European Archives of Oto-Rhino-Laryngology (2024) 281:2145–2151
https://doi.org/10.1007/s00405-023-08423-w

MISCELLANEOUS

ChatGPT vs UpToDate: comparative study of usefulness and reliability of Chatbot in common clinical presentations of otorhinolaryngology–head and neck surgery

Ziya Karimov[1] · Irshad Allahverdiyev[2] · Ozlem Yagiz Agayarov[3] · Dogukan Demir[3] · Elvina Almuradova[4,5]

Received: 2 September 2023 / Accepted: 18 December 2023 / Published online: 13 January 2024

Abstract

Purpose The usage of Chatbots as a kind of Artificial Intelligence in medicine is getting to increase in recent years. UpToDate® is another well-known search tool established on evidence-based knowledge and is used daily by doctors worldwide. In this study, we aimed to investigate the usefulness and reliability of ChatGPT compared to UpToDate in Otorhinolaryngology and Head and Neck Surgery (ORL–HNS).

Materials and methods ChatGPT-3.5 and UpToDate were interrogated for the management of 25 common clinical case scenarios (13 males/12 females) recruited from literature considering the daily observation at the Department of Otorhinolaryngology of Ege University Faculty of Medicine. Scientific references for the management were requested for each clinical case. The accuracy of the references in the ChatGPT answers was assessed on a 0–2 scale and the usefulness of the ChatGPT and UpToDate answers was assessed with 1–3 scores by reviewers. UpToDate and ChatGPT 3.5 responses were compared.

Results ChatGPT did not give references in some questions in contrast to UpToDate. Information on the ChatGPT was limited to 2021. UpToDate supported the paper with subheadings, tables, figures, and algorithms. The mean accuracy score of references in ChatGPT answers was 0.25–weak/unrelated. The median ($Q1$–$Q3$) was 1.00 (1.25–2.00) for ChatGPT and 2.63 (2.75–3.00) for UpToDate, the difference was statistically significant ($p < 0.001$). UpToDate was observed more useful and reliable than ChatGPT.

Conclusions ChatGPT has the potential to support the physicians to find out the information but our results suggest that ChatGPT needs to be improved to increase the usefulness and reliability of medical evidence-based knowledge.

Keywords Artificial intelligence · Chatbot · ChatGPT · ENT · UpToDate · Otorhinolaryngology and head and neck surgery

Introduction

The application of Artificial Intelligence (AI) in medicine is getting to increased last decade. Several studies reported the application of AI in clinical grading systems, assessment of cochlear implant function, parathyroid recognition, and prediction of clinical prognosis in otorhinolaryngology–head and neck surgery (ORL–HNS) [1–5]. Ethical concerns such as autonomy, beneficence, nonmaleficence, and justice were emphasized in the paper by Arambula et al. [6].

Chatbots are one of the trending topics of the AI nowadays. ChatGPT (by OpenAI) is one of the most commonly used Chatbots due to the literature. Several studies investigated the application of ChatGPT in medical exams, making a clinical diagnosis, article writing, etc. [4, 7–9].

UpToDate® is a well-known medical knowledge source for physicians that is used in daily clinical practice in

✉ Ziya Karimov
dr.ziya.karimov@gmail.com

1 Medicine Program, Ege University Faculty of Medicine, 35100 Izmir, Türkiye

2 Medicine Program, Istanbul University, Istanbul Faculty of Medicine, Istanbul, Türkiye

3 Department of Otolaryngology-Head and Neck Surgery, Izmir Tepecik Education and Research Hospital, Health Sciences University, Izmir, Türkiye

4 Department of Medical Oncology, Ege University Faculty of Medicine, Izmir, Türkiye

5 Department of Oncology, Medicana International Hospital, Izmir, Türkiye

worldwide and our hospital [10]. Studies reported its effectiveness on health care quality, decreasing diagnostic error and mortality, association with shorter length of hospital stay, and lower complication rate [11–14]. Another study reported that UpToDate was faster and gave detailed knowledge compared to similar database systems [15].

