European Archives of Oto-Rhino-Laryngology (2025) 282:1631–1639
https://doi.org/10.1007/s00405-024-08985-3

MISCELLANEOUS

Chasing sleep physicians: ChatGPT-4o on the interpretation of polysomnographic results

Christopher Seifen[1] · Tilman Huppertz[1] · Haralampos Gouveris[1] · Katharina Bahr-Hamm[1] · Johannes Pordzik[1] · Jonas Eckrich[1] · Harry Smith[2] · Tom Kelsey[2] · Andrew Blaikie[3] · Christoph Matthias[1] · Sebastian Kuhn[4] · Christoph Raphael Buhr[1,3]

Received: 23 June 2024 / Accepted: 27 August 2024 / Published online: 20 October 2024

Abstract

Background From a healthcare professional's perspective, the use of ChatGPT (Open AI), a large language model (LLM), offers huge potential as a practical and economic digital assistant. However, ChatGPT has not yet been evaluated for the interpretation of polysomnographic results in patients with suspected obstructive sleep apnea (OSA).

Aims/objectives To evaluate the agreement of polysomnographic result interpretation between ChatGPT-4o and a board-certified sleep physician and to shed light into the role of ChatGPT-4o in the field of medical decision-making in sleep medicine.

Material and methods For this proof-of-concept study, 40 comprehensive patient profiles were designed, which represent a broad and typical spectrum of cases, ensuring a balanced distribution of demographics and clinical characteristics. After various prompts were tested, one prompt was used for initial diagnosis of OSA and a further for patients with positive airway pressure (PAP) therapy intolerance. Each polysomnographic result was independently evaluated by ChatGPT-4o and a board-certified sleep physician. Diagnosis and therapy suggestions were analyzed for agreement.

Results ChatGPT-4o and the sleep physician showed 97% (29/30) concordance in the diagnosis of the simple cases. For the same cases the two assessment instances unveiled 100% (30/30) concordance regarding therapy suggestions. For cases with intolerance of treatment with positive airway pressure (PAP) ChatGPT-4o and the sleep physician revealed 70% (7/10) concordance in the diagnosis and 44% (22/50) concordance for therapy suggestions.

Conclusion and significance Precise prompting improves the output of ChatGPT-4o and provides sleep physician-like polysomnographic result interpretation. Although ChatGPT shows some shortcomings in offering treatment advice, our results provide evidence for AI assisted automation and economization of polysomnographic interpretation by LLMs. Further research should explore data protection issues and demonstrate reproducibility with real patient data on a larger scale.

Keywords ChatGPT · ChatGPT-4o · Large language models · Artificial intelligence · Obstructive sleep apnea · OSA · Polysomnography · PSG · Digital health · Sleep medicine

Christopher Seifen, Christoph Raphael Buhr have contributed equally.

✉ Christoph Raphael Buhr
buhrchri@uni-mainz.de

[1] Sleep Medicine Center & Department of Otolaryngology, Head and Neck Surgery, University Medical Center Mainz, Mainz, Germany

[2] School of Computer Science, University of St Andrews, St Andrews, UK

[3] School of Medicine, University of St Andrews, St Andrews, UK

[4] Institute for Digital Medicine, University Hospital of Giessen and Marburg, Philipps-University Marburg, Marburg, Germany

Introduction

With the introduction of artificial intelligence (AI) large language models (LLMs) healthcare professionals are experiencing a technical revolution. Launched in November 2022 by OpenAI Inc., Chat Generative Pre-Trained Transformer (ChatGPT) has become one of the most popular LLMs [1]. ChatGPT has garnered interest for its capability to analyze and convert textual and graphical inputs into text-based outputs, simulating human-like conversations and generating human-like content [2]. The inclusion of AI, such as ChatGPT, in routine clinical practice holds great potential: from the identification of research topics to support in routine

diagnostic processes, such as the analysis and interpretation of medical data sets, to the development of personalized therapy suggestions. In addition, AI can improve the situation of patients by providing them with easily accessible and understandable information. However, current shortcomings, limitations, and barriers to introducing LLMs in clinical practice include concerns about data privacy, accuracy and reliability of AI-generated insights, integration with existing medical systems, and the need for rigorous validation and regulation to ensure patient safety [3, 4]. ChatGPT has attracted more than a billion users in a remarkably short time, and a PubMed search on May 30, 2024 returned more than 3300 ChatGPT-related results, demonstrating the huge interest in this technology in the medical field.

