European Archives of Oto-Rhino-Laryngology (2025) 282:2669–2677
https://doi.org/10.1007/s00405-025-09255-6

MISCELLANEOUS

ChatGPT 4.0 and algor in generating concept maps: an observational study

Antonino Maniaci[1,2,10,11] · Caterina Gagliano[1] · Valerio Salerno[1] · Nicole Cilia[1] · Salvatore Lavalle[1] · Alberto Maria Saibene[2,3] · Giovanni Cammaroto[2,4] · Carlos Chiesa-Estomba[2,5] · Thomas Radulesco[2,6] · Luigi Vaira[2,7] · Giannicola Iannella[2,8] · Nicolas Fakhry[2,9] · Jerome Rene Lechien[2,10]

Received: 30 November 2024 / Accepted: 22 January 2025 / Published online: 20 February 2025

Abstract

Background To evaluate the performance of two AI systems, ChatGPT 4.0 and Algor, in generating concept maps from validated otolaryngology clinical practice guidelines.

Methods Concept maps were generated by ChatGPT 4.0 and Algor from four American Academy of Otolaryngology-Head and Neck Surgery Foundation (AAO-HNSF) clinical practice guidelines. Eight otolaryngology specialists evaluated the generated concept maps using the AI-Map questionnaire, covering concept identification, relationship establishment, hierarchical structure representation, and visual presentation. Chi-square tests and Kendall's tau coefficient were used for statistical analysis.

Results While no consistent superiority was observed across all guidelines, both AI systems demonstrated unique strengths. ChatGPT excelled in representing cross-connections between concepts and layout optimization, particularly for the Rhinoplasty guidelines (χ^2=6.000, p=0.050 for cross-connections). Algor showed strengths in capturing main themes and distinguishing general/abstract concepts, especially in the BPVV and Tympanostomy Tube guidelines (χ^2=8.000, p=0.046 for main themes in BPVV). Statistically significant differences were found in representing dynamic nature (favouring H&NMass-GPT, χ^2=7.571, p=0.023) and overall value and usefulness (favouring H&NMass-Algor, χ^2=7.905, p=0.019) for the H&N Masses guidelines.

Conclusion AI systems showed potential in automating concept map creation from otolaryngology guidelines, with performance varying across different medical topics and evaluation criteria. Further research is required to optimize AI systems for medical education and knowledge representation, highlighting their promise and current limitations.

Keywords Artificial intelligence · Concept mapping · Otolaryngology guidelines · Medical education · Knowledge representation

✉ Antonino Maniaci
antonino.maniaci@unikore.it

1 Department of Medicine and Surgery, University of Enna "Kore", Enna 94100, Italy

2 Study Group of Young-Otolaryngologists of the International Federations of Oto-rhino-laryngological Societies (YO-IFOS), Paris, France

3 Otolaryngology Unit, Santi Paolo e Carlo Hospital, Department of Health Sciences, Università degli Studi di Milano, Milan 26900, Italy

4 Department of Otolaryngology-Head and Neck Surgery, Forli Hospital, Forli 47122, Italy

5 Department of Otolaryngology-Head and Neck Surgery, San Sebastian University Hospital, San Sebastian, Spain

6 APHM, CNRS, IUSTI, La Conception University Hospital, ENT-HNS Department, Aix Marseille University, Marseille 13005, France

7 Maxillofacial Surgery Operative Unit, Department of Medicine Surgery and Pharmacy, University of Sassari, Sassari 07100, Italy

8 Department of 'Organi di Senso', University "Sapienza", Viale dell' Università, 33, Rome 00185, Italy

9 Department of Otolaryngology and Head and Neck Surgery, Aix-Marseille University, AP-HM, La Conception 33 Hospital, Marseille 13005, France

10 Department of Anatomy and Experimental Oncology, Mons School of Medicine, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons 7000, Belgium

11 School of Medicine University Enna Kore, Enna, Italy

Introduction

The application of artificial intelligence (AI) in healthcare has created new opportunities for the synthesis and sharing of knowledge. Huge language models like ChatGPT have proven to be remarkably effective at several tasks, including summarizing and creating content [1]. At the same time, idea mapping has gained popularity in the medical community to visualize intricate linkages and complex data inside clinical guidelines [2]. Concept maps are useful resources for medical education and practice because they make it easier to comprehend and implement clinical practice recommendations [3]. These concept maps are essential for improving students' understanding, retention, and critical thinking abilities [4]. Modern language models like ChatGPT have demonstrated promise in producing text, deciphering medical instructions and providing conceptual illustrations [1]. In the same way, Algor, an additional AI system created for knowledge representation (https://www.algoreducation.com/it), demonstrated concept mapping capabilities, especially in the context of medical education and the interpretation of guidelines. The purpose of this observational study was to evaluate how well ChatGPT 4.0 and Algor, two AI systems, perform while mapping four clinical practice guidelines from the American Academy of Otolaryngology-Head and Neck Surgery Foundation (AAO-HNSF) onto concept maps. The accuracy of idea recognition, the creation of conceptual linkages, the depiction of hierarchical structures, and the general visual presentation and user experience of the created concept maps were the main areas of focus for our investigation.