In this study, we aimed to compare the ChatGPT to UpToDate® for their usefulness and reliability in common clinical presentations of ORL–HNS.

Materials and methods

Study design: cross-sectional comparative

Study description

ChatGPT version 3.5 [accessed on 27 August 2023 (1–6 cases) and 23 October 2023 (7–25 cases)] and UpToDate® [accessed on 28 August 2023 (1–6 cases) and 23 October 2023 (7–25 cases)] were used for the study. We created 25 case scenarios that are related to the subspecialties of the ORL–HNS. We consider common clinical presentations of the ORL–HNS in the literature while making them [16–23]. These case scenarios include almost equal ratios of the sexes—female/male is 12:13—and different age segments 7 decades of life—of the patients. Clinical presentations are described in Table 1. Then, we asked the ChatGPT "Tell me how would you manage a "number of the age"-year-old male/female patient comes with "... symptoms" that started/for/since day/week/month. Give me references at the end of your response." and the meantime searched the case on UpToDate.

We assessed the accuracy of the references in the ChatGPT answers. The scale is: 0—the reference is not available with the described DOI number and source link or is not correct; 1—the reference is available with the described DOI number and source link but not so related to the specific topic; 2—the reference is available with the described DOI number and source link and strongly related to the topic.

Table 1 Clinical presentations

Case number	Case presentation
1	An 8-year-old male patient comes with a sudden hearing loss that started two days ago
2	A 41-year-old female patient comes with dizziness for a month
3	A 36-year-old male patient comes with recurrent epistaxis
4	A 17-year-old male patient comes with septal deviation and difficulty breathing
5	A 53-year-old female patient comes with snoring during sleep for two months
6	A 26-year-old female patient comes with a painless anterior cervical mass
7	A 22-year-old male patient comes with sneezing, nasal congestion, and, rhinorrhea for 3 days
8	A 33-year-old female patient comes with a runny nose with clear, thin fluid-like water
9	A 14-year-old male patient comes with nasal obstruction, malodorous, and sensation of a foreign body movement within the nose
10	A 55-year-old female patient comes with otalgia for 15 weeks
11	A 66-year-old male patient comes with a painless swelling in the cheek and difficulty in opening the mouth and swallowing
12	A 38-year-old female patient comes with a facial drop in the right that includes the eyelid
13	A 51-year-old male patient comes with ringing in the left ear for one week
14	A 48-year-old female patient comes with a painless, firm, hard thyroid mass for two months
15	A 19-year-old male patient comes with painful swelling in the gingiva since yesterday
16	A 62-year-old female patient comes with nasal obstruction, anosmia, epistaxis, facial pain and swelling, periorbital numbness, and rhinorrhea
17	A 30-year-old male patient is transferred from another rural medical center for consideration of primary hyperparathyroidism as a diagnosis
18	An 18-year-old male patient comes with painless, nonpruritic, bluish, darkly pigmented nodules/plaques on the oral mucosa and face
19	A 69-year-old female patient comes with painless, foul-smelling otorrhea, and conductive hearing loss in the left side
20	A 1-year-old female patient comes with otalgia and fever in the right ear for two days and tender mastoid
21	A 49-year-old male patient comes with dysphonia and difficult breathing for 4 months
22	A 39-year-old female patient comes with anosmia for 6 days
23	A 13-year-old male patient comes with recurrent epistaxis and unilateral nasal obstruction
24	A 2-year-old female patient comes with a fever, trismus, limited cervical neck extension, and dyspnea
25	A 28-year-old male patient comes with preauricular, intermittent, sharp pain, limited jaw motion, and clicking of the temporomandibular joint

Then, we calculate the mean score for each answer. In addition, we used the score from 1 to 3 to assess the usefulness of the ChatGPT and UpToDate answers; the scale was reported by Johnson et al. [24]: 1—incomplete answer and not useful; 2—semi-complete answer, somewhat useful but should need some extra knowledge; and 3—complete answer and useful in management.

