

# Diagnostic and Management Applications of ChatGPT in Structured Otolaryngology Clinical Scenarios

Roy W. Qu, MD<sup>1</sup> , Uneeb Qureshi, MD<sup>1</sup> ,  
Garrett Petersen, BS<sup>2</sup> , and Steve C. Lee, MD, PhD<sup>1</sup> 

OTO Open  
2023, Vol. 7(3):e67  
© 2023 The Authors. OTO Open  
published by Wiley Periodicals LLC  
on behalf of American Academy of  
Otolaryngology-Head and Neck  
Surgery Foundation.  
DOI: 10.1002/oto.2.67  
<http://oto-open.org>

WILEY

## Abstract

**Objective.** To evaluate the clinical applications and limitations of chat generative pretrained transformer (ChatGPT) in otolaryngology.

**Study Design.** Cross-sectional survey.

**Setting.** Tertiary academic center.

**Methods.** ChatGPT 4.0 was queried for diagnoses and management plans for 20 physician-written clinical vignettes in otolaryngology. Attending physicians were then asked to rate the difficulty of the clinical vignettes and agreement with the differential diagnoses and management plans of ChatGPT responses on a 5-point Likert scale. Summary statistics were calculated. Univariate ordinal regression was then performed between vignette difficulty and quality of the diagnoses and management plans.

**Results.** Eleven attending physicians completed the survey (61% response rate). Overall, vignettes were rated as very easy to neutral difficulty (range of median score: 1.00-4.00; overall median 2.00). There was a high agreement with the differential diagnosis provided by ChatGPT (range of median score: 3.00-5.00; overall median: 5.00). There was also high agreement with treatment plans (range of median score: 3.00-5.00; overall median: 5.00). There was no association between vignette difficulty and agreement with differential diagnosis or treatment. Lower diagnosis scores had greater odds of having lower treatment scores.

**Conclusion.** Generative artificial intelligence models like ChatGPT are being rapidly adopted in medicine. Performance with curated, easy-to-moderate difficulty otolaryngology scenarios indicate high agreement with physicians for diagnosis and management. However, a decreased quality in diagnosis is associated with decreased quality in management. Further research is necessary on ChatGPT's ability to handle unstructured clinical information.

## Keywords

artificial intelligence, ChatGPT, large language model, otolaryngology

Received July 13, 2023; accepted July 19, 2023.

Artificial intelligence (AI) has rapidly made inroads across many industries with health care being no exception with applications ranging from diagnosis to drug development and treatment plans. Large language models (LLMs), colloquially known as chatbots, utilize natural language processing to generate human-like conversations and have been ubiquitously described in popular media.<sup>1</sup> In medicine, it is crucial we understand this technology as a tool used by clinicians as well as by the public. As one of the most advanced and publicly available chatbots, chat generative pretrained transformer (ChatGPT) was mainly trained using publicly available data until September 2021 and has been shown to have high accuracy in generating human-like responses to a wide range of questions in seconds.<sup>2</sup> Within medicine, recent studies have shown that ChatGPT can pass the United States Medical Licensing Exam.<sup>3</sup> However, it is unclear how well ChatGPT can provide a diagnosis and treatment plan when provided in an open-ended clinical situation. Similarly, little is known about how its responses compare with human evaluation.

Early evidence demonstrates a mixed picture of the role of AI in diagnostics. One study found that AI models had sufficient accuracy in diagnosing and providing treatment plans and suggested the possibility of using AI as a tool in health care.<sup>4</sup> While another study compared the diagnostic accuracy of an AI model with that of internal medicine physicians in diagnosing and treating common chief complaints and found that the AI model had a

<sup>1</sup>Department of Otolaryngology-Head and Neck Surgery, Loma Linda University Medical Center, Loma Linda, California, USA

<sup>2</sup>Department of Otolaryngology-Head and Neck Surgery, Loma Linda School of Medicine, Loma Linda University Medical Center, Loma Linda, California, USA

## Corresponding Author:

Steve C. Lee, MD, PhD, Department of Otolaryngology-Head and Neck Surgery, Loma Linda University Medical Center, Room 2584, 11234 Anderson Street, Loma Linda, CA 92354, USA.  
Email: [stelee@llu.edu](mailto:stelee@llu.edu)

significantly lower diagnostic and treatment accuracy than human physicians.<sup>5</sup>

With public interest at a high, ChatGPT may become an increasingly utilized tool for both patients and clinicians. Its efficacy in diagnosis and treatment for specialty care, such as otolaryngology, has not been well evaluated. In this study, we compare the diagnostic accuracy and the treatment plans of ChatGPT with that of human physicians using a series of otolaryngology clinical vignettes. The results of this study highlight a relatively high accuracy in both diagnosis and treatment for common otolaryngology pathology but suggest that treatment accuracy decreases when a poor list of differential diagnoses is generated.

## Methods

### *Institutional Review Board (IRB) Determination*

This study was exempt from review by the Loma Linda University Health IRB as it does not constitute human subject research.

### *ChatGPT Interface and Survey Generation*

Twenty clinical vignettes were conceptualized and prepared by the authors of this paper. These represented clinical presentations across multiple specialties within otolaryngology with the intent of having varying degrees of difficulty. While these were uniquely created for this study primarily using clinical experience, materials such as otolaryngology textbooks and question banks were also utilized. The clinical presentations were designed to be similar to the type that would be encountered during in-service tests and board examinations. These vignettes were then fed into ChatGPT 4.0 from April 26, 2023 to April 27, 2023 in 2 stages. First, we provided the prompt in the following format and asked for differential diagnoses:

For the following scenario, give me 5 differential diagnoses in order of likelihood using primary scientific literature only. List references you used for each diagnosis. [Prompt]

After ChatGPT provided an answer, we next asked it to provide a treatment plan:

What is your treatment plan for your most likely diagnosis? Please include any consults, laboratory or radiographic studies, and treatments (medical and/or surgical) in your response

Example queries and outputs are provided in Supplemental Figure S1, available online. We collated the ChatGPT responses and created a survey asking participants to rate the difficulty of the prompt and report how much they agreed with ChatGPT outputs on a 5-point Likert scale. The scale for difficulty is as follows: very easy (1), somewhat easy (2), neutral (3), somewhat difficult (4),

and very difficult (5). The scale for agreement is as follows: strongly disagree (1), somewhat disagree (2), neutral (3), somewhat agree (4), and strongly agree (5). We distributed the survey to all attending physicians at our tertiary academic center via Research Electronic Data Capture, a secure web-based application for surveys and databases. Full ChatGPT outputs and survey are also provided in Supplemental Figure S1, available online. ChatGPT 3.5 was utilized to begin writing the introduction of this paper. The original output for the introduction is provided in Supplemental Figure S2, available online.

### *Statistical Analysis*

Data were analyzed using IBM SPSS Statistics Version 27. Summary statistics are reported for vignette difficulty, diagnostic accuracy, and treatment plan for each item and overall: median, interquartile range, and range. Univariate ordinal regression was then performed examining the relationship between vignette difficulty and diagnostic and treatment accuracy. Odds ratios with a 95% confidence interval are reported. Statistical significance was determined at  $p < .05$ .