Methods

Study design

To assess the effectiveness of two AI systems, ChatGPT 4.0 and Algor, in generating concept maps from clinical practice guidelines, we performed a cross-sectional observational design. The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) recommendations are adhered to in this methods section to guarantee thorough and lucid reporting of our observational study [5].

Setting

Online AI platforms and digital copies of clinical practice guidelines were used in the study, which was carried out in a virtual setting. The study period was from July 1, 2024, to 1 September 2024.

Participants

Eight otolaryngology experts rated the created concept maps using the AI-Map questionnaire. Evaluators were 3 consultants with more than 10 years of experience, and 5 mid-career specialists (5–10 years). Their subspecialty expertise covered rhinology ($n=2$), otology ($n=2$), head and neck surgery ($n=2$), and general otolaryngology ($n=2$). All evaluators had active academic appointments and regular involvement in resident education. This diverse composition of evaluators was chosen to ensure comprehensive assessment across different levels of expertise and subspecialty perspectives. We used for our study four clinical practice guidelines published by the American Academy of Otolaryngology-Head and Neck Surgery Foundation (AAO-HNSF) [6–9].

Variables

The main outcomes measured in this study were the accuracy and quality of concept maps produced by each AI system. We assessed the accuracy of concept identification, the correctness of relationships between concepts, hierarchical structure representation, and the overall clarity and appearance of the presentation.

Data sources and measurement

All clinical practice guidelines were given in full text to ChatGPT 4.0 and Algor. The following prompt was given to both AI systems: *Based on the information and suggestions from the clinical practice guideline attached released by the American Academy of Otolaryngology-Head & Neck Surgery Foundation (AAO-HNSF), kindly construct an extensive concept map. The main ideas, links, hierarchies, and any connections between the many topic areas covered in the guideline should all be visually represented in the concept map. The concept map should be optimized for clarity, ease of understanding, and visual appeal.* We adopted a standardized assessment framework [10] to use validated metrics for concept map evaluation, including hierarchical structure (0–5), cross-linkages (0–5), and conceptual accuracy (0–5), alongside our current evaluation criteria. A group of 8 otolaryngology specialists assessed the generated concept maps using a predefined evaluation instrument, the AI-Map questionnaire (Suppl File I). It is composed of a 15-question questionnaire covering a range of important topics, including concept identification, relationship

representation, hierarchical organization, visual presentation, and overall efficacy. Raters use a 4-point Likert scale (0 = Poor, 1 = Fair, 2 = Good, and 3 = Excellent) to assign a score to each question. A total score is obtained by the 15 questions-scores sum. To reduce bias, the identity of the AI system that created each concept map was concealed from the expert. For every assessor, there was a different concept map evaluation order. Consequently, the specialist answer was analyzed by three independent judges (G.I, C.G., S.L.) to assess Inter-rater Reliability.

Statistical analysis

The Statistical Package for the Social Sciences for Windows (v.29.0; IBM Corp.) was used. Descriptive statistics were employed. To compare ChatGPT 4.0 and Algor's performance for each variable, chi-square tests were run. A statistically significant p-value was defined as one less than 0.05. Using Kendall tau, the judge's consistency (interrater reliability) for AI-Map ratings was evaluated.

Results

Basic concept map creation (Q1-Q3, Q6)

Rhinoplasty guidelines show fully comparable performance between systems (all $p > 0.405$, χ^2 ranging from 0.444 to 1.833) (Table 1). BPVV guidelines demonstrated Algor's most substantial advantage, showing superior performance in key concepts ($\chi^2 = 3.091$, $p = 0.378$), structure/content ($\chi^2 = 3.091$, $p = 0.378$), and significantly better performance in capturing main themes ($\chi^2 = 8.000$, $p = 0.046$). While both systems performed comparably in structural elements across Tympanostomy and H&N Masses guidelines ($p > 0.513$), Algor showed a consistent trend toward better performance in main theme capture across multiple guidelines (Tympanostomy: $\chi^2 = 3.818$, $p = 0.148$; H&N Masses: $\chi^2 = 3.000$, $p = 0.223$), except for H&N Masses key concepts where ChatGPT trended better ($\chi^2 = 2.333$, $p = 0.127$).