Afterward, four reviewers assessed each case scenario for ChatGPT answers and related UpToDate papers regarding the search result. Reviewers were blinded to each other's assessment results.

Ethical approval

Not applicable to this study because of not include patient data.

Statistical analysis

The frequencies and percentages were given for categorical variables; and median (IQR: $Q1$–$Q3$) values were given for numerical variables as descriptive statistics. The agreement among the usefulness responses of reviewers for ChatGPT and UpToDate was determined using the coefficients of agreement of "Percent agreement (PA), Fleiss's κ and Gwet AC_1" [25–27]. All coefficients were presented with 95% confidence intervals (CI). Especially, due to the problems encountered with the Kappa coefficient [28], the Gwet AC_1 coefficient, which gives more consistent and reliable results, was preferred, but according to the published guide [26], the other two coefficients were also given to present more than one coefficient of agreement. The interpretation of the coefficients was carried out by Gwet's probabilistic method according to the Landis and Koch scale [29]. The McNemar–Bowker test was used to test the symmetry between ChatGPT and UpToDate usefulness responses of each reviewer. In addition, the Wilcoxon rank signed test was used to compare ChatGPT–UpToDate usefulness response means calculated over reviewers.

Statistical significance was assessed at $p < 0.05$ and all statistical analyses were performed using R software (R software, version 4.0.5, packages: arsenal-irrcac-ggplot2, R Foundation for Statistical Computing, Vienna, Austria; http://r project.org).

Results

A comparison of ChatGPT answers to UpToDate search results is described in Appendix 1 in supplementary material.

UpToDate supported its information with references from peer-reviewed journals, conference papers, book chapters, etc. However, ChatGPT did not give references in some questions. The overall mean accuracy score of references in ChatGPT answers was 0.25–weak/unrelated; the mean score of each question was described in Appendix 1 in supplementary material.

Table 2 Distribution of usefulness score in ChatGPT and UpToDate

Usefulness score	ChatGPT *n* (%)	UpToDate *n* (%)	*p* value
1	54 (54.0%)	0 (0%)	–
2	42 (42.0%)	27 (27.0%)	
3	4 (4.0%)	73 (73.0%)	
Median (*Q*1–*Q*3)	1.00 (1.25–2.00)	2.63 (2.75–3.00)	< 0.001

Table 3 Agreement among the usefulness responses of reviewers for ChatGPT and UpToDate

Coefficient	Value	95% CI	Interpretation
ChatGPT			
Gwet's AC_1	0.86	(0.78–0.93)	Substantial
Percent agreement	0.92	(0.89–0.97)	Almost perfect
Kappa	0.65	(0.41–0.90)	Moderate
UpToDate			
Gwet's AC_1	0.55	(0.32–0.78)	Fair
Percent agreement	0.73	(0.61–0.84)	Substantial
Kappa	0.30	(0.02–0.60)	Slight

AC agreement coefficient

The mean usefulness score was 1.5 ± 0.51 for ChatGPT and 2.73 ± 0.31 for UpToDate. Each reviewer scored the UpToDate responses 2 or 3 points; therefore, UpToDate had a higher overall mean score than ChatGPT. The median (Q1–Q3) was 1.00 (1.25–2.00) for ChatGPT and 2.63 (2.75–3.00) for UpToDate, and the difference was statistically significant (Wilcoxon test, $p < 0.001$) (Tables 2, 3 and Fig. 1). When the usefulness scores were compared for two groups for each reviewer, the result was found to be statistically significant (McNmear–Bowker p values for each reviewer, $p < 0.001$). The mean usefulness score distribution for ChatGPT and UpToDate is also described in Figs. 2 and 3, respectively.

UpToDate supported the topic with algorithms, figures, and tables that are different from ChatGPT. ChatGPT supported many answers by declaring "I am not a doctor" and advising to ask physicians for professional medical advice (highlighted in bold in Appendix 1 in supplementary material). The knowledge by the ChatGPT was extracted from sources with limited to older date, 2021 year (please look at the end of the answer of the first case scenario in Appendix 1 in supplementary material).