## Results

Eleven attending physicians provided complete responses to our survey, which represents a response rate of 61%. The specialty representation included 5 head and neck surgeons, 1 facial plastics surgeon, 1 laryngologist, 1 neuro-otologist, 1 rhinologist, and 2 pediatric otolaryngologists (data not shown). **Table 1** illustrates twenty clinical vignettes with the median difficulty rating, interquartile range (IQR), and full range. The lowest median score was 1.00 for prompts #3, #9, #15, and #17. The highest median score was 4.00 prompt #5. The spread varied for each vignette with IQR ranging from 0 to 2.00. Overall, the median difficulty was 2.00 with an IQR of 1.00. This suggests that attending physicians typically viewed the vignettes as easy but there was modest variation.

**Table 2** displays the 5 differential diagnoses provided by ChatGPT for each clinical vignette and the attending physician's agreement with the provided differential diagnoses. There was high agreement across all vignettes with most achieving a median score of 5.00. The lowest median score was 3.00 for prompt #19. The IQR for each prompt was small, usually 1.00 or less. Prompt #19 had the highest spread with an IQR of 2.00. Overall, there was again high agreement across all the vignettes with a median of 5.00 and IQR of 1.00.

**Table 3** reports abbreviated management plans for the most likely differential diagnosis provided by ChatGPT, not the treatment plan best suited for the vignette. Full ChatGPT outputs are provided as supplemental information (Supplemental Figure S1, available online). Again, there was high agreement with the treatment plan for what ChatGPT felt was the most likely diagnosis with the vast majority achieving a median score of at least 4.00. The

**Table 1.** Clinical Vignettes and Physician-Rated Difficulty of Prompt

| Prompt  | Median (IQR) | Range     |
|---|--------------|-----------|
| 1. A 25 year-old male was punched in his right eye while intoxicated at a bar. He reports right eye pain and double vision when looking up but denies vision loss. On physical exam, his right eye is soft and does not exhibit proptosis, but there is superior gaze restriction.  | 2.00 (1.00)  | 1.00-3.00 |
| 2. A 21 year-old female presents with bilateral facial pain for 2 weeks. She states it started with a runny nose, cough and facial pressure 2 weeks ago. Her symptoms started to spontaneously improve after 7 days while taking Motrin. She started to develop worsening bilateral facial pain, particularly on her cheeks and forehead, associated with a fever and purulent nasal drainage. On physical exam, her frontal and maxillary sinuses are tender to palpation. There is green mucopurulent nasal drainage bilaterally, and there is edematous and erythematous mucosa throughout her nasal cavity.   | 2.00 (1.00)  | 1.00-3.00 |
| 3. A 65 year-old male with a 75 pack-year tobacco smoking history presents with a 5 cm, painful, ulcerative, and raised lateral tongue mass. He says it has continued to grow for the past 3 months and sometimes bleeds. On physical exam, the mass is friable, fixed and is isolated to the lateral ventral tongue. There are no palpable cervical lymph nodes.   | 1.00 (1.00)  | 1.00-3.00 |
| 4. A 33 year-old obese female with a 3 year history of bifrontal headache. These headaches are worse in the morning and when she lays down. She presents for evaluation of intermittent rhinorrhea for 3 months. It is clear, only comes from the left side, and does not have a particular taste. She has mild bilateral nasal congestion but denies facial pressure and change or loss of smell. She has tried intranasal steroid and anticholinergic nasal sprays which have not improved her symptoms. She was in a minor car accident 6 months ago, but the airbags were not deployed and she did not hit her head. On physical exam, there is no active nasal drainage in a head neutral or chin tucked position. Her nasal mucosa is not erythematous and there is no turbinate hypertrophy.   | 2.00 (1.00)  | 1.00-4.00 |
| 5. A 22 year-old male presents with left-sided facial pain for 2 days. He has been feeling abnormally fatigued for the past 3 months, has had 20 lbs of unintended weight loss, and has been soaking through his sheets overnight. He has also noticed some bumps around his waist while putting on his belt. The facial pain is associated with mild chunky nasal drainage. He otherwise denies facial numbness and vision changes. On physical exam, he has a low grade fever of 99F. His facial sensation, extraocular movements, and visual acuity are intact. His conjunctiva is also clear. His left nasal cavity has some mild crusting and mucinous nasal drainage. The head of the left inferior turbinate head appears dusky, as does a 1 cm area of his left hard palate. He has reports normal sensation in these areas but these areas do not bleed when pricked. There is nontender inguinal lymphadenopathy. | 4.00 (1.00)  | 2.00-5.00 |
| 6. A 40 year-old female comes in with a 12 month history of progressive right hearing loss. It is associated with a high frequency non-pulsatile tinnitus and intermittent vertigo. She denies otalgia and otorrhea. Recently, she has noticed the right-side of her face drooping over the past 3 months. On physical exam, her external auditory canals and tympanic membranes are clear. There is no middle ear effusion. Her tuning fork exam lateralizes to the left, and air conduction is greater than bone conduction bilaterally. She has a House-Brackmann grade 3 right facial paralysis.  | 2.00 (2.00)  | 2.00-5.00 |
| 7. A 34 year-old male with a history of allergic rhinitis, eustachian tube dysfunction, and recurrent episodes of right acute otitis media presents with 6 months of progressive right sided hearing loss and clear otorrhea. He denies otalgia, vertigo, and tinnitus. On physical exam, the right external auditory canal is clear. The right tympanic membrane has a superior retraction pocket with a possible tympanic membrane perforation, erosion of the malleus, and a small serous middle ear effusion. There is also a round, pearly mass in the right middle ear. His left external auditory canal and tympanic membrane are clear. His tuning fork exam lateralizes to the right, and air conduction is greater than bone conduction on the left but bone conduction is greater than air conduction on the right. His facial nerve is intact.  | 2.00 (1.00)  | 1.00-4.00 |
| 8. A 34 year-old male with no past medical history presents with left-sided neck mass for the past year. It has been slowly growing for the past year and is nontender. During the same time, he has noticed episodes of spontaneous palpitations and sweating. During a visit with his primary care doctor last month, his blood pressure was noted to be 190/110s, and he was sent to the emergency room. On physical exam, there is a firm 3 cm mobile left level 2 neck mass. There are no other palpable masses, including thyroid masses. His oral cavity and oropharynx exam are normal and his vital signs are normal.  | 2.00 (1.00)  | 1.00-4.00 |
| 9. A 25 year-old female presents with left otalgia 4 hours after a boxing match where she sustained a strike to the left side of her head. She states it has become swollen and increasingly swollen. On physical exam, the helix and antihelix of the left pinna are swollen, fluctuant and have mild ecchymosis. The external auditory canal and tympanic membrane are clear bilaterally.   | 1.00 (0)     | 1.00-2.00 |
| 10. A 29 year-old male presents with 2 day history of right hearing loss. He was working at a coffee shop and noticed right-sided hearing loss when he went home 2 days ago. He was not doing anything out of the ordinary that day. He recently came back from a trip from Colorado 1 week ago where he spent most of his time skiing. He has no other complaints and has not had hearing problems prior to 2 days ago. On physical exam, his pinna, external auditory canal, and tympanic membrane are normal. He brings an audiogram obtained yesterday that shows a 30 dB hearing loss from 250-3000 Hz and a downsloping 30-60 dB hearing loss from 4000-8000 Hz in the right ear. Air and bone conduction thresholds are equivalent. Hearing in the left ear is normal.   | 2.00 (1.00)  | 1.00-4.00 |