The implementation of ChatGPT in the field of sleep medicine has recently been successfully tested, e.g. in the processing of patient questions on obstructive sleep apnea (OSA) from everyday clinical practice and in the evaluation of questions on OSA-specific surgeries [5–7]. OSA is the most common type of sleep-disordered breathing with increasing prevalence in the general adult population around the globe [8, 9]. OSA is characterized by upper airway collapses that clinically result in daytime sleepiness and fatigue [10]. The association with serious comorbidities such as arterial hypertension [11], coronary artery disease [12], or stroke [13] make OSA a major public health concern. Its diagnostic approach includes home sleep apnea testing (HSAT) or full-night polysomnography (PSG) in a sleep laboratory. The first-line therapy of moderate and severe OSA is the use of positive airway pressure (PAP) [14, 15].

Current clinical practice requires that HSAT or polysomnographic data are evaluated by experienced sleep medicine specialists. This procedure ensures high quality and therefore safety for the patient, but is time-consuming and requires trained personnel. In this context, ChatGPT could represent a valuable tool, e.g. for evaluating sleep medicine data or interpreting it in selected cases. However, it is currently uncertain whether ChatGPT can interpret polysomnographic results in such a way that a correct diagnosis is made and a guideline-oriented therapy suggestion is given. As we are convinced that LLMs, e.g. in the form of ChatGPT, will sooner or later be implemented in everyday clinical practice, this proof-of-concept study was designed to test whether ChatGPT is fundamentally capable of accurately evaluating selected data from fictitious polysomnographic results.

Material and methods

For this study, we generated $n = 40$ fictitious polysomnographic results from $n = 40$ consecutive fictitious patients. Patients 1–30 were designed to be simple cases, while patients 31–40 were more complex.

We defined the following polysomnographic parameters for each of the 40 fictitious patients:

- age,
- sex,
- body mass index (BMI),
- apnea hypopnea index (AHI, apneas and hypopneas per hour of sleep),
- apnea index (AI, apneic events per hour of sleep),
- hypopnea index (HI, hypopnea events per hour of sleep),
- cumulative apnea and hypopnea duration during sleep,
- oxygen desaturation index (ODI, desaturation episodes as a decrease in mean oxygen saturation of $\geq 3\%$ per hour of sleep),
- average oxygen saturation,
- percentage of cumulative time with oxygen saturation below 90% during sleep (T90),
- total sleep time (TST),
- sleep efficacy (ratio total sleep time to time in bed), and
- ratio AHI in supine position to not supine position.

Polysomnography is a standardized examination and its results are presented in a standardized format. For this study, we followed the format of the institutions' own accredited sleep laboratory. Figure 1 shows the template we used to generate the fictitious polysomnographic results. The detailed polysomnographic results of all 40 fictitious patients can be found in the supplemental materials.

The following general assumptions applied to patients 1–30 (simple cases) in order to minimize important confounders:

- first time polysomnographic recording due to assumed OSA-typical clinical complaints,
- the polysomnographic recording was performed technically correct in an accredited sleep laboratory under the supervision of a licensed technician,
- the polysomnographic recording had no technical problems and was not interrupted,
- no presence of comorbidities of any kind,
- no intake of medication of any kind,
- no sleep-disordered breathing other than OSA (e.g. periodic breathing or Cheyne-Stokes breathing), and
- the patient slept in all positions, with at least 60 min of the total sleep time spent on the back.

After testing different prompts, the polysomnographic data of patients 1–30 were passed to ChatGPT-4o (latest version) using the prompt shown in Fig. 2. ChatGPT-4o was given the general assumptions as mentioned before, asked to make a diagnosis and decide whether automatic PAP (aPAP) therapy is necessary for each fictitious patient based on their polysomnographic data. In each case, ChatGPT-4o was

Personal details	
Name	
Sex	
Age	
Body mass index	
Respiratory assessment	
Apnea hypopnea index	
Apnea index	
Hypopnea index	
Cumulative apnea and hypopnea duration	
Evaluation of pulse oximetry	
Oxygen desaturation index	
Average oxygen saturation	
T90	
Neurological evaluation	
Total sleep time	
Sleep efficacy	
Evaluation of body position	
Ratio apnea hypopnea index in supine position to not supine position	

Fig. 1 The template used to generate fictitious polysomnographic results. T90: percentage of cumulative time with oxygen saturation below 90% during sleep

The following basic assumptions apply to the following patient:
- There are no comorbidities.
- There is no long-term medication.
- There is a first polysomnography to clarify a sleep-related breathing disorder due to clinical signs of (obstructive) sleep apnoea.
- The polysomnography was performed technically correctly.
- The supine position is at least 60 min of the total time of sleep.