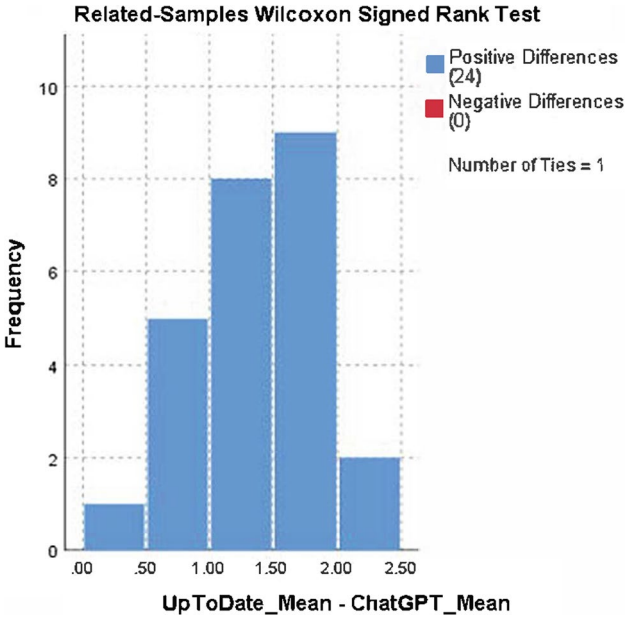

Fig. 1 Wilcoxon rank signed test result for comparison of ChatGPT-UpToDate usefulness response means

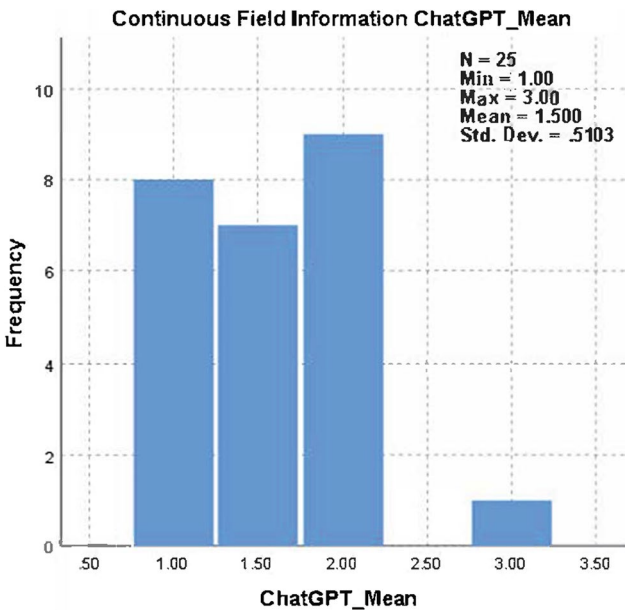

Fig. 2 The mean usefulness score distribution for ChatGPT

Discussion

The usage of AI in medicine is increasing. Its application to surgical fields has been on trend in recent years. ChatGPT (version 3.5) is a free AI Chatbot and was released by OpenAI at the end of last year. Afterward, it became a trended research topic for doctors and researchers very quickly. Over a thousand article is found in PubMed while searching with the "ChatGPT" keyword right now (accessed on 28 Aug 2023).