(continued)

Table 1. (continued)

| Prompt  | Median (IQR) | Range     |
|---|--------------|-----------|
| 11. A 37 year-old female presents with progressive shortness of breath. 5 months ago, she was in a housefire where she sustained severe inhalational injuries and was intubated for 14 days. After that incident, she has had progressive shortness of breath and has recently developed noisy, high-pitched breathing over the past 3 weeks that is worse when she exercises. On physical exam, her voice is hoarse and weak. She has mild suprasternal retractions, biphasic stridor, and requires 2 L/min of supplemental oxygen to maintain oxygen saturations above 92%.   | 2.00 (1.00)  | 1.00-4.00 |
| 12. A 39 year-old female presents with a history of recurrent bilateral cheek swelling. These episodes are accompanied by intermittent low grade fever, up to 100F. She has had these symptoms for the past 2 months. She has also had to drink more water during these episodes, particularly when she eats. She is currently undergoing a work up for shortness of breath and pulmonary hilar fullness discovered on a chest X-ray 3 months ago. Sometimes she has vision changes and facial weakness with these episodes, but these are rare. She states she is currently feeling normal. On physical, she is afebrile. Her parotid and submandibular glands are normal to palpable, and there is no palpable cervical lymphadenopathy.  | 2.00 (2.00)  | 1.00-5.00 |
| 13. A 60 year-old female with a history of Hashimoto's thyroiditis presents with a 1 year history of an enlarging neck. She has noticed the left side of her neck become more prominent, which bothers her, and has found it more difficult to swallow solids in the past 3 months. She has not had unintended weight loss. On physical exam, she has a 6 cm nodular neck mass, just left of midline, that moves with swallowing. There is no palpable lymphadenopathy, and her voice is normal.  | 2.00 (1.00)  | 1.00-4.00 |
| 14. A 4 year-old male presents with double vision for 1 day. His older brother developed a cough and nasal congestion 2 weeks ago which improved without intervention after 5 days. One week ago, the patient developed similar symptoms and nasal congestion. He woke up with a headache and forehead swelling this morning. On physical exam, he is febrile to 101.2F. His forehead is tense, swollen, tender and mildly fluctuant. His right eye has mild conjunctival injection, has mild proptosis. There is mild right-sided abduction with lateral gaze, but visual acuity is intact. There is mild clear bilateral mucinous rhinorrhea. There is no neck stiffness.   | 2.00 (0)     | 1.00-4.00 |
| 15. A 13 year-old male presents with a 7 day history of sore throat. It is associated with odynophagia, poor oral intake, and a fever up to 101.4F. He denies cough. This has happened multiple times in the past, all of which have resolved with antibiotics. This time, however, his symptoms have not improved despite starting a course of amoxicillin/clavulanic acid 4 days ago. On physical exam, he is febrile, tachycardic, and has dry mucous membranes. His tonsils are enlarged and erythematous. There is significant leftward uvular deviation, and his voice is muffled. There is bilateral level 2 tender cervical adenopathy, right more than the left.   | 1.00 (1.00)  | 1.00-3.00 |
| 16. A 42 year-old female presents with a 1 year history of a hoarse and strained voice. She reports a 2 week history of cough and sore throat during the winter last year after which her voice became hoarse and strained. It waxes and wanes but is rarely normal. She also reports clearing her throat more frequently. She is a teacher and finds it difficult to teach because of the strain in her voice. Sometimes she also has difficulty yelling at her kids from across the house. On physical exam, her voice is hoarse and mildly strained. She has equal difficulty saying either sentence: "A dog dug a new bone" and "Harry is happy because he has a new horse." There are no breaks in her voice, and there is no stridor.   | 3.00 (2.00)  | 1.00-4.00 |
| 17. A 55 year-old male underwent a coronary artery bypass graft 5 days ago and presents with a hoarse voice. He has had difficulty with eating, particularly with thin liquids. His voice has not improved since surgery. On physical exam, his voice is hoarse and breathy. When he takes a drink of water, he has overt signs of aspiration. He finds it easier to drink water when he turns his head to the left and tucking his chin.   | 1.00 (1.00)  | 1.00-4.00 |
| 18. A 3 week-old male presents with a 1 week history of left neck mass. Prenatal history is significant for gestational diabetes. The patient was born post-term via forceps-assisted vaginal delivery. The neck mass has been firm and slowly progressive in size. The patient has otherwise been doing well and has been feeding and gaining weight appropriately. On physical exam, the patient's head is turned slightly to the right. There is a 4 ×3 cm dense mass on the left neck that moves with head turning.   | 2.00 (2.00)  | 1.00-4.00 |
| 19. A 10 year-old female presents with a painful neck mass over the past 3 days. She states this has happened once before and resolved after a course of oral antibiotics. It is associated with a fever, swelling over the affected area, and pain with swallowing. On physical exam, there is a 3 cm midline tender neck mass that moves when swallowing. It is firm and partially fixed to the overlying skin which has mild erythema. There is bilateral tender cervical lymphadenopathy.   | 2.00 (1.00)  | 1.00-4.00 |
| 20. A 41 year-old female presents with 1 week history of left otalgia. It has been progressively worse since it started and has spread to her left eye today. She also endorses double vision and hearing loss but denies tinnitus, otorrhea and vertigo. She has insulin-dependent diabetes mellitus. On physical exam, she appears fatigued. She has a mild left eye abduction deficit. She has normal pinna and external auditory canals bilaterally. There is a milky left middle ear effusion, but her tympanic membrane appears relatively normal. She has a temperature of 101.4F. His tuning fork exam lateralizes to the left, and air conduction is greater than bone conduction on the right but bone conduction is greater than air conduction on the left. Her facial nerve is intact. | 2.00 (2.00)  | 1.00-4.00 |
| Overall   | 2.00 (1.00)  | 1.00-5.00 |

Abbreviation: IQR, interquartile range.