Structural and hierarchical representation (Q7-Q10)

ChatGPT showed excellence in two domains: cross-linking in Rhinoplasty maps, $\chi^2 = 6.000$, $p = 0.050$ (Fig. 1), and representation of dynamic nature in H&N Masses, $\chi^2 = 7.571$, $p = 0.023$, $\tau = -0.585$ (Table 2). Algor showed promising trends, $\tau > 0.4$, in hierarchical structure for Tympanostomy, $\chi^2 = 3.091$, $p = 0.213$, $\tau = 0.459$, and distinction of general/abstract concept in BPVV guidelines, $\chi^2 = 3.091$, $p = 0.378$, $\tau = 0.452$, though the differences were not statistically significant. Both systems were statistically equal on the rest of the aspects, including the dynamic nature representation of Tympanostomy $\chi^2 = 0.000$, $p = 1.000$ and general/abstract distinction for H&N Masses $\chi^2 = 0.000$, $p = 1.000$.

Visual presentation and user experience (Q4, Q11-Q13)

No statistically significant differences were found between ChatGPT and Algor across all guidelines, though several notable trends emerged with strong effect sizes ($\tau > 0.4$). Algor showed stronger trends in clear labelling for Rhinoplasty ($\tau = 0.532$) (Fig. 2) and BPVV maps ($\tau = 0.452$), as well as visual elements incorporation in Tympanostomy maps ($\tau = 0.471$) (Table 3). ChatGPT demonstrated better trends in layout optimization, particularly in Rhinoplasty ($\chi^2 = 7.571$, $p = 0.056$, $\tau = -0.432$) and BPVV guidelines

Table 1 Basic concept map creation performance across guidelines

Evaluation aspect	Guideline	Performance	Statistical values
Key concept identification	Rhinoplasty	Comparable	$\chi^2 = 0.444$, $p = 0.801$
	BPVV	Algor superior*	$\chi^2 = 3.091$, $p = 0.378$
	Tympanostomy	Algor trending better*	$\chi^2 = 3.091$, $p = 0.213$
	H&N Masses	ChatGPT trending better	$\chi^2 = 2.333$, $p = 0.127$
Relationship establishment	Rhinoplasty	Comparable	$\chi^2 = 1.833$, $p = 0.608$
	BPVV	Comparable	$\chi^2 = 1.200$, $p = 0.753$
	Tympanostomy	Algor trending better*	$\chi^2 = 3.091$, $p = 0.213$
	H&N Masses	Comparable	$\chi^2 = 1.333$, $p = 0.513$
Structure/content reflection	Rhinoplasty	Comparable	$\chi^2 = 1.810$, $p = 0.405$
	BPVV	Algor superior*	$\chi^2 = 3.091$, $p = 0.378$
	Tympanostomy	Comparable	$\chi^2 = 1.333$, $p = 0.513$
	H&N Masses	Comparable	$\chi^2 = 0.343$, $p = 0.842$
Main theme capture	Rhinoplasty	Comparable	$\chi^2 = 1.091$, $p = 0.580$
	BPVV	Algor significantly superior†	$\chi^2 = 8.000$, $p = 0.046$
	Tympanostomy	Algor trending better*	$\chi^2 = 3.818$, $p = 0.148$
	H&N Masses	Algor trending better*	$\chi^2 = 3.000$, $p = 0.223$

Abbreviation: *, noteworthy trend ($\tau > 0.4$); †, statistical significance ($p < 0.05$)