Order:
For the polysomnography data below, a diagnosis of obstructive sleep apnoea syndrome should be made. Sleep disorders such as central apnoea should not be taken into account, only the diagnosis of obstructive sleep apnoea syndrome is required. The answer should not exceed 200 characters.

The findings should be structured as follows:
Diagnosis: XX (If available, the sleep apnoea syndrome should be classified as 'none', 'mild', 'moderate' and 'severe').
Treatment recommendation: XX (aPAP therapy recommended / no aPAP therapy necessary).

Fig. 2 Prompt for patients 1–30, translated by DeepL (Cologne, Germany)

asked to use a maximum of 200 characters for the answer. Consequently, for patients 1–30 the prompt of Fig. 2 followed by the patients specific table (as text data) exemplary illustrated in Fig. 1 was passed to ChatGPT-4o.

For patients 31–40 (more complex cases) the following general assumptions were assumed in order to minimize important confounders:

– no presence of comorbidities of any kind,
– no intake of medication of any kind,
– at least mild obstructive sleep apnoea already confirmed by external polysomnography/polygraphy,
– the external polysomnography/polygraphy was performed due to clinical signs of (obstructive) sleep apnea,
– PAP therapy has already been initiated and trialed,
– PAP therapy was not tolerated and therefore rejected by patient,
– various mask fits have already been trialed (e.g. full-face mask, nasal mask, nasal cushion mask),
– presentation in our sleep laboratory for renewed polysomnography,
– the polysomnography was performed technically correctly, and
– the patient slept in all positions, with at least 60 min of the total sleep time spent on the back.

Thus, the prompt was slightly adjusted for patients 31–40 with PAP intolerance, see Fig. 3. ChatGPT-4o was given the general assumptions as mentioned before, asked to make a diagnosis and provide a therapy alternative for each fictitious patient based on their polysomnographic data. In each case, ChatGPT-4o was asked to use a maximum of 200 characters for the answer. Again, for patients 31–40 the prompt of Fig. 3 followed by the patients specific table (as text data) exemplary illustrated in Fig. 1 was passed to ChatGPT-4o.

In a second step, each fictitious polysomnographic result was interpreted by a board-certified sleep physician according to the same general assumptions. Based on the standard guidelines of the American Academy of Sleep Medicine (AASM) [16], a diagnosis was made and

The following basic assumptions apply to the following patient:
- There are no comorbidities
- There is no long-term medication.
- At least mild obstructive sleep apnoea already confirmed by external polysomnography/polygraphy
- External polysomnography/polygraphy was performed due to clinical signs of (obstructive) sleep apnoea
- PAP therapy has already been initiated and trialled
- PAP therapy is not tolerated and therefore rejected by patients
- Various mask fits have already been trialled (e.g. full face mask, nasal mask, nasal cushion mask)
- Presentation in our sleep laboratory for renewed polysomnography
- Polysomnography was performed technically correctly
- Supine position for at least 60 min of TST (total time of sleep)

Order:
For the polysomnography data below, a diagnosis of obstructive sleep apnoea syndrome should be made. Sleep disorders such as central apnoea should not be taken into account, only the diagnosis of obstructive sleep apnoea syndrome is required. The answer should not exceed 200 characters.

The findings should be structured as follows:
Diagnosis: XX (If present, the sleep apnoea syndrome should be classified as 'none', 'mild', 'moderate' and 'severe').
Treatment recommendation: If PAP intolerance exists: XX (alternative to PAP therapy, as this was not tolerated)

Fig. 3 Prompt for patients 31–40, translated by DeepL (Cologne, Germany)

an appropriate therapy was suggested. The diagnostic and therapeutic evaluation of the data by the board-certified sleep physician was not submitted to ChatGPT-4o in any of the 40 cases. In all cases, ChatGPT-4o only received the data as shown in Fig. 1.

In a final step, we compared the diagnosis and therapy suggestions of ChatGPT-4o to those of the sleep physician. Figure 4 shows the workflow of this study.

All data was collected in Microsoft Word and Excel sheets (Microsoft, Redmond, WA, USA). GraphPad Prism version 5.01 (GraphPad Software, Boston, MA, USA) was used for statistical analysis and graphical illustration.