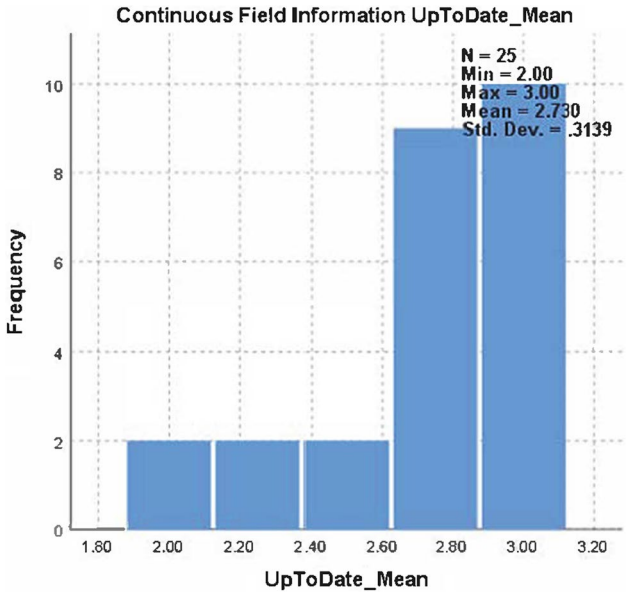

Fig. 3 The mean usefulness score distribution for UpToDate

There are a limited number of studies evaluating the ChatGPT in ENT&HNS in the literature. Most of them focused the exam-based work. Brennan et al. reported ChatGPT benefit on ear, nose, and throat (ENT) surgical education [30]. Qu et al. evaluated the diagnostic application of ChatGPT and reported the low quality of the Chatbot [4]. Hoch et al. assessed ChatGPT skills in single and multiple choice ENT board questions, and it performed a low correct answer percentage [8]. Other studies evaluated the triage and radiologic diagnosis accuracy of ChatGPT, but the accurate decision ratio was below that of physicians [31, 32]. Ayoub et al. compared the ChatGPT with Google Search and reported the first one had a good result for general medical knowledge but a worse result for medical advice than the second one [33].

UpToDate differs from ChatGPT with a subscription fee—institutional or personal [34]. However, ChatGPT was free access for people when released date, and version 3.5—used in our study—is still free, which makes it useful and reachable for all physicians. However, the upper version requires payment [35]. In addition, ChatGPT can search for more databases/websites and extract knowledge from various sources and languages. UpToDate supports sixteen languages (accessed 28 Aug 2023), but ChatGPT can extract data from more than 25 languages (accessed 28 Aug 2023). The papers' contents are the same in all languages in UpToDate. However, the answer may change with a wide range of different languages in ChatGPT.

Another nuance is that ChatGPT's answer depends on the question style and writing format. It requires "well-written" questions to get better answers. We should emphasize that answers to the same question also could result in a wary range depending on the question style. We tried the different versions of the question style and finally unanimously decided on "Tell me how would you manage a "number of the age"-year-old male/female patient comes with "... symptoms" that started/for/since day/week/month. Give me references at the end of your response" format. This nuance is subjective and could be a bias for studies asking open questions to ChatGPT like our study. When we decided to question format, we considered the details of the answers, and in addition, asked for references to improve sources. Because, when we asked ChatGPT a question without the phrase "Give me references at the end of your response", it did not give any references. Therefore, if a physician wants to get a reference to find out more information related to the topic, he/she should write an extra sentence while asking the question. This decreases the usefulness and reliability of the ChatGPT. Supporting the knowledge with references from peer-reviewed journals, conference papers, and book chapters increases the reliability and makes the knowledge transparent in UpToDate. Besides, promoting the topic with algorithms, figures, and tables makes UpToDate more systematic and beneficial.

UpToDate's search tool finds the related paper from its database regarding the search keywords. However, ChatGPT searches for many websites and databases. Papers in UpToDate included main subheadings that ease the physician's work to find the wanted information quickly within the paper. In addition, ChatGPT gave a subheading while asking about the management of patients, however, this heading contains non-specific sentences. Therefore, it looks like a useful feature of UpToDate. On the contrary, ChatGPT replies to the questions quickly differ from UpToDate and decrease the time to reach out for knowledge. It is one of the strong features of ChatGPT. UpToDate requires finding related papers and headings/subheadings within the papers manually and takes time.