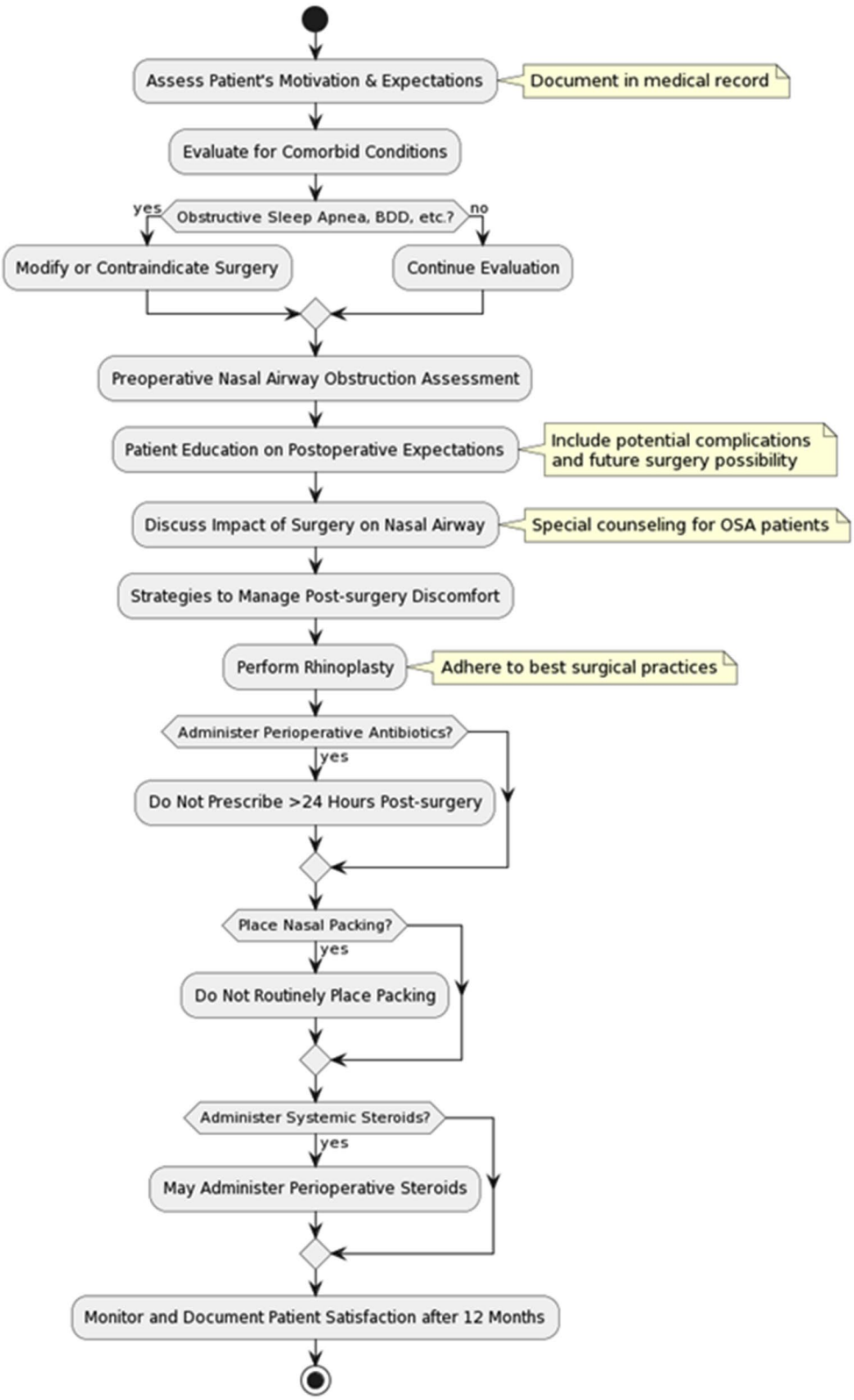

Fig. 1 Concept maps generated by ChatGPT 4.0 for the rhinoplasty clinical practice guideline

Table 2 Structural and hierarchical representation performance across guidelines

Evaluation aspect	Guideline	Performance	Statistical values
Hierarchical structure	Rhinoplasty	Comparable	χ^2=4.400, p=0.111, τ=-0.240
	BPVV	Comparable	χ^2=5.086, p=0.166, τ=-0.092
	Tympanostomy	Algor trending better*	χ^2=3.091, p=0.213, τ=0.459
	H&N Masses	Comparable	χ^2=0.343, p=0.842, τ=0.112
General/abstract concepts	Rhinoplasty	ChatGPT trending better	χ^2=4.952, p=0.084, τ=-0.308
	BPVV	Algor trending better*	χ^2=3.091, p=0.378, τ=0.452
	Tympanostomy	Comparable	χ^2=1.091, p=0.580, τ=0.193
	H&N Masses	Comparable	χ^2=0.000, p=1.000, τ=0.000
Cross-connections	Rhinoplasty	ChatGPT superior†	χ^2=6.000, p=0.050, τ=-0.347
	BPVV	Comparable	χ^2=2.952, p=0.399, τ=-0.161
	Tympanostomy	Comparable	χ^2=2.000, p=0.368, τ=0.371
	H&N Masses	Comparable	χ^2=1.091, p=0.580, τ=-0.145
Dynamic nature	Rhinoplasty	Comparable	χ^2=1.333, p=0.513, τ=0.044
	BPVV	Comparable	χ^2=2.143, p=0.543, τ=0.306
	Tympanostomy	Identical	χ^2=0.000, p=1.000, τ=0.000
	H&N Masses	ChatGPT superior†	χ^2=7.571, p=0.023, τ=-0.585

Abbreviations: *, noteworthy trend (τ>0.4); †, statistical significance ($p \leq 0.05$)