**Table 2.** ChatGPT-Provided Differential Diagnoses and Physician Agreement

| Prompt | Differential diagnoses  | Median (IQR) | Range     |
|--------|---|--------------|-----------|
| 1      | 1. Orbital blowout fracture<br>2. Orbital contusion<br>3. Superior rectus/levator palpebrae superioris muscle injury<br>4. Cranial nerve palsy (specifically, oculomotor nerve palsy)<br>5. Periorbital hematoma  | 5.00 (1.00)  | 4.00-5.00 |
| 2      | 1. Acute bacterial sinusitis<br>2. Acute viral rhinosinusitis<br>3. Allergic rhinitis<br>4. Fungal sinusitis<br>5. Dental infection with sinus involvement  | 5.00 (1.00)  | 4.00-5.00 |
| 3      | 1. Squamous cell carcinoma (SCC)<br>2. Verrucous carcinoma: A rare, well-differentiated variant of SCC<br>3. Oral pyogenic granuloma<br>4. Major aphthous ulcer (Sutton's disease)<br>5. Oral lichen planus   | 5.00 (0)     | 4.00-5.00 |
| 4      | 1. Spontaneous cerebrospinal fluid leak<br>2. Idiopathic intracranial hypertension (pseudotumor cerebri)<br>3. Chronic tension-type headache<br>4. Migraine without aura<br>5. Sinus headache (noninfectious)   | 5.00 (1.00)  | 2.00-5.00 |
| 5      | 1. Granulomatosis with polyangiitis (formerly known as Wegener's granulomatosis)<br>2. Extranodal natural killer/T-cell lymphoma, nasal type<br>3. Invasive fungal sinusitis (eg, mucormycosis or aspergillosis)<br>4. Tuberculosis infection with sinonasal involvement<br>5. Syphilis with sinonasal manifestations | 4.00 (3.00)  | 1.00-5.00 |
| 6      | 1. Vestibular schwannoma (acoustic neuroma)<br>2. Ramsay Hunt syndrome (herpes zoster oticus)<br>3. Ménière's disease<br>4. Temporal bone fracture<br>5. Autoimmune inner ear disease   | 5.00 (1.00)  | 4.00-5.00 |
| 7      | 1. Chronic otitis media with cholesteatoma<br>2. Chronic serous otitis media<br>3. Adhesive otitis media<br>4. Tympanosclerosis<br>5. Congenital cholesteatoma  | 5.00 (0)     | 5.00-5.00 |
| 8      | 1. Pheochromocytoma with a cervical paraganglioma<br>2. Cervical lymphadenopathy due to metastatic malignancy<br>3. Schwannoma of the cervical sympathetic chain<br>4. Branchial cleft cyst<br>5. Tuberculous lymphadenitis   | 5.00 (1.00)  | 4.00-5.00 |
| 9      | 1. Auricular hematoma<br>2. Perichondritis<br>3. Traumatic ear laceration<br>4. Acute otitis externa<br>5. Temporomandibular joint dysfunction  | 5.00 (0)     | 5.00-5.00 |
| 10     | 1. Sudden sensorineural hearing loss<br>2. Acoustic trauma<br>3. Barotrauma<br>4. Eustachian tube dysfunction<br>5. Meniere's disease (early stage)   | 5.00 (0)     | 5.00-5.00 |
| 11     | 1. Subglottic stenosis<br>2. Tracheomalacia<br>3. Vocal cord dysfunction<br>4. Bronchiolitis obliterans<br>5. Reactive airway dysfunction syndrome  | 4.00 (1.00)  | 3.00-5.00 |
| 12     | 1. Sjögren's syndrome<br>2. Sarcoidosis<br>3. Mikulicz's disease (IgG4-related disease)<br>4. Recurrent parotitis<br>5. Lymphoma  | 4.00 (1.00)  | 1.00-5.00 |
| 13     | 1. Goiter associated with Hashimoto's thyroiditis<br>2. Thyroid nodule(s)   | 5.00 (0)     | 4.00-5.00 |

(continued)

**Table 2.** (continued)

| Prompt  | Differential diagnoses  | Median (IQR) | Range     |
|---------|---|--------------|-----------|
| 14      | <ol style="list-style-type: none"> <li>3. Thyroid cyst</li> <li>4. Thyroid cancer</li> <li>5. Thyroid lymphoma</li> <li>1. Orbital cellulitis</li> <li>2. Preseptal cellulitis (periorbital cellulitis)</li> <li>3. Frontal sinusitis with subperiosteal abscess</li> <li>4. Ethmoid sinusitis with orbital extension</li> <li>5. Allergic rhinitis with a secondary bacterial infection</li> </ol> | 4.00 (1.00)  | 3.00-5.00 |
| 15      | <ol style="list-style-type: none"> <li>1. Peritonsillar abscess</li> <li>2. Infectious mononucleosis (Epstein-Barr virus infection)</li> <li>3. Bacterial tonsillitis</li> <li>4. Lemierre's syndrome</li> <li>5. Viral pharyngitis</li> </ol>  | 5.00 (0)     | 4.00-5.00 |
| 16      | <ol style="list-style-type: none"> <li>1. Vocal fold nodules</li> <li>2. Muscle tension dysphonia</li> <li>3. Chronic laryngitis</li> <li>4. Reinke's edema (polypoid corditis)</li> <li>5. Vocal fold paralysis or paresis</li> </ol>  | 4.00 (1.00)  | 2.00-5.00 |
| 17      | <ol style="list-style-type: none"> <li>1. Recurrent laryngeal nerve injury or palsy</li> <li>2. Vocal cord paralysis</li> <li>3. Aspiration pneumonia</li> <li>4. Laryngeal edema</li> <li>5. Laryngopharyngeal reflux</li> </ol>   | 5.00 (1.00)  | 2.00-5.00 |
| 18      | <ol style="list-style-type: none"> <li>1. Congenital muscular torticollis</li> <li>2. Sternocleidomastoid tumor (fibromatosis colli)</li> <li>3. Lymphatic malformation (cystic hygroma)</li> <li>4. Branchial cleft cyst</li> <li>5. Hemangioma</li> </ol>   | 5.00 (1.00)  | 4.00-5.00 |
| 19      | <ol style="list-style-type: none"> <li>1. Acute suppurative thyroiditis</li> <li>2. Subacute (De Quervain's) thyroiditis</li> <li>3. Lymphadenitis</li> <li>4. Thyroglossal duct cyst infection</li> <li>5. Abscess</li> </ol>  | 3.00 (2.00)  | 1.00-5.00 |
| 20      | <ol style="list-style-type: none"> <li>1. Acute otitis media with complications</li> <li>2. Mastoiditis</li> <li>3. Gradenigo's syndrome</li> <li>4. Cholesteatoma with complications</li> <li>5. Temporal bone osteomyelitis</li> </ol>  | 4.00 (1.00)  | 2.00-5.00 |
| Overall |   | 5.00 (1.00)  | 1.00-5.00 |

Abbreviations: ChatGPT, chat generative pretrained transformer; IgG4, immunoglobulin G4; IQR, interquartile range.

lowest median score was 3.00 for prompt #19. The highest median score was 5.00. The lowest IQR was 0 for prompts #6, #7, and #9. The highest IQR was 4.00 for prompt #5. Prompts #12 and #19 had an IQR of 3.00. Agreement with treatment was similarly high. Overall, the median agreement with treatment was 5.00 with an IQR of 1.00.

To elucidate the potential limitations of ChatGPT in diagnosis and treatment, we performed univariate ordinal regression between prompt difficulty, diagnosis score, and treatment score. Strongly agree (Likert score of 5) was used as the reference level. Odds ratios with 95% confidence intervals and *p* value are presented in **Table 4**. The prompt difficulty was not a significant predictor of the diagnostic score as the odds of having a higher diagnostic score did not vary with increasing Likert score for prompt difficulty. Similarly, there was no association between prompt difficulty and treatment score. The odds of having a higher diagnostic score did not vary with an increasing Likert

score for prompt difficulty. The diagnostic score did appear to be significantly related to the treatment score. Having a lower diagnostic score had lower odds of having a higher treatment score. More simply put, a lower diagnostic score was more likely associated with lower treatment scores. Odds ratios for diagnostic scores of 1, 2, 3, and 4 were 0.381, 0.330, 0.289, and 0.301 respectively. Diagnostic scores of 2 and 4 reach statistical significance ( $p < .001$  and  $p = .050$ , respectively) while a score of 3 trended toward significance ( $p = .052$ ). Having a Likert score of 1 likely did not reach significance ( $p = .307$ ) due to the small sample size, as evidenced by the wide confidence interval for this level. Overall, this suggests that if there is no strong agreement with the diagnosis, agreement with treatment will also tend to be lower. Again, we emphasize that the treatment score is based on the top differential diagnosis selected by ChatGPT and not necessarily the diagnosis most otolaryngologists would select.