ChatGPT's information base is limited to 2021 due to its training; therefore, it is a weak feature of Chatbot regarding further and most updated knowledge [33]. In addition, we observed the same result while looking at the references of Chatbot's answers. ChatGPT emphasized that in some answers reference parts its last knowledge was updated in September 2021. Informing the users on this issue is a good point regarding ethics. On the other hand, medical knowledge in UpToDate is reviewed and updated by doctors, well-experienced specialists, and academicians continuously.

Interestingly, ChatGPT cited and recommended the UpToDate while answering our questions in the 7th and 25th cases.

It was observed that ChatGPT give medical recommendation in contrast to basic medical knowledge in the reported studies [33]. This is an important concern for the safety of patients. In our study, we did not observe it. In addition, UpToDate gives medical recommendations, but these are evidence-based and supported by studies. In our study, most of the references in ChatGPT answers were unrelated to the question and some of them were inaccessible/unavailable. ChatGPT supported many answers by declaring "I am not a doctor" and advises referral to physicians for professional medical advice. This is a good point for ethical issues related to the AI. In addition, repeating sentences in the same answer in ChatGPT may be wordy while reading.

ChatGPT's answers may vary on different computers, in different locations, and at different times. The questions in our paper were answered differently according to this issue. We used the same computer device for asking the question to ChatGPT.

Twenty-five clinical case scenarios were investigated in the study which is a limited number. ChatGPT summarized the result itself, but we searched and selected the appropriate monograph in the UpToDate. Hence, it is a subjective factor of the authors' selection. Because there are several monographs for the same search result in the UpToDate. In this study, UpToDate had more usefulness scores and reliability than ChatGPT with statistical significance.

Conclusion

In this study, we aimed to investigate the usefulness and reliability of ChatGPT in comparison with UpToDate in common clinical presentations of otorhinolaryngology–head and neck surgery. In this stage, UpToDate looks more useful and reliable than ChatGPT. Developers need to improve the ChatGPT with evidence-based search and analysis skills and update its database.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00405-023-08423-w.