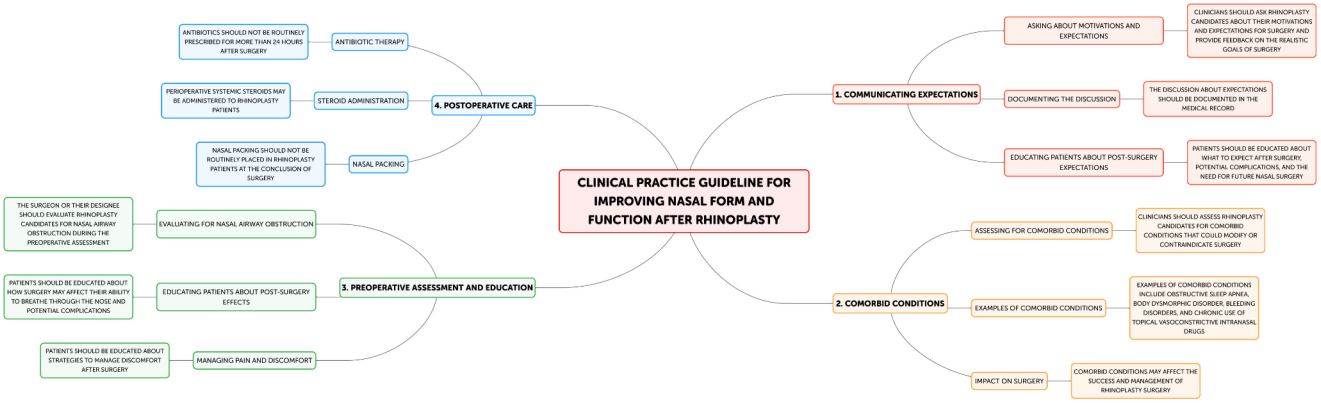

Fig. 2 Concept maps generated by algor for the rhinoplasty clinical practice guideline

Table 3 Visual presentation and user experience performance across guidelines

Evaluation aspect	Guideline	Performance	Statistical values
Ease of understanding	Rhinoplasty	Comparable	χ^2=5.111, p=0.164, τ=-0.297
	BPVV	Comparable	χ^2=3.600, p=0.308, τ=-0.258
	Tympanostomy	Comparable	χ^2=0.444, p=0.801, τ=0.120
	H&N Masses	Identical	χ^2=0.000, p=1.000, τ=0.000
Visual elements	Rhinoplasty	Comparable (both poor)	χ^2=2.444, p=0.485, τ=-0.099
	BPVV	Comparable	χ^2=1.111, p=0.774, τ=0.177
	Tympanostomy	Algor trending better*	χ^2=3.300, p=0.192, τ=0.471
	H&N Masses	Comparable	χ^2=1.167, p=0.558, τ=-0.265
Clear labelling	Rhinoplasty	Algor trending better*	χ^2=4.400, p=0.221, τ=0.532
	BPVV	Algor trending better*	χ^2=3.091, p=0.378, τ=0.452
	Tympanostomy	Comparable	χ^2=1.091, p=0.580, τ=-0.145
	H&N Masses	Identical	χ^2=0.000, p=1.000, τ=0.000
Layout Optimization	Rhinoplasty	ChatGPT trending better*	χ^2=7.571, p=0.056, τ=-0.432
	BPVV	ChatGPT trending better	χ^2=7.333, p=0.062, τ=-0.384
	Tympanostomy	ChatGPT slightly better	χ^2=2.400, p=0.301, τ=-0.258
	H&N Masses	Comparable	χ^2=0.444, p=0.801, τ=-0.160

Abbreviations: *, noteworthy trend (τ>0.4)

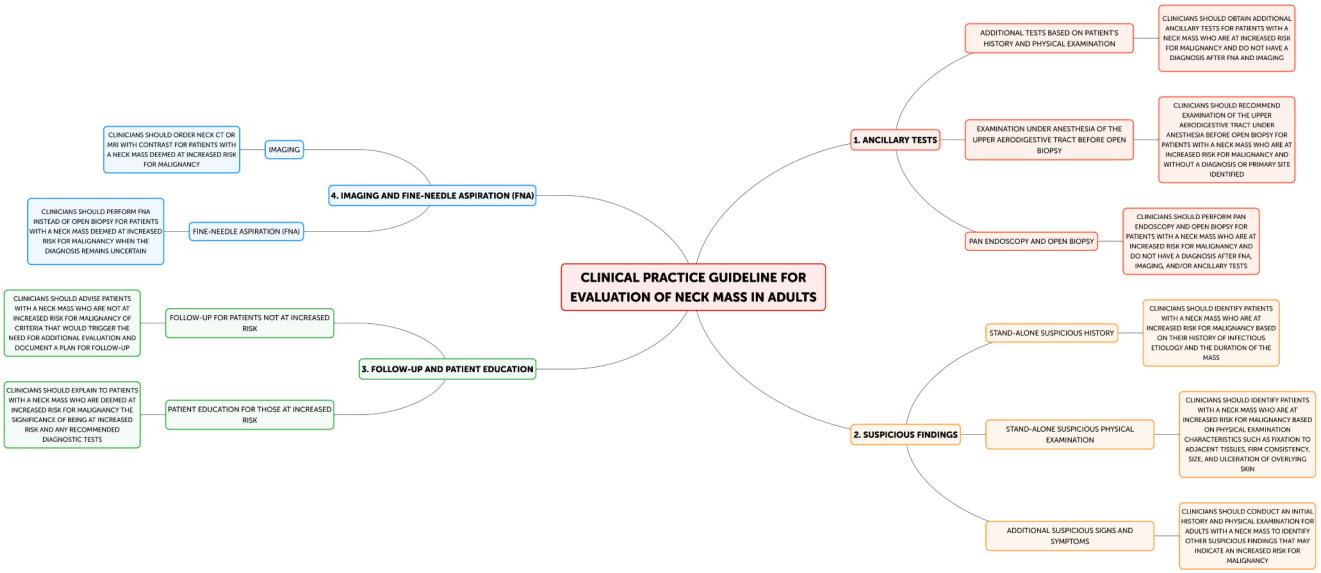

Fig. 3 Concept maps generated by Algor for the head and neck masses clinical practice guideline

