- and/or chemotherapy		2 patients (2/25 = 8%)	5 patients (5/25 = 20%)	1 patient (1/25 = 4%)	
Further DIAGNOSTIC		2 patients (2/25 = 8%)	–	1 patient (1/25 = 4%)	
Best supportive care		1 patient (1/25 = 4%)	–	1 patient (1/25 = 4%)	
Therapy regimen considered beyond first-line					
MDT		ChatGPT-4o		Llama 3	
Therapy regimen considered beyond first-line		Concordance with MDT		Concordance with MDT	
Surgery	16 patients (16/25 = 64%)	14 patients (14/25 = 56%)	84% (21/25)	20 patients (20/25 = 80%)	76% (19/25)
Radiotherapy	10 (patients 10/25 = 40%)	21 patients (21/25 = 84%)	56% (14/25)	19 patients (19/25 = 76%)	56% (14/25)
Chemoradiotherapy	12 (patients 12/25 = 48%)	12 patients (12/25 = 48%)	76% (19/25)	14 patients (14/25 = 56%)	56% (14/25)
Chemotherapy	4 (patients 4/25 = 16%)	10 patients (10/25 = 40%)	76% (19/25)	13 patients (13/25 = 52%)	52% (13/25)
Immunotherapy	4 (patients 4 = 25 = 16%)	18 patients (18/25 = 72%)	44% (11/25)	13 patients (13/25 = 52%)	40% (10/25)

recommendations as the MDT as first-line in 52% (13/25) of the cases, Llama 3 did so in 48% (12/25). Whereas ChatGPT-4o stated the first-line recommendation of the MDT in parts as first-line in 12% (3/25) more cases, Llama 3 achieved the same in 8% (2/25) more cases. Accordingly, in 72% (18/25) ChatGPT-4o and in 68% (17/25) of the cases Llama 3 stated at least one identical MDT first-line therapy

regimen as first-line. This competence of the current LLMs both on and off-line is surprisingly sufficient, especially when considering that previous studies in ORL were more negative regarding individualized therapy recommendations of LLMs [16, 17].

First-line surgery as sole therapy was most frequently suggested by the MDT and LLMs alike, with the MDT for

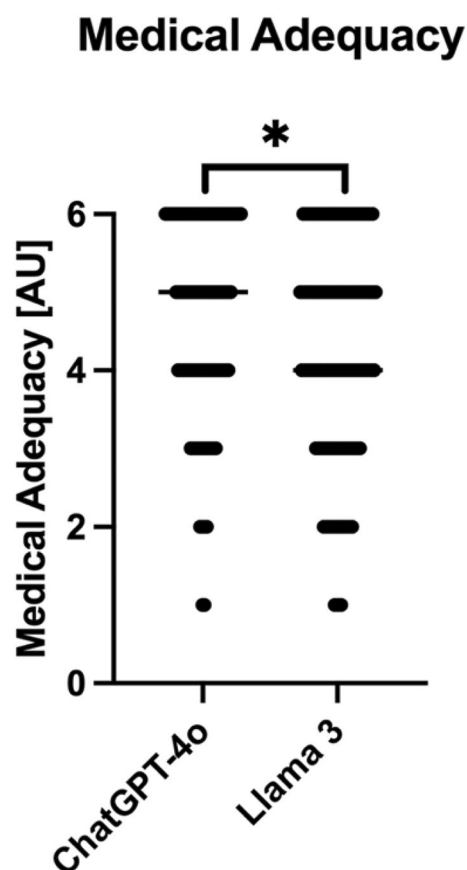


Fig. 3 Medical Adequacy for recommendations stated by ChatGPT-4o (mean 4.7; range 1–6) and Llama 3 (mean 4.3; range 1–6) as rated by the 4 MDT members. Data shown as a scatter dot blot with each point resembling an absolute value (bar width resembling a high amount of individual values). Normality distribution was tested with the D’Agostino and Pearson test. Group comparisons were performed using the Mann Whitney Test. * $p < 0.05$

48% (12/25), ChatGPT-4o for 52% (13/25) and Llama for 64% (16/25) of the cases. While the MDT provided a more detailed surgical therapy plan i.e. “local resection and neck dissection ipsilateral” the LLMs frequently lack specification of the surgery. This is particularly problematic if the LLM, like Llama did in an exemplary case (M1), concentrates on the resection of the primary tumor and does not mention a necessary neck dissection.

To provide a realistic setting, we intentionally implemented some obstacles into the patient profiles to provide a more realistic setup of flawed and/or misleading submissions to the MDT. In one case a patient with cT3 cN3 cM0 Hypopharynx Carcinoma (P3) without any actual histopathological confirmation, both ChatGPT-4o and Llama 3 suggested treatment plans whereas the MDT recommended a panendoscopy with subsequent biopsy first. In this case, ChatGPT-4o suggested a palliative therapy for a tumor which was not histologically proven at that time. Llama

3 recommended primary chemoradiotherapy as first line therapy although neither histological evaluation nor a surgeon’s statement regarding resectability was apparent.