All personal and polysomnographic patient data used in this study are fictitious. They do not correspond to real patient data.

Results

The complete patient cohort contained 40 adult individuals, 20 (50%) male and 20 (50%) female. The mean age was 50.18 ± 15.31 years, BMI was 28.45 ± 3.64 kg/m2 and AHI was 26.46 ± 19.52/h. According to AASM standard guidelines, three patients had no OSA (AHI < 5/h), twelve patients had mild OSA (AHI ≥ 5/h but < 15/h), ten patients had moderate OSA (AHI 15–30/h) and 15 patients had severe OSA (AHI > 30/h).

In the patient group with no OSA ($n = 3$), the sleep physician did not recommend any specific therapy.

In the patient group with mild OSA ($n = 12$), the sleep physician did not recommend any specific therapy in the simple cases. In the more complex cases, however, therapy suggestions included positional therapy, mandibular protrusion, or OSA surgery (case-specific therapy suggestions can be found in the supplemental material). In the patient group with mild OSA, the highest AHI was 13.8/h.

In the patient group with moderate OSA ($n = 10$), the lowest AHI was 17.5/h and the highest AHI was 29.5/h. For all patients, the sleep physician recommended treatment with aPAP in the pressure range of 5–9 millibar (mbar), except for one patient with a BMI of 32 kg/m2, who was recommended to use the pressure range of 5–10 mbar. Furthermore, all patients with moderate OSA were indicated for a polysomnographic control of the aPAP settings three months after the start of aPAP therapy.

In the patient group with severe OSA ($n = 15$), the lowest AHI was 30.8/h and the highest AHI was 82.5/h. In all patients with first time polysomnography (simple cases), the sleep physician recommended treatment with aPAP in the pressure range of 5–9 mbar, except for three patients with a BMI of ≥ 32 kg/m2, for whom the pressure range of 5–10 mbar was recommended. Again, polysomnographic control of the aPAP settings were indicated three months after the start of aPAP therapy. In the more complex cases, where former PAP therapy was not tolerated, the sleep physician indicated the evaluation of an alternative therapy by means of neurostimulation of the hypoglossal nerve.

As an overview, Fig. 5 shows patients 1–30 (simple cases) with regard to diagnosis and treatment decisions by the sleep

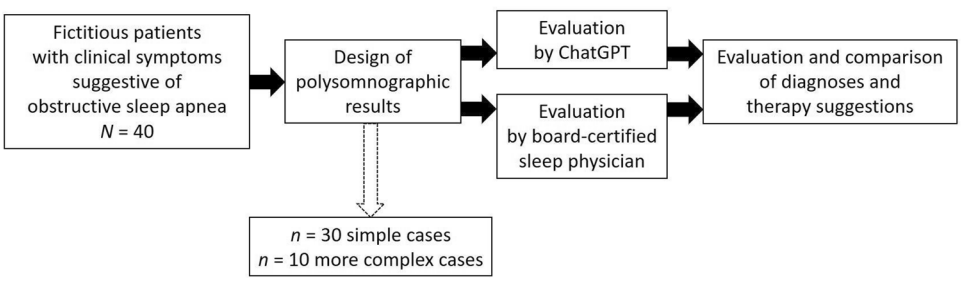

Fig. 4 Workflow of the study

board-certified sleep physician		
Patient	Diagnosis	Therapy (aPAP)
1	no	no
2	mild	no
3	mild	no
4	mild	no
5	mild	no
6	no	no
7	mild	no
8	mild	no
9	mild	no
10	no	no
11	moderate	yes
12	moderate	yes
13	moderate	yes
14	moderate	yes
15	moderate	yes
16	moderate	yes
17	moderate	yes
18	moderate	yes
19	moderate	yes
20	moderate	yes
21	severe	yes
22	severe	yes
23	severe	yes
24	severe	yes
25	severe	yes
26	severe	yes
27	severe	yes
28	severe	yes
29	severe	yes
30	severe	yes

ChatGPT-4o		
Patient	Diagnosis	Therapy (aPAP)
1	no	no
2	mild	no
3	mild	no
4	mild	no
5	mild	no
6	no	no
7	mild	no
8	mild	no
9	mild	no
10	no	no
11	moderate	yes
12	moderate	yes
13	moderate	yes
14	moderate	yes
15	moderate	yes
16	moderate	yes
17	moderate	yes
18	moderate	yes
19	moderate	yes
20	moderate	yes
21	severe	yes
22	moderate	yes
23	severe	yes
24	severe	yes
25	severe	yes
26	severe	yes
27	severe	yes
28	severe	yes
29	severe	yes
30	severe	yes