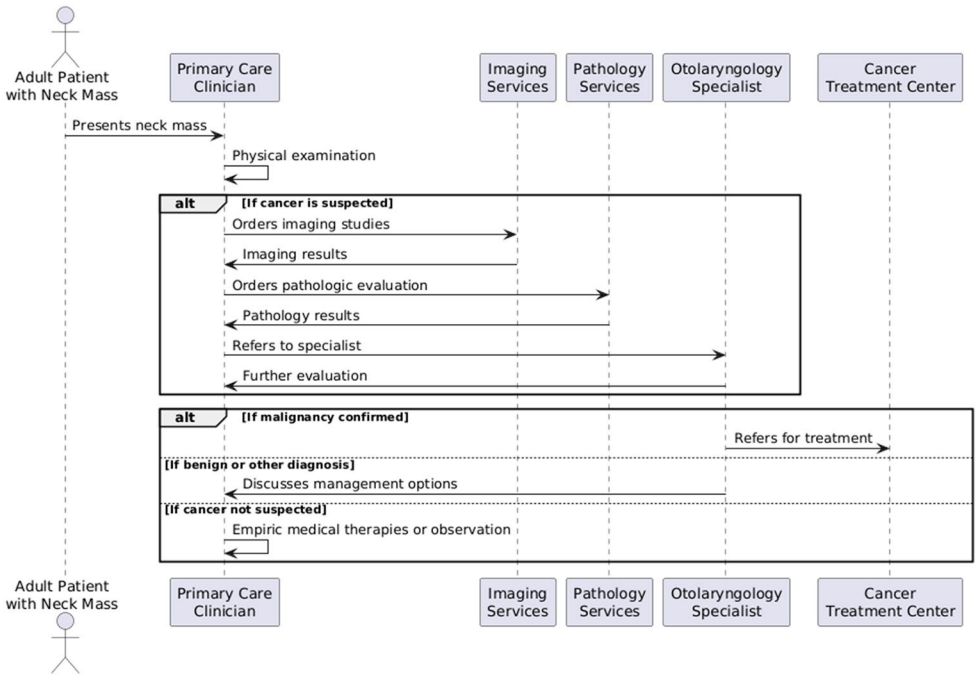

Fig. 4 Concept maps generated by Chat-GPT 4.0 for the head and neck masses clinical practice guideline

(χ^2=7.333, p=0.062, τ=-0.384), while both systems performed identically in several aspects of H&N Masses maps (ease of understanding and clear labelling: χ^2=0.000, p=1.000, τ=0.000) (see Figs.3 and 4).

Overall effectiveness and adaptability (Q5, Q14, Q15)

Algor demonstrated stronger performance in overall value and usefulness, showing greater outcomes in H&N Masses maps (χ^2=7.905, p=0.019) and positive trends in both BPVV (τ=0.452) and Tympanostomy guidelines (τ=0.459) (Table 4). Instead, ChatGPT demonstrated slightly better trends in overall effectiveness in Rhinoplasty maps (χ^2=3.300, τ=-0.196), though these differences were not statistically significant. Both systems were comparable in responsiveness to feedback across all guidelines, with notably poor ratings in Tympanostomy maps (χ^2=3.091, p=0.213, τ=-0.121), and no significant differences in overall effectiveness for BPVV, Tympanostomy, or H&N Masses maps.

Table 4 Overall effectiveness and adaptability performance across guidelines

Evaluation aspect	Guideline	Performance	Statistical values
Overall effectiveness	Rhinoplasty	ChatGPT trending better	χ^2=3.300, p=0.192, τ=-0.196
	BPVV	Comparable	χ^2=3.091, p=0.378, τ=0.167
	Tympanostomy	Comparable (both "Good")	χ^2=2.000, p=0.368, τ=0.029
	H&N Masses	Comparable	χ^2=2.000, p=0.368, τ=-0.371
Overall value/usefulness	Rhinoplasty	ChatGPT trending better	χ^2=4.400, p=0.111, τ=-0.240
	BPVV	Algor trending better*	χ^2=3.091, p=0.378, τ=0.452
	Tympanostomy	Algor trending better*	χ^2=3.091, p=0.213, τ=0.459
	H&N Masses	Algor superior†	χ^2=7.905, p=0.019, τ=0.311
Responsiveness to feedback	Rhinoplasty	Comparable	χ^2=3.333, p=0.343, τ=-0.167
	BPVV	Comparable	χ^2=2.000, p=0.572, τ=0.064
	Tympanostomy	Comparable (both "Poor")	χ^2=3.091, p=0.213, τ=-0.121
	H&N Masses	Comparable	χ^2=1.333, p=0.513, τ=-0.044