**Table 3.** ChatGPT-Provided Treatment and Physician Agreement

| Prompt | Treatment  | Median (IQR) | Range     |
|--------|--|--------------|-----------|
| 1      | Based on the given scenario, the most likely diagnosis is an orbital blowout fracture. Here's a suggested treatment plan:<br><i>Consults</i> —ophthalmology, OMFS<br><i>Radiographic studies</i> —CT scan of orbits and facial bones<br><i>Medical treatment</i> —pain control, tetanus prophylaxis if immunization is not up to date, antibiotics may be prescribed<br><i>Surgical treatment</i> —depending on the extent of the fracture, muscle entrapment, and any functional or cosmetic concerns, surgical repair may be necessary within 1-2 weeks of injury. Surgery typically involves reconstruction of the orbital floor or wall with implants or grafts, as well as the release of any entrapped muscles<br><i>Follow-up</i> —regular follow-up appointments with ophthalmology and maxillofacial surgery  | 4.00 (2.00)  | 1.00-5.00 |
| 2      | The most likely diagnosis, in this case, is acute bacterial sinusitis. Here's a suggested treatment plan:<br><i>Consults</i> —ENT if symptoms persist despite medical therapy<br><i>Radiographic studies</i> —CT sinus if there is no significant improvement within 48-72 hours of starting antibiotics<br><i>Medical treatment</i> —antibiotics (amoxicillin-clavulanate for 5-10 days in adults; doxycycline or a respiratory fluoroquinolone if the patient has a penicillin allergy)<br><i>Symptomatic relief</i> —pain relief, oral or topical nasal decongestant no longer than 3 days to avoid rebound congestion, nasal saline irrigation<br><i>Follow up</i> —educate the patient on the importance of completing the full course of antibiotics, follow up in 10-14 days  | 5.00 (1.00)  | 4.00-5.00 |
| 3      | Given the patient's history and presentation, the most likely diagnosis is squamous cell carcinoma (SCC) of the tongue. Here's a suggested treatment plan:<br><i>Consults</i> —ENT and OMFS<br><i>Radiographic studies</i> —contrast-enhanced CT or MRI of the head and neck, CXR, or CT chest may be considered to rule out distant metastasis<br><i>Other studies</i> —biopsy of the lesion should be performed, along with imaging can be used to stage the tumor for treatment<br><i>Treatment</i> —may include a combination of surgery, radiation therapy, and chemotherapy <ul style="list-style-type: none"> <li>• Surgery: resection with margins for early-stage SCC of the tongue is the primary treatment; neck dissection if there is evidence of regional lymph node involvement or with advanced-stage disease</li> <li>• Radiation: postoperative radiation for high-risk or advanced stage; definitive radiation as an alternative for patients who are not surgical candidates</li> <li>• Chemotherapy: neoadjuvant or adjuvant chemotherapy for advanced stage or high risk of recurrence</li> </ul> <i>Follow up</i> —regular follow-up with ENT, OMFS, and oncologist | 5.00 (1.00)  | 4.00-5.00 |
| 4      | The most likely diagnosis, in this case, is a spontaneous cerebrospinal fluid leak. Here's a suggested treatment plan:<br><i>Consults</i> —ENT and/or NSGY if CSF leak does not resolve with conservative measures<br><i>Radiographic studies</i> —high-resolution CT of sinuses, MRI with intrathecal gadolinium<br><i>Other studies</i> —LP to measure opening pressure and obtain CSF for analysis<br><i>Medical treatment</i> —bed rest with head elevation, analgesics, and prophylactic antibiotics to minimize the risk of meningitis if necessary<br><i>Surgical treatment</i> —endoscopic surgical repair if the leak does not resolve with conservative measures, lumbar drain or shunt if the leak is not identifiable or if multiple leaks are present<br><i>Follow up</i> —weight loss/obesity management, monitor recurrence of symptoms which if persist or worsen then consider repeat imaging and re-evaluating treatment plan  | 5.00 (2.00)  | 2.00-5.00 |
| 5      | The most likely diagnosis, in this case, is granulomatosis with polyangiitis. Here's a suggested treatment plan:<br><i>Consults</i> —rheumatology, ENT<br><i>Radiographic studies</i> — CXR, CT Chest<br><i>Labs</i> —CBC, CMP, ESR, CRP, ANCA, UA<br><i>Biopsy</i> —nasal or sinus biopsy<br><i>Medical treatment</i> —induction high-dose corticosteroids with a DMARD; maintenance with less toxic DMARD<br><i>Follow up</i> —rheumatology, monitor laboratory markers, assess for end organ complications  | 4.00 (4.00)  | 1.00-5.00 |
| 6      | Based on the patient's history and presentation, the most likely diagnosis is a vestibular schwannoma (acoustic neuroma). Here's a suggested treatment plan:<br><i>Consults</i> —neuro-otologist, NSGY, audiologist<br><i>Radiographic studies</i> —MRI head w/contrast  | 5.00 (0)     | 1.00-5.00 |

(continued)

Table 3. (continued)