Concordance		
Patient	Diagnosis	Therapy (aPAP)
1	yes	yes
2	yes	yes
3	yes	yes
4	yes	yes
5	yes	yes
6	yes	yes
7	yes	yes
8	yes	yes
9	yes	yes
10	yes	yes
11	yes	yes
12	yes	yes
13	yes	yes
14	yes	yes
15	yes	yes
16	yes	yes
17	yes	yes
18	yes	yes
19	yes	yes
20	yes	yes
21	yes	yes
22	no	yes
23	yes	yes
24	yes	yes
25	yes	yes
26	yes	yes
27	yes	yes
28	yes	yes
29	yes	yes
30	yes	yes

Fig. 5 Visualization of diagnosis and therapy suggestions for patients 1–30 as stated by the sleep physician and ChatGPT-4o. Additionally, concordance between the two assessment instances is shown. aPAP: automatic positive airway pressure

physician as well as ChatGPT-4o In addition, the concordance between the two assessment instances is visualized. While the two assessment instances show 100% (30/30) concordance regarding the therapy suggestion, there is a disagreement on a single patient's diagnosis (Patient 22), leading to an overall concordance of 97% (29/30). In this specific case of disagreement, the patient reveals an AHI of 32.0/h, which was incorrectly diagnosed by ChatGPT-4o as a moderate OSA while being rated as severe OSA by the sleep physician according to AASM standard guidelines.

The patient cases 31–40 (more complex cases) represent patients not tolerating PAP therapy. Diagnosis and therapy suggestions as provided by the sleep physician and ChatGPT-4o as well as the concordance between the two assessment instances are shown in Fig. 6. Regarding the diagnosis, the sleep physician and ChatGPT-4o show 70% (7/10) concordance in the more complex patients with intolerance of PAP therapy. Whereas the sleep physician diagnosed mild OSA for patients 31–35 and severe OSA for patients 36–40, ChatGPT-4o incorrectly diagnosed three patients with a subsequent AHI of 12.4/h, 13.8/h and 12.8/h as moderate OSA. In terms of therapy suggestions, the sleep physician and ChatGPT-4o revealed 44% (22/50) concordance. While the sleep physician recommended no weight loss at all, ChatGPT-4o gave the recommendation for two patients. In reverse, ChatGPT-4o suggested a mandibular advancement device for every patient not tolerating PAP therapy, while the sleep physician recommended this form of therapy for three patients with mild OSA. ChatGPT-4o did not recommend the evaluation of an alternative therapy by means of neurostimulation of the hypoglossal nerve for any patient, while this procedure was suggested for every patient with severe OSA and PAP intolerance by the sleep physician.

Discussion

Different studies have shown and evaluated useful applications for LLMs in clinical practice [17–32]. However, studies addressing the use of ChatGPT in the interpretation of polysomnographic results are so far lacking. The present proof-of-concept study therefore aimed to investigate the extent to which ChatGPT is able to accurately evaluate selected data from fictitious polysomnographic results of patients with suspected OSA. We provide evidence that ChatGPT may serve as a valuable tool in everyday clinical

board-certified sleep physician						
Patient	Diagnosis	Therapy				
		lose weight	Mandibular advancement device	Side-lying therapy	surgery (UPPP)	HGNS
31	mild	no	yes	no	yes	no
32	mild	no	no	yes	no	no
33	mild	no	yes	no	yes	no
34	mild	no	no	yes	no	no
35	mild	no	yes	no	yes	no
36	severe	no	no	no	no	yes
37	severe	no	no	no	no	yes
38	severe	no	no	no	no	yes
39	severe	no	no	no	no	yes
40	severe	no	no	no	no	yes

ChatGPT-4o						
Patient	Diagnosis	Therapy				
		lose weight	Mandibular advancement device	Side-lying therapy	surgery (UPPP)	HGNS
31	mild	no	yes	yes	no	no
32	moderate	no	yes	yes	no	no
33	moderate	no	yes	yes	no	no
34	mild	no	yes	yes	no	no
35	moderate	no	yes	yes	no	no
36	severe	no	yes	no	yes	no
37	severe	no	yes	yes	yes	no
38	severe	yes	yes	yes	yes	no
39	severe	no	yes	no	yes	no
40	severe	yes	yes	yes	yes	no