Abbreviations: *, noteworthy trend (τ>0.4); †, Indicates statistical significance ($p \leq 0.05$)

Inter-rater reliability

For the Rhino maps, the tau values ranged from −0.432 to 0.532, with the highest agreement observed for clear labelling of concepts and relationships (τ=0.532) and the lowest agreement for layout optimization (τ = -0.432). The BPVV maps showed a wider range of tau values, from −0.384 to 0.688, with the highest agreement for capturing main themes (τ=0.688, p=0.000) and the lowest for layout optimization (τ = -0.384, p=0.138). The Tube maps demonstrated a range of tau values from −0.258 to 0.507, with the highest agreement also observed for capturing main themes (τ=0.507, p=0.022) and the lowest for layout optimization (τ = -0.258, p=0.309). Lastly, the H&N Masses maps exhibited tau values ranging from −0.585 to 0.449, with the highest agreement for representing dynamic nature (τ = -0.585, p=0.004) and the lowest for both distinguishing general/abstract concepts and clear labelling (τ=0.000, p=1.000 for both).

Discussion

This observational study offers insightful information about how well ChatGPT 4.0 and Algor perform while creating concept maps based on clinical practice guidelines for otolaryngology. ChatGPT performed exceptionally well overall in illustrating the relationships between concepts and layout optimization—especially when it came to the Rhino guidelines. Our findings are corroborated by research by *Qadir et al.* [11] on AI-generated concept maps for computer science education, which discovered that AI systems were especially good at recognizing and illustrating intricate links between concepts. On the other hand, Algor demonstrated proficiency in identifying general abstract concepts and capturing major themes, particularly in the BPVV and Tube recommendations. This feature is essential because, as Novak and Cañas [3] pointed out, hierarchical concept map structuring is critical to efficient knowledge representation. In this sense, our results align with those of Wang et al. [12]., who found that AI-powered idea-mapping tools were very good at recognizing overarching themes and hierarchical structures in the literature related to medicine. The variation in performance across various medical issues, probably related to architectural AI differences, was highlighted by the statistically significant variations we found in representing the dynamic nature (favouring ChatGPT) and overall value and utility (favouring Algor) for the H&N Masses guidelines. ChatGPT due to its transformer-based model, appears to better capture long-range dependencies and complex relationships between concepts. This architecture is of special value for surgical guidelines involving interflowing aesthetics and functional outcomes. Likewise, its superior performance for representing the dynamic nature in H&N Masses (χ^2=7.571, p=0.023) indicated its ability to learn from sequential data and temporal connections in time series data. On the other hand, Algor outperformed BPVV guidelines (χ^2=8.000, p=0.046 for principal themes) and Tympanostomy Tube guidelines possibly because its knowledge representation framework is more specifically tailored towards an organized, hierarchical expression of blackboard design and appropriate theme structure. This is particularly useful for guidelines with defined diagnostic and treatment pathways. This heterogeneity is consistent with the findings of Nesbit and Adesope [4], who pointed out that subject matter and learning context might affect how effective concept maps are. Like our findings across several otolaryngology recommendations, a study by Lugo et al. [13] on AI-generated concept maps in nursing education discovered that AI performance differed significantly among different medical specialities. Our investigation uncovered areas where both AI systems needed to be improved, especially when it came to adding visual components and accurately capturing the dynamic nature of medical concepts. These restrictions

align with research by Zhang et al. [14], who noted comparable difficulties with AI-generated concept maps for biology instruction. They observed that whereas AI systems performed exceptionally well in text analysis and concept extraction, they had difficulty representing biological processes visually and expressing their dynamic interactions. Our study's varied levels of inter-rater agreement for the various evaluation criteria point to the continued subjectivity of the process of judging the calibre of concept maps produced by AI. This subjectivity creates problems for standardizing assessment techniques and emphasizes the need for future research to use more sophisticated and objective assessment instruments. Kim et al. [10] suggested a uniform rubric for assessing AI-generated idea maps in STEM education and emphasized the need for consistent assessment techniques, lends more credence to our findings in this regard.

Study limitation

Study limitations included any multi-stakeholder validation beyond expert evaluators, and restriction to otolaryngology guidelines. Although splitting it into four different guidelines provided useful insight, examining other specialities, such as cardiology and neurology, would give a better overview of the performance level of AI when faced with more-complex diagnostic and treatment scenarios. Content (main themes, $\tau = 0.688$) achieved stronger inter-rater agreement than design elements (layout, $\tau = -0.432$), in line with the findings of Kim et al. [10] indicating hierarchical organization reliability. These challenges of achieving consensus across complex algorithms are clear where H&N Masses maps show little agreement among algorithms ($\tau = 0.000$). While the composition of raters from diverse clinical backgrounds limited agreement, it provided valuable multi-perspective assessment. Standardized metrics for future research should be provided, including quantitative measures for visual elements and tools to automatically assess quality.