A further pitfall for the LLMs was a cancer of unknown primary (CUP) Syndrome (CUP1). Here none of the LLMs suggested a CUP panendoscopy or a PET CT like the MDT. In addition, Llama did not realize that the primary tumor localisation is in fact unknown and therefore its resection is not feasible.

Furthermore, in one instance (M2) Llama did not realize that the patient had already undergone surgery with curative intent and that the presentation was only for adjuvant therapy. Here, one can argue that the prompt should have been adapted, as MDT submissions usually also provide information about a pre- or post-therapeutic presentation. However, the simulated case represents a scenario of an imperfect submission, which unfortunately does occur in everyday clinical practice.

Moreover, we noticed that Llama 3 ignored the German input by answering in English instead of German all the time. The same happened with regards to the word limit, not only did Llama 3 exceed the word limit nearly every time, in one case Llama 3 even claimed to respect the word limit even though it was exceeded. These examples suggest that LLMs tend to ignore instruction if they are not capable of fulfilling them, rather than admitting deficiencies. Especially for tumor treatment plans this is hazardous and should be considered when consulting LLMs in these matters. This fact underlines the inference of previous studies (in ORL and beyond) that LLMs are very promising for augmenting MDT but cannot and should not replace them [14–17].

Despite some fundamental flaws, the LLMs received respectable ratings for medical adequacy on a 6-point Likert-scale by our 4 evaluating MDT members visualized by a mean score of 4.7 (IQR 4–6) for ChatGPT-4o and 4.3 (IQR 3–5) for Llama 3 (Fig. 3). This might be explained by the fact that allegedly less important facts, such as ECOG status or smoking history, can easily fade into the background in a MDT whilst they were frequently highlighted by the LLMs. In our study, the MDT had suggested a curative treatment regimen for a patient with ECOG stage 3 (L2). Here, two raters stated that the information from the LLM might have changed the recommendation of the MDT. Raters found it positive that the LLMs highlighted chronic diseases of the patients and addiction history. Overall, raters stated in 17% (33/150) of the ratings, that the information of the LLM might have changed the recommendation of the board. In individual cases, the use of LLMs can therefore potentially have a very positive influence on the fate of a patient. Particularly when considering the fact that the resources for an MDT can vary enormously depending on the environment of MDTs [2].

Obviously, this study has some limitations, including the fact that constructed cases instead of real patient profiles were used. However, even after anonymization of data sets, entry of real cases into a web-based LLM was not approved by our data protection officer. As a comparison of the web-based LLM ChatGPT-4o and the locally run LLM Llama 3 was a key objective of this study we were limited to constructed patient profiles. Moreover, one could argue that a MDT resembling members from different affiliations is no realistic scenario, however, we choose this design by default in order to limit subjective and institutional biases within the MDT decision. In fact, we noticed local differences in default treatment regimens, for instance whereas in one institution a dental renovation is implemented prior to radiation therapy, other institutions waive the step in order to accelerate the start of the treatment. This emphasizes exemplary the virtue of a multicentric approach.

Taking these limitations into account, the present study is the first study evaluating a locally run LLM (Llama 3) on ORL, head and neck surgery MDT recommendations. Interestingly, the results of the local LLM (Llama 3) were inferior to the web-based ChatGPT-4o. With regard to the decision on a curative or palliative procedure, the agreement between the MDT and the local Llama 3 LLM was even higher. Since this is the first study in ORL, head and neck surgery evaluating a local LLM, the potential for optimization is certainly not exhausted yet. First, newer versions and the comparison of different local LLMs might improve results. Secondly, special training on specific ORL, head and neck literature and aligned prompting might further improve the outcome. It is therefore likely that local LLMs will catch up with the web-based versions in the near future. Especially in medicine, passing highly sensitive private patient data to privately owned, web-based LLMs is hardly imaginable in clinical practice [9]. Despite data privacy, the open-access structure of local LLMs is quite cost-effective requiring just a sole standard computer available for less than €1000. Furthermore, local LLMs provide the opportunity of deployment in remote regions without internet connection. Thus, these models represent also a great opportunity for low-resource settings in low and middle income countries (LMICs). Future research regarding medical application should therefore focus on local LLMs.

With regards to the local and open-source character of Llama 3, the results of this study might reflect an important step to actual implementation of locally run, data protection compliant LLMs into real clinical practice. Despite our promising findings, it is unlikely that MDTs are close to being replaced by LLMs. Human assessment in tumor board decision-making currently proves more reliable than LLM-driven assessments, especially as they will be reluctant to put any decision making regarding their therapeutic faith in the hands of a computer [19]. Additionally, the medicolegal