Acknowledgements We acknowledge Semiha Ozgul, PhD for her help in the biostatistics analysis of the data.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability Not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Knoedler L, Baecher H, Kauke-Navarro M, Prantl L, Machens HG, Scheuermann P, Palm C, Baumann R, Kehrer A, Panayi AC, Knoedler S (2022) Towards a reliable and rapid automated grading system in facial palsy patients: facial palsy surgery meets computer science. J Clin Med 11(17):4998. https://doi.org/10.3390/jcm11174998
2. Crowson MG, Dixon P, Mahmood R, Lee JW, Shipp D, Le T, Lin V, Chen J, Chan TCY (2020) Predicting postoperative cochlear implant performance using supervised machine learning. Otol Neurotol 41(8):e1013. https://doi.org/10.1097/MAO.0000000000002710
3. Wang B, Zheng J, Yu JF, Lin SY, Yan SY, Zhang LY, Wang SS, Cai SJ, Abdelhamid Ahmed AH, Lin LQ, Chen F, Randolph GW, Zhao WX (2022) Development of artificial intelligence for parathyroid recognition during endoscopic thyroid surgery. Laryngoscope 132(12):2516–2523. https://doi.org/10.1002/lary.30173
4. Qu RW, Qureshi U, Petersen G, Lee SC (2023) Diagnostic and management applications of chatgpt in structured otolaryngology clinical scenarios. OTO Open 7(3):e67. https://doi.org/10.1002/oto2.67
5. Lim SJ, Jeon E, Baek N, Chung YH, Kim SY, Song I, Rah YC, Oh KH, Choi J (2023) Prediction of hearing prognosis after intact canal wall mastoidectomy with tympanoplasty using artificial intelligence. Otolaryngol Neck Surg. https://doi.org/10.1002/ohn.472
6. Arambula AM, Bur AM (2020) Ethical considerations in the advent of artificial intelligence in otolaryngology. Otolaryngol Neck Surg 162(1):38–39. https://doi.org/10.1177/0194599819889686
7. Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, Sanchez-Barrueco A, Saga-Gutierrez C (2023) Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol. https://doi.org/10.1007/s00405-023-08104-8
8. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, Cotofana S, Alfertshofer M (2023) ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 280(9):4271–4278. https://doi.org/10.1007/s00405-023-08051-4
9. D'Amico RS, White TG, Shah HA, Langer DJ (2023) I asked a ChatGPT to write an editorial about how we can incorporate Chatbots into neurosurgical research and patient care…. Neurosurgery 92(4):663–664. https://doi.org/10.1227/neu.0000000000002414
10. Kinengyere AA, Rosenberg J, Pickard O, Kamya M (2021) Utilization and uptake of the UpToDate clinical decision support tool at the Makerere University College of Health Sciences (MakCHS), Uganda. Afr Health Sci 21(2):904. https://doi.org/10.4314/ahs.v21i2.52
11. Shimizu T, Nemoto T, Tokuda Y (2018) Effectiveness of a clinical knowledge support system for reducing diagnostic errors in outpatient care in Japan: a retrospective study. Int J Med Inf 109:1–4. https://doi.org/10.1016/j.ijmedinf.2017.09.010
12. Isaac T, Zheng J, Jha A (2012) Use of UpToDate and outcomes in US hospitals. J Hosp Med 7(2):85–90. https://doi.org/10.1002/jhm.944
13. Addison J, Whitcombe J, William GS (2013) How doctors make use of online, point-of-care clinical decision support systems: a case study of UpToDate©. Health Inf Libr J 30(1):13–22. https://doi.org/10.1111/hir.12002
14. Bonis PA, Pickens GT, Rind DM, Foster DA (2008) Association of a clinical knowledge support system with improved patient safety, reduced complications and shorter length of stay among Medicare beneficiaries in acute care hospitals in the United States. Int J Med Inf 77(11):745–753. https://doi.org/10.1016/j.ijmedinf.2008.04.002
15. Ahmadi SF, Faghankhani M, Javanbakht A, Akbarshahi M, Mirghorbani M, Safarnejad B, Baradaran H (2011) A comparison of answer retrieval through four evidence-based textbooks (ACP PIER, Essential Evidence Plus, First Consult, and UpToDate): a randomized controlled trial. Med Teach 33(9):724–730. https://doi.org/10.3109/0142159X.2010.531155
16. Neuhauser HK (2016) The epidemiology of dizziness and vertigo. Handb Clin Neurol 137:67–82. https://doi.org/10.1016/B978-0-444-63437-5.00005-4
17. Chandrasekhar SS, Tsai Do BS, Schwartz SR, Bontempo LJ, Faucett EA, Finestone SA, Hollingsworth DB, Kelley DM, Kmucha ST, Moonis G, Poling GL, Roberts JK, Stachler RJ, Zeitler DM, Corrigan MD, Nnacheta LC, Satterfield L (2019) Clinical practice guideline: sudden hearing loss (Update). Otolaryngol Head Neck Surg 161(1_suppl):S1–S45. https://doi.org/10.1177/0194599819859885
18. Franklin KA, Lindberg E (2015) Obstructive sleep apnea is a common disorder in the population—a review on the epidemiology of sleep apnea. J Thorac Dis 7(8):1311–1322. https://doi.org/10.3978/j.issn.2072-1439.2015.06.11
19. Hayois L, Dunsmore A (2023) Common and serious ENT presentations in primary care. InnovAiT 16(2):79–86. https://doi.org/10.1177/17557380221140131
20. Kaya Z, Mutlu V, Durna D (2023) KBB Acilleri. Akademisyen Kitabevi. https://books.google.com.tr/books?hl=en&lr=&id=X9rBEAAAQBAJ&oi=fnd&pg=PP1&dq=info:SOdGLzfFuzsJ:scholar.google.com&ots=q-8ecsmRgr&sig=mp8jowYQICS2zmzdv_LPWuskQ5g&redir_esc=y#v=onepage&q&f=false
21. Tunkel DE, Anne S, Payne SC, Ishman SL, Rosenfeld RM, Abramson PJ, Alikhaani JD, Benoit MM, Bercovitz RS, Brown MD, Chernobilsky B, Feldstein DA, Hackell JM, Holbrook EH, Holdsworth SM, Lin KW, Lind MM, Poetker DM, Riley CA, Schneider JS, Seidman MD, Vadlamudi V, Valdez TA, Nnacheta LC, Monjur TM (2020) Clinical practice guideline: nosebleed (Epistaxis). Otolaryngol Neck Surg 162(1_suppl):S1–S38. https://doi.org/10.1177/0194599819890327
22. Tunkel DE, Bauer CA, Sun GH, Rosenfeld RM, Chandrasekhar SS, Cunningham ER Jr, Archer SM, Blakley BW, Carter JM, Granieri EC, Henry JA, Hollingsworth D, Khan FA, Mitchell S, Monfared A, Newman CW, Omole FS, Phillips CD, Robinson SK, Taw MB, Tyler RS, Waguespack R, Whamond EJ (2014) Clinical practice guideline: tinnitus. Otolaryngol Neck Surg 151(S2):S1–S40. https://doi.org/10.1177/0194599814545325
23. Topuz MF (2022) Kulak Burun Boğaz Hastalıklarına Giriş. Akademisyen Kitabevi. https://books.google.com.tr/books?id=mVGmEAAAQBAJ&printsec=frontcover#v=onepage&q&f=false
24. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, Chang S, Berkowitz S, Finn A, Jahangir E, Scoville E, Reese T, Friedman D, Bastarache J, van der Heijden Y, Wright J, Carter N, Alexander M, Choe J, Chastain C, Zic J, Horst S, Turker