Conclusion

Our research indicates that ChatGPT 4.0 and Algor show potential in automating the construction of concept maps based on otolaryngology standards. However, their efficacy differs depending on the medical issue and evaluation criteria. To optimize AI systems for medical education and knowledge representation in the future, our study lays the groundwork for future research in these areas. Further studies should investigate the performance of fine-tuned AI models, including feedback loops via medical concept mapping and domain-specific training datasets, for tasks where current systems fail. In addition, Interim analyses should be performed on longitudinal follow-up over 12–24 months to determine how fast the AI system can pivot to user input, how performance improves across multiple updates of guidelines, including characteristics of learning curve and concepts maps temporal stability.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00405-025-09255-6.

Funding None.

Declarations

Competing interests The author Jerome R. Lechien was not involved with the peer review process of this article.

References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Amodei D (2020) Language models are few-shot learners. arXiv preprint arXiv:2005.14165
2. Daley BJ, Durning SJ, Torre DM (2016) Using concept maps to create meaningful learning in medical education. MedEdPublish 5(1):19
3. Novak JD, Cañas AJ (2008) The theory underlying concept maps and how to construct and use them. Technical Report IHMC CmapTools 2006-01 Rev 01-2008, Florida Institute for Human and Machine Cognition
4. Nesbit JC, Adesope OO (2006) Learning with concept and knowledge maps: a meta-analysis. Rev Educ Res 76(3):413–448
5. Vandenbroucke JP, Pocock SJ, Gøtzsche PC, von Elm E, Egger M, Altman DG, Initiative STROBE (2008) The STROBE declaration provides criteria for reporting observational studies in epidemiology. Strength Report Observational Stud Epidemiol 61(4):344–349 J Clin Epidemiol
6. Bhattacharyya N, Gubbels SP, Schwartz SR, Edlow JA, El-Kashlan H, Fife T, Holmberg JM, Mahoney K, Hollingsworth DB, Roberts R, Seidman MD, Prasaad Steiner RW, Tsai Do B, Voelker CC, Waguespack RW, Corrigan MD (2017) Clinical practice Guideline: Benign Paroxysmal positional Vertigo (update) executive Summary. Otolaryngol Head Neck Surg 156(3):403–416. https://doi.org/10.1177/0194599816689660
7. Pynnonen MA, Gillespie MB, Roman B, Rosenfeld RM, Tunkel DE, Bontempo L, Brook I, Chick DA, Colandrea M, Finestone SA, Fowler JC, Griffith CC, Henson Z, Levine C, Mehta V, Salama A, Scharpf J, Shatzkes DR, Stern WB, Youngerman JS, Corrigan MD (2017) Clinical practice Guideline: evaluation of the Neck Mass in adults executive Summary. Otolaryngol Head Neck Surg 157(3):355–371. https://doi.org/10.1177/0194599817723609
8. Rosenfeld RM, Tunkel DE, Schwartz SR, Anne S, Bishop CE, Chelius DC, Hackell J, Hunter LL, Keppel KL, Kim AH, Kim TW, Levine JM, Maksimoski MT, Moore DJ, Preciado DA, Raol NP, Vaughan WK, Walker EA, Monjur TM (2022) Executive Summary of Clinical Practice Guideline on Tympanostomy tubes in children (update). Otolaryngol Head Neck Surg 166(2):189–206. https://doi.org/10.1177/01945998211065661
9. Ishii LE, Tollefson TT, Basura GJ, Rosenfeld RM, Abramson PJ, Chaiet SR, Davis KS, Doghramji K, Farrior EH, Finestone SA,

Ishman SL, Murphy RX Jr, Park JG, Setzen M, Strike DJ, Walsh SA, Warner JP, Nnacheta LC (2017) Clinical practice Guideline: improving nasal form and function after Rhinoplasty Executive Summary. Otolaryngol Head Neck Surg 156(2):205–219. https://doi.org/10.1177/0194599816683156

10. Kim M, Pathak SA, Jacobson MJ, Zhang B, Gobert JD (2015) Cycles of exploration, reflection, and consolidation in model-based learning of genetics. J Sci Edu Technol 24(6):789
11. Qadir J, Taha AEM, Yau KLA, Ponciano J, Hussain S, Al-Fuqaha A (2020) Leveraging the force of formative assessment & feedback for effective engineering education. IEEE Access 8:45814–45835
12. Wang S, Ororbia A, Wu Z, Williams K, Liang C, Pursel B, Giles CL (2016), June Using prerequisites to extract concept maps from textbooks. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 317–326)
13. Lugo RG, Gjøsæter T, Khorram-Manesh A, Nordby A (2022) AI-generated concept maps for nursing education: a scoping review. Nurse Educ Today 109:105218
14. Zhang Y, Li X, Zhao L, Zou C (2021) Automatic generation of concept maps from texts: a comparative study of deep learning models. IEEE Access 9:47385–47395