implications are entirely unclear at this point. Rather MDTs are likely to be augmented by locally run LLMs helping to reduce geographical bias of tumor boards [2, 20]. Future studies should thus evaluate the integration of (local) LLMs in the MDT process.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The relevant data to support the results are included in the manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Mano MS, Çitaku FT, Barach P (2021) Implementing multidisciplinary tumor boards in oncology: a narrative review. *Future Oncol* 18(3):375–384
2. Berardi R, Morgese F, Rinaldi S, Torniai M, Mentrasti G, Scortichini L et al (2020) <p>benefits and limitations of a multidisciplinary approach in cancer patient management</p>. *Cancer Manag Res* 12:9363–9374
3. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider CR, Forte AJ (2024) Clinical and surgical applications of large language models: a systematic review. *J Clin Med* 13(11):3041
4. Liu J, Wang C, Liu S (2023) Utility of ChatGPT in clinical practice. *J Med Internet Res* 25:e48568
5. Buhr CR, Smith H, Huppertz T, Bahr-Hamm K, Matthias C, Blaikie A et al (2023) ChatGPT versus consultants: blinded evaluation on answering otorhinolaryngology case-based questions. *JMIR Med Educ* 9:e49183
6. Buhr CR, Smith H, Huppertz T, Bahr-Hamm K, Matthias C, Cuny C et al (2024) Assessing unknown potential—quality and limitations of different large language models in the field of otorhinolaryngology. *Acta Otolaryngol* 144(3):237–242
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al (2017) Attention is all you need. *Advances in neural information processing systems*. <https://doi.org/10.48550/arXiv.1706.03762>
8. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
9. Meskó B, Topol EJ (2023) The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 6(1):120. <https://doi.org/10.1038/s41746-023-00873-0>
10. Knoppers BM, Thorogood AM (2017) Ethics and big data in health. *Curr Opin Syst Biol* 4:53–57

11. Mello MM, Lieou V, Goodman SN (2018) Clinical trial participants' views of the risks and benefits of data sharing. *N Engl J Med* 378(23):2202–2211
12. Price WN, Cohen IG (2019) Privacy in the age of medical big data. *Nat Med* 25(1):37–43
13. Nowak S, Sprinkart AM (2024) Große Sprachmodelle von OpenAI, Google, Meta, X und Co. *Die Radiologie*
14. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N et al (2023) Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer* 9(1):44. <https://doi.org/10.1038/s41523-023-00557-8>
15. 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* 13(10):1502. <https://doi.org/10.3390/jpm13101502>
16. Lechien JR, Chiesa-Estomba C-M, Baudouin R, Hans S (2023) Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. *Eur Arch Otorhinolaryngol* 281(4):2105–2114
17. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T et al (2024) Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for primary head and neck cancer cases. *Front Oncol* 14:1353031. <https://doi.org/10.3389/fonc.2024.1353031>
18. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T et al (2024) Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. *Eur Arch Otorhinolaryngol* 281(11):6099–6109. <https://doi.org/10.1007/s00405-024-08828-1>
19. Reis M, Reis F, Kunde W (2024) Influence of believed AI involvement on the perception of digital medical advice. *NatMed* 30(11):3098–3100. <https://doi.org/10.1038/s41591-024-03180-7>
20. Bazoukis G, Hall J, Loscalzo J, Antman EM, Fuster V, Armoundas AA (2022) The inclusion of augmented intelligence in medicine: a framework for successful implementation. *Cell Rep Med* 3(1):100485. <https://doi.org/10.1016/j.xcrm.2021.100485>

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

Authors and Affiliations

Christoph Raphael Buhr^{1,2}  · Benjamin Philipp Ernst³ · Andrew Blaikie² · Harry Smith⁴ · Tom Kelsey⁴ · Christoph Matthias¹ · Maximilian Fleischmann⁵ · Florian Jungmann⁶ · Jürgen Alt⁷ · Christian Brandts⁸ · Peer W. Kämmerer⁹ · Sebastian Foersch¹⁰ · Sebastian Kuhn¹¹ · Jonas Eckrich¹

✉ Christoph Raphael Buhr
buhrchri@uni-mainz.de

¹ Department of Otorhinolaryngology, University Medical Center of the Johannes Gutenberg-University Mainz, Langenbeckstraße 1, 55131 Mainz, Germany

² School of Medicine, University of St Andrews, St Andrews, UK

³ Department of Otorhinolaryngology, University Medical Center Frankfurt, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany

⁴ School of Computer Science, University of St Andrews, St Andrews, UK

⁵ Department of Radiotherapy and Oncology, University Medical Center Frankfurt, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany

⁶ Outpatient Department for Radiology and Nuclear Medicine at Marienhaus Hospital Saarlouis, Xcare Group, Kapuzinerstraße 4, 66740 Saarlouis, Germany

⁷ Department of Hematology and Medical Oncology, University Medical Center Mainz of the Johannes Gutenberg-University Mainz, Langenbeckstraße 1, 55131 Mainz, Germany

⁸ Department of Hematology and Medical Oncology, University Medical Center Frankfurt, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany

⁹ Department of Oral and Maxillofacial Surgery—Plastic Surgery, University Medical Center Mainz of the Johannes Gutenberg-University Mainz, Langenbeckstraße 1, 55131 Mainz, Germany

¹⁰ Institute of Pathology, University Medical Center of the Johannes Gutenberg-University Mainz, Langenbeckstraße 1, 55131 Mainz, Germany

¹¹ Institute for Digital Medicine, Philipps-University Marburg and University Hospital of Giessen and Marburg, Marburg, Germany