| Prompt | Treatment   | Median (IQR) | Range     |
|--------|---|--------------|-----------|
| 7      | <p><i>Treatment options</i>—observation (small tumors with mild symptoms), stereotactic radiosurgery (serviceable hearing and tumor &lt; 3 cm), surgical resection (large tumors with significant symptoms or when hearing preservation is not possible)</p> <p><i>Follow-up</i>—facial nerve rehabilitation, neuro-otology, NSGY</p> <p>The most likely diagnosis is Chronic Otitis Media with Cholesteatoma. Here is a treatment plan for this condition:</p> <p><i>Consults</i>—ENT, audiologist</p> <p><i>Radiographic studies</i>—high-resolution CT of the temporal bones</p> <p><i>Medical treatment</i>—treat concurrent infections with antibiotics and/or corticosteroid ear drops</p> <p><i>Surgical treatment</i>—excision which may include canal-wall-up or canal-wall-down mastoidectomy with tympanoplasty</p> <p><i>Follow up</i>—monitor for postoperative complications, regular follow up with otoscopic exams and audiograms, manage eustachian tube dysfunction and allergic rhinitis</p> | 5.00 (0)     | 1.00-5.00 |
| 8      | <p>The most likely diagnosis, in this case, is pheochromocytoma with a cervical paraganglioma. Here's a treatment plan for this patient:</p> <p><i>Consults</i>—endocrinology, ENT</p> <p><i>Radiographic studies</i>—CT or MRI of the neck with contrast, abdominal CT or MRI</p> <p><i>Labs</i>—plasma or 24-hour urine metanephrines and catecholamines, CBC, renal function tests, and electrolytes</p> <p><i>Medical treatment</i>—alpha-blockers and subsequently beta-blockers</p> <p><i>Surgical treatment</i>—laparoscopic or open adrenalectomy and neck dissection</p> <p><i>Follow-up</i>—monitor patient's blood pressure and catecholamine levels, adjust antihypertensive medications, imaging and laboratory testing for recurrence</p>   | 4.00 (1.00)  | 2.00-5.00 |
| 9      | <p>For the most likely diagnosis of auricular hematoma, the treatment plan would include the following steps:</p> <p><i>Consults</i>—ENT</p> <p><i>Treatment</i>—cold compress, aspiration or incision and drainage, pressure dressing or bolster after drainage, antibiotic prophylaxis, pain management</p> <p><i>Follow-up</i>—ENT within 1 week to monitor healing and removal of pressure dressing, avoid contact sports or other potentially traumatic activities until healing, and protective headgear in the future for prevention</p>   | 5.00 (0)     | 4.00-5.00 |
| 10     | <p>The most likely diagnosis, in this case, is sudden sensorineural hearing loss. Here's a treatment plan for this patient:</p> <p><i>Consults</i>—ENT</p> <p><i>Radiographic studies</i>—MRI brain and IAC with contrast</p> <p><i>Labs</i>—CBC, renal function tests, electrolytes, autoimmune labs, and infectious etiologies (eg, Lyme disease, syphilitic) if indicated</p> <p><i>Treatment</i>—oral corticosteroids preferably within 72 hours of onset with a 10-14 day taper, intratympanic corticosteroid injections if oral treatment is contraindicated or ineffective, supportive care including hearing aids or assistive listening devices as necessary</p> <p><i>Follow up</i>—follow up audiograms and adjust treatments as necessary</p>   | 4.00 (1.00)  | 3.00-5.00 |
| 11     | <p>The most likely diagnosis in this case is subglottic stenosis. The treatment plan for subglottic stenosis includes the following:</p> <p><i>Consults</i>—ENT, pulmonology</p> <p><i>Radiographic studies</i>—CT neck and chest</p> <p><i>Other studies</i>—flexible laryngoscopy, pulmonary function tests</p> <p><i>Medical treatment</i>—inhaled corticosteroids and bronchodilators</p> <p><i>Surgical options</i>—endoscopic dilation, endoscopic lysis of adhesions, laser surgery, open surgical repair</p> <p><i>Follow-up</i>—speech therapy, pulmonary rehabilitation, regular follow-up with ENT and pulmonology to monitor for recurrence</p>   | 5.00 (1.00)  | 3.00-5.00 |
| 12     | <p>The most likely diagnosis, in this case, is Sjögren's syndrome. Here's a treatment plan for this patient:</p> <p><i>Consults</i>—rheumatology, ophthalmology</p> <p><i>Radiographic studies</i>—CXR or CT test, parotid gland ultrasound</p> <p><i>Labs</i>—CBC, renal function tests, electrolytes, liver function tests, ANA, anti-Ro/SSA, anti-La/SSB, ESR, CRP, immunoglobulins</p> <p><i>Other studies</i>—Schirmer's test, minor salivary gland biopsy if clinically indicated</p> <p><i>Treatment</i>—artificial tears and lubricating eye ointments, saliva substitutes, sugar-free gum or lozenges, hydroxychloroquine, NSAIDs, pilocarpine, or cevimeline</p> <p><i>Follow-up</i>—close monitoring for potential complications, regular dental care, and good oral hygiene</p>   | 4.00 (3.00)  | 1.00-5.00 |
| 13     | <p>The most likely diagnosis, in this case, is a goiter associated with Hashimoto's thyroiditis. The treatment plan for goiter in the context of Hashimoto's thyroiditis includes the following:</p> <p><i>Consults</i>—endocrinology</p>   | 5.00 (1.00)  | 2.00-5.00 |

(continued)



**Table 3.** (continued)

| Prompt | Treatment   | Median (IQR) | Range     |
|--------|---|--------------|-----------|
| 14     | <p><i>Radiographic studies</i>—thyroid ultrasound</p> <p><i>Labs</i>—thyroid function tests</p> <p><i>Other studies</i>—fine-needle aspiration biopsy</p> <p><i>Medical treatment</i>—levothyroxine for hypothyroidism and optionally for goiter size reduction</p> <p><i>Surgical treatment</i>—thyroidectomy (partial or total) if there are compressive symptoms, malignancy is confirmed or strongly suspected, goiter is cosmetically bothersome</p> <p><i>Follow-up</i>—endocrinology, periodic thyroid ultrasound, and thyroid function tests</p> <p>The most likely diagnosis, in this case, is orbital cellulitis. Here's a treatment plan for this patient:</p> <p><i>Consults</i>—ENT, ophthalmology</p> <p><i>Radiographic studies</i>—CT or orbit and sinuses with contrast</p> <p><i>Labs</i>—CBC, blood cultures, culture from sinus or conjunctival swab if possible</p> <p><i>Medical treatment</i>—hospital admission, broad-spectrum intravenous antibiotics to cover staphylococcus aureus and streptococcus species, adjust antibiotics when culture results are available, analgesics and antipyretics, nasal decongestants</p> <p><i>Surgical treatment</i>—consider drainage of abscess or endoscopic sinus surgery in cases of severe infection, abscess, or lack of clinical improvement with conservative management</p> <p><i>Follow-up</i>—monitor for complications including vision loss, cavernous sinus thrombosis, and intracranial spread of infection</p> | 5.00 (1.00)  | 1.00-5.00 |
| 15     | <p>The most likely diagnosis, in this case, is a peritonsillar abscess. Here is a suggested treatment plan:</p> <p><i>Consults</i>—ENT</p> <p><i>Radiographic studies</i>—CT neck with contrast</p> <p><i>Labs</i>—CBC, blood cultures</p> <p><i>Medical treatment</i>—intravenous hydration, intravenous broad-spectrum antibiotics to cover group A <i>Streptococcus</i> and anaerobes, analgesics and antipyretics, adjust antibiotics based on culture and clinical response</p> <p><i>Surgical treatment</i>—needle aspiration, incision, and drainage, or tonsillectomy</p> <p><i>Follow-up</i>—monitor patient's clinical response to treatment, follow up with ENT for recovery, and discuss elective tonsillectomy if recurrent tonsillitis is an issue</p>  | 5.00 (1.00)  | 2.00-5.00 |
| 16     | <p>The most likely diagnosis, in this case, is vocal fold nodules. Here's a treatment plan for this patient:</p> <p><i>Consults</i>—ENT, speech-language pathologist</p> <p><i>Other studies</i>—laryngoscopy, videostroboscopy</p> <p><i>Medical treatment</i>—voice therapy, vocal hygiene, voice rest, vocal warm-ups and cool-downs</p> <p><i>Surgical treatment</i>—microlaryngoscopy with excision or laser ablation if conservative management fails to yield improvement or in severe cases</p> <p><i>Follow-up</i>—monitor for improvement with voice therapy and consider re-evaluating diagnosis and exploring alternative treatments if there is no significant improvement in a few months</p>   | 4.00 (1.00)  | 1.00-5.00 |
| 17     | <p>Based on the most likely diagnosis of recurrent laryngeal nerve injury or palsy following coronary artery bypass graft surgery, I can suggest a general approach for management:</p> <p><i>Consults</i>—ENT</p> <p><i>Radiographic studies</i>—CT or MRI of neck and chest</p> <p><i>Other studies</i>—laryngoscopy</p> <p><i>Medical treatment</i>—voice and swallow therapy</p> <p><i>Surgical treatment</i>—injection laryngoplasty, medialization thyroplasty, or arytenoid adduction</p> <p><i>Follow-up</i>—aspiration precautions, pneumonia prevention, follow-up with ENT and speech-language pathology</p>   | 5.00 (1.00)  | 4.00-5.00 |
| 18     | <p>The most likely diagnosis, in this case, is congenital muscular torticollis. Here's a treatment plan for this patient:</p> <p><i>Consults</i>—pediatrician, physical therapist</p> <p><i>Radiographic studies</i>—neck ultrasound</p> <p><i>Medical treatment</i>—physical therapy, parent education for strengthening and stretching exercises, head positioning techniques</p> <p><i>Surgical treatment</i>—surgical release of the sternocleidomastoid muscle in severe cases or cases where physical therapy fails to yield improvement</p> <p><i>Follow-up</i>—pediatrician, physical therapist</p>   | 5.00 (1.00)  | 2.00-5.00 |
| 19     | <p>Based on the most likely diagnosis of acute suppurative thyroiditis, I can suggest a general approach for management:</p> <p><i>Consults</i>—endocrinology, ENT</p> <p><i>Radiographic studies</i>—neck ultrasound, CT neck with contrast</p> <p><i>Labs</i>—CBC, thyroid function tests</p> <p><i>Other studies</i>—fine-needle aspiration</p> <p><i>Medical treatment</i>—intravenous broad-spectrum antibiotics until culture results are available after which adjust the antibiotics for a 2-3 week course, pain and fever management</p>   | 3.00 (3.00)  | 1.00-5.00 |