I, Agarwal R, Osmundson E, Idrees K, Kieman C, Padmanabhan C, Bailey C, Schlegel C, Chambless L, Gibson M, Osterman T, Wheless L (2023) Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq. https://doi.org/10.21203/rs.3.rs-2566942/v1
25. Gwet KL (2014) Handbook of Inter-rater reliability: the definitive guide to measuring the extent of agreement among raters, 4th edn. Advances Analytics, LLC
26. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL (2011) Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Int J Nurs Stud 48(6):661–671. https://doi.org/10.1016/j.ijnurstu.2011.01.016
27. Ph.d KLG. K. Gwet's Inter-Rater Reliability Blog: Benchmarking Agreement CoefficientsInter-rater reliability: Cohen kappa, Gwet AC1/AC2, Krippendorff Alpha. K. Gwet's Inter-Rater Reliability Blog (2023). Published December 12, 2014. https://inter-rater-reliability.blogspot.com/2014/12/benchmarking-agreement-coefficients.html. Accessed November 15, 2023
28. Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43(6):543–549. https://doi.org/10.1016/0895-4356(90)90158-l
29. Conger AJ (1980) Integration and generalization of kappas for multiple raters. Psychol Bull 88(2):322–328. https://doi.org/10.1037/0033-2909.88.2.322
30. Brennan LJ, Balakumar R, Bennett WO (2023) The role of ChatGPT in enhancing ENT surgical training: a trainees' perspective. J Laryngol Otol. https://doi.org/10.1017/S0022215123001354
31. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, Beam A (2023) The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. medRxiv. https://doi.org/10.1101/2023.01.30.23285067
32. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD (2023) Evaluating ChatGPT as an adjunct for radiologic decision-making. MedRxiv Prepr Serv Health Sci. https://doi.org/10.1101/2023.02.02.23285399
33. Ayoub NF, Lee YJ, Grimm D, Divi V (2023) Head-to-head comparison of ChatGPT versus google search for medical knowledge acquisition. Otolaryngol-Head Neck Surg. https://doi.org/10.1002/ohn.465
34. UpToDate Subscription Options (2023). https://www.wolterskluwer.com/en/solutions/uptodate/subscription-payment-options. Accessed 27 Nov 2023
35. Pricing (2023). https://openai.com/pricing. Accessed 27 Nov 2023