(continued)

**Table 3.** (continued)

| Prompt  | Treatment   | Median (IQR) | Range     |
|---------|---|--------------|-----------|
| 20      | <p><i>Surgical treatment</i>—drainage of abscess or in rare cases thyroidectomy (partial or total)</p> <p><i>Follow-up</i>—endocrinology, ENT</p> <p>The most likely diagnosis, in this case, is acute otitis media with complications. Here's a treatment plan for this patient:</p> <p><i>Consults</i>—ENT, audiology</p> <p><i>Imaging studies</i>—CT of the temporal bone</p> <p><i>Other studies</i>—CBC, CRP, ESR</p> <p><i>Medical treatment</i>—broad-spectrum antibiotics, analgesics, antipyretics</p> <p><i>Surgical treatment</i>—myringotomy with or without tympanostomy tube placement or a mastoidectomy if complications such as mastoiditis or cholesteatoma are identified</p> <p><i>Follow-up</i>—close monitoring for improvement, complications, and need for surgery</p> | 4.00 (2.00)  | 2.00-5.00 |
| Overall |   | 5.00 (1.00)  | 1.00-5.00 |

Abbreviations: ANA, antinuclear antibody; ANCA, antineutrophil cytoplasmic antibody; CBC, complete blood count; ChatGPT, chat generative pretrained transformer; CMP, comprehensive metabolic panel; CRP, C-reactive protein; CSF, cerebrospinal fluid; CT, computed tomography; CXR, chest X-ray; DMARD, disease-modifying antirheumatic drug; ENT, otolaryngology; ESR, erythrocyte sedimentation rate; IAC, internal auditory canal; IQR, interquartile range; LP, lumbar puncture; MRI, magnetic resonance imaging; NSAIDs, nonsteroidal anti-inflammatory drugs; NSGY, neurosurgery; OMFS, oral maxillofacial surgery; SCC, squamous cell carcinoma; UA, urinalysis.

**Table 4.** Ordinal Regression Between Vignette Difficulty, Diagnostic Scores, and Treatment Scores

| Likert score | Dependent: Diagnosis score<br>Independent: Prompt difficulty |                |         | Dependent: Treatment score<br>Independent: Prompt difficulty |                |         | Dependent: Treatment score<br>Independent: Diagnosis score |                |         |
|--------------|--|----------------|---------|--|----------------|---------|--|----------------|---------|
|              | OR   | 95% CI         | p Value | OR   | 95% CI         | p Value | OR   | 95% CI         | p Value |
| 1            | 2.01   | (0.275, 14.30) | .495    | 0.482  | (0.046, 5.00)  | .542    | 0.381  | (0.060, 2.44)  | .307    |
| 2            | 0.853  | (0.121, 5.99)  | .874    | 0.273  | (0.027, 2.80)  | .274    | 0.330  | (0.110, 1.00)  | .050    |
| 3            | 0.865  | (0.106, 7.03)  | .892    | 0.868  | (0.071, 10.70) | .912    | 0.289  | (0.082, 1.01)  | .052    |
| 4            | 0.660  | (0.087, 5.00)  | .162    | 0.232  | (0.021, 2.53)  | .231    | 0.301  | (0.170, 0.533) | <.001   |
| 5            |  | -              | -       |  | -              | -       |  | -              | -       |

Abbreviations: CI, confidence interval; OR, odds ratio.

## Discussion

As a result of increasing data availability and accessibility to high-performance AI technologies, AI is rapidly integrating across many industries, including health care. One of the most promising applications of AI in health care is diagnostics. The incorporation of publicly available and culturally ubiquitous AI systems, such as ChatGPT, will be inevitable in medicine. Whether the user is laymen or a clinician, it is imperative that physicians, particularly those in specialized care, understand its potential and limitations.

Several authors have begun to examine how ChatGPT might become involved in medicine. Some emphasize its role in supporting clinical tasks by helping create medical reports, patient-specific forms and handouts, medical publishing, administration, and provider education.<sup>6</sup> In fact, we used ChatGPT to assist in writing the introduction. While it required significant human editing, ChatGPT provided a strong framework to start with (Supplemental Figure S2, available online).

To date, no study has evaluated ChatGPT's performance in clinical otolaryngology, but several studies have

found it to perform well in adjacent fields. Despite not being trained on a specific data set, ChatGPT performed at the level of a first-year resident in plastic surgery on the in-service training exam.<sup>7,8</sup> In neurosurgery, ChatGPT performed worse than the average user on Self-Assessment Neurosurgery questions but better than residents in some topics.<sup>9</sup> Clearly, there is already some rudimentary capacity in providing specialty care. This is consistent with the findings of our study which demonstrated a high rating from physicians for easy to moderately difficult clinical vignettes.

As evidenced in our study, when presented with clinical vignettes of well-known clinical issues using medical jargon and curated relevant history, physical exam, and radiographic and laboratory findings, ChatGPT provides a very accurate differential diagnosis and reasonable treatment plans. This is likely due to the similarity of the vignettes provided in this study to the type of writing found in textbooks, scientific papers, and other data sources the AI model was trained on. This also explains why we failed to identify an association between prompt difficulty and diagnosis and treatment scores. ChatGPT

performed better than the authors of this study expected for the difficulty of the prompts provided. If more difficult prompts were provided, we suggest that variations in diagnosis and treatment scores would become more apparent. Also, the apparent discordance between human-assigned difficulty and the performance of the AI system is not surprising when we understand that the relative strengths and weaknesses between human intelligence and AI are different. Humans will tend to assign more difficulty to rare and esoteric conditions and treatments because the amount of data we can store is limited and we will prioritize more common and useful information. Computers do not have these limitations and can have access to more information than any person can have in their mind and obscurity is a trivial barrier for them. Nuanced diagnosis and treatment of more common clinical problems presented by patients in nonmedical jargon will likely be perceived as an easy problem by human evaluators but be difficult for AI systems.

Furthermore, it remains unclear how well it would perform if provided with real-world, open-ended free-form histories, complete exams, and uncurated data which would often contain irrelevant, extraneous, and contradictory information. Unfiltered clinical information may prove too challenging for current LLMs to narrow into meaningful diagnoses, particularly in specialty fields. Several studies have demonstrated its impressive but inferior performance to human physicians in a variety of clinical vignettes.<sup>5,10</sup> Future studies may compare how ChatGPT performs with open-ended inputs provided by otolaryngology patients (eg, “Why do I have nasal obstruction?”). Finally, some studies note that ChatGPT may generate different responses for the same prompt entered in multiplicate which undermines its clinical reliability.<sup>11</sup>

The role of ChatGPT from a patient's perspective in certain situations has been explored. ChatGPT is able to answer a variety of patient questions about colonoscopy in a digestible and generally satisfactory manner.<sup>12</sup> It is also able to provide high-quality answers to patient questions in pre- and postoperative care in oral and maxillofacial surgery, though the author suggests it be used in conjunction with surgeon experience.<sup>13</sup> Within otolaryngology, ChatGPT provides patients with procedure-specific instructions equivalent to institutional standards.<sup>14</sup> However, ChatGPT instructions were equal to and inferior to Google's in terms of understandability and actionability, respectively.<sup>14</sup> Patients are likely to find ChatGPT to be a useful and accessible resource. And while the quality of its answers for patients is surprisingly high, it cannot be used as a direct substitute for a physician's counseling. As alluded to above and by other authors, it may serve as a useful adjunct or starting point for patient handouts. Patients using ChatGPT should be counseled on these limitations. Furthermore, its use in guiding general practice providers in specialty care has not been assessed but remains an interesting concept.

Some limitations in ChatGPT may be inherent to its underlying data modeling. For instance, it lacks knowledge of events occurring after September 2021 and does not learn from its own experience.<sup>2</sup> Additionally as its creators note ChatGPT may produce artificial hallucinations (ie, “produce content that is nonsensical or untruthful in relation to certain sources”) which can be difficult to discern when juxtaposed with ever-increasing believability.<sup>2</sup> Within biomedical research, this phenomenon is apparent when ChatGPT returns fake references when it is asked to cite its sources.<sup>15</sup> How this impacts its capacity in medical decision-making is not readily apparent.

Finally, from an ethics perspective, it should be reiterated that AI systems like ChatGPT in medicine are niche, and generalizability is still in its infancy. Moreover, unknown bias in ChatGPT can have significant and unintended consequences for patient outcomes. Inherent to AI is the training set they are based on which can in turn perpetuate disparities and biases in race, sex, and culture.<sup>16-18</sup> Thus, the need for large amounts of high-quality data that accurately and equitably represent a diverse patient population and understanding intrinsic bias when applying ChatGPT is imperative. Physicians that consider using ChatGPT in clinical practice must understand this aspect. And the burgeoning field of bioethics will also have to contemplate the risks and benefits of the use of AI in health care.<sup>19</sup>

## Conclusion

Overall, ChatGPT is a promising technology for both patients and physicians. Our study demonstrated that within otolaryngology, given highly curated vignettes, it provided differential diagnoses and treatment plans for easy to moderately difficult clinical scenarios that physicians highly agreed with. However, both physicians and patients need to be aware of the challenges and limitations of this LLM before implementing it in clinical practice. Physicians, medical societies, and patients, among other important stakeholders, should be involved in the development and application of these infant technologies in medicine.

## Author Contributions

**Roy W. Qu**, study design, data acquisition, statistical analysis, data interpretation, and manuscript preparation; **Uneeb Qureshi**, data acquisition, study design, data interpretation, and manuscript preparation; **Garrett Petersen**, data acquisition, study design, data interpretation, and manuscript preparation; **Steve C. Lee**, study design, data interpretation, and manuscript preparation. All authors approve the manuscript and agree to be accountable for all aspects of the work presented herein.

## Disclosures


**Competing interests:** The authors have no conflicts of interest to declare.


**Funding source:** None.


## Supplemental Material


Additional supporting information is available in the online version of the article.

### ORCID iD

Roy W. Qu  <http://orcid.org/0000-0002-4689-819X>

Uneeb Qureshi  <http://orcid.org/0000-0003-3185-0449>

Garrett Petersen  <http://orcid.org/0009-0004-5501-7114>

Steve C. Lee  <http://orcid.org/0000-0002-4460-4303>

### References

- Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. *Front Comput Intell Syst.* 2023;2(2):81-83. doi:10.54097/fcis.v2i2.4465
- OpenAI. *GPT-4 technical report.* OpenAI; 2023.
- Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9:e45312. doi:10.2196/45312
- Feng Y, Wang Y, Zeng C, Mao H. Artificial intelligence and machine learning in chronic airway diseases: focus on asthma and chronic obstructive pulmonary disease. *Int J Med Sci.* 2021;18(13):2871-2889. doi:10.7150/ijms.58191
- Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 Chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health.* 2023;20(4):3378. doi:10.3390/ijerph20043378
- Biswas S. ChatGPT and the future of medical writing. *Radiology.* 2023;307(2):e223312. doi:10.1148/radiol.223312
- Humar P, Asaad M, Bengur FB, Nguyen V. TEMPORARILY REMOVED: ChatGPT is equivalent to first-year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service examination. *Aesthet Surg J.* 2023. doi:10.1093/asj/sjad130
- Gupta R, Herzog I, Park JB, et al. TEMPORARILY REMOVED: performance of ChatGPT on the plastic surgery in-service training examination. *Aesthet Surg J.* 2023. doi:10.1093/asj/sjad128
- Hopkins BS, Nguyen VN, Dallas J, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *J Neurosurg.* 2023. doi:10.3171/2023.2.JNS23419
- Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv.* 2023. doi:10.1101/2023.02.21.23285886
- Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectrum.* 2023;7(2):pkad010. doi:10.1093/jncics/pkad010
- Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT answers common patient questions about colonoscopy. *Gastroenterology.* 2023;165:509-511. doi:10.1053/j.gastro.2023.04.033
- Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? *J Stomatol Oral Maxillofac Surg.* 2023:101471. doi:10.1016/j.jormas.2023.101471
- Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison between ChatGPT and Google search as sources of postoperative patient instructions. *JAMA Otolaryngol Head Neck Surg.* 2023;149:556. doi:10.1001/jamaoto.2023.0704
- Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus.* 2023;15(2):e35179. doi:10.7759/cureus.35179
- Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digital Health.* 2023;5:e333-e335. doi:10.1016/S2589-7500(23)00083-3
- Lucy L, Bamman D. *Gender and Representation Bias in GPT-3 Generated Stories.* Association for Computational Linguistics; 2021.
- Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. *Nat Mach Intell.* 2021;3(6):461-463. doi:10.1038/s42256-021-00359-2
- Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak.* 2020;20(1):310. doi:10.1186/s12911-020-01332-6