Concordance						
Patient	Diagnosis	Therapy				
		lose weight	Mandibular advancement device	Side-lying therapy	surgery (UPPP)	HGNS
31	yes	yes	yes	no	no	yes
32	no	yes	no	yes	yes	yes
33	no	yes	yes	no	no	yes
34	yes	yes	no	yes	yes	yes
35	no	yes	yes	no	no	yes
36	yes	yes	no	yes	no	no
37	yes	yes	no	no	no	no
38	yes	no	no	no	no	no
39	yes	yes	no	yes	no	no
40	yes	no	no	no	no	no

Fig. 6 Visualization of diagnosis and therapy suggestions for patients 31–40 (patients with intolerance of therapy with positive airway pressure; more complex cases) as stated by the sleep physician and ChatGPT-4o. Additionally, concordance between the two assessment instances is shown

sleep medicine practice. Regarding the interpretation of fictitious polysomnographic data, ChatGPT provides highly accurate diagnostic and in most cases therapy suggestions that correspond to those of a board-certified sleep physician.

In the field of sleep medicine, ChatGPT was first challenged in November 2023: A prospective cross-sectional study showed that ChatGPT gave the same answers to ten questions on OSA-specific surgical interventions as 97 otolaryngologists with an expertise in OSA in 75% of cases [7]. Another study from December 2023 showed that ChatGPT was able to provide correct answers to questions about OSA in most cases, regardless of the prompting. The answers generated by the LLM were rated as being appropriate for patients [5]. In a third study from February 2024, the five most common questions asked by patients about PAP therapy were answered by sleep medicine specialists and by ChatGPT. The answers were then rated by up to 42 patients and, interestingly, it was found that the specialist's answer was preferred in four out of five cases [6]. The authors of this study explained the result of the study primarily by the fact that ChatGPT can only imitate, but not understand complex social interactions that are based on empathy and direct human contact, including elements such as body language.

Following the key objective of our study, evaluating the quality of ChatGPT-4o´s diagnosis and therapy recommendation of fictitious polysomnographic results, we tested different prompts initially. Prompts based on an open question did provide various different therapy options without a specific recommendation. Thus, we chose a prompt asking for a clear classification of OSA severity and a specific statement whether PAP (e.g. aPAP) therapy as the gold standard first line therapy of moderate to severe OSA is necessary. This setup was meant to provide a diagnosis and treatment recommendation that is as close as possible to that of a board-certified sleep physician. ChatGPT-4o and the board-certified sleep physician revealed 100% (30/30) concordance on recommendations for PAP therapy. Although there were some discrepancies between the two assessment instances in the classification of OSA severity, with only 97% (29/30) agreement, this did not affect the correct therapy suggestions for patients suitable for PAP therapy. Thus, the brief diagnosis and therapy suggestions of ChatGPT-4o for patients 1–30 could have been used literally in clinical practice without a single patient being treated incorrectly.

For patients 31–40 with PAP therapy intolerance, the prompt was modified to an open question asking about alternative treatment strategies, as the gold standard of first line PAP therapy was not applicable. In this context, the board-certified sleep physician and ChatGPT-4o showed a lower agreement of 70% (7/10) for diagnosis and 44% (22/50) for therapy suggestions. Interestingly, ChatGPT-4o did not mention the evaluation of an alternative therapy by neurostimulation of the hypoglossal nerve (hypoglossal nerve stimulation, HGNS) [33]. The neglect of HGNS therapy might be explained by the fact that there is much less literature available regarding this relatively new form of OSA therapy. ChatGPT-4o may lack familiarity with the topic and thus underestimates the relevance of this therapy, due to little training data about HGNS therapy. Selecting a suitable treatment alternative for patients with PAP therapy intolerance is a highly complex procedure. In routine clinical practice, this involves not only the assessment of specific polysomnographic data, but also personal preferences, individual anatomical factors, medical history, existing collaborations with medical supply distributors or available surgical approaches. One of the most important factors may be the clinical experience of the sleep physician.

We acknowledge several limitations of the present study. First, due to data protection aspects, all data used in this study are fictitious and were not collected from real polysomnographic recordings. Although our patient profiles represent a broad and typical spectrum of cases, ensuring a balanced distribution of demographics and clinical characteristics, this circumstance may have influenced the presented results independently. Secondly, general assumptions were defined that only apply in this form to a small number of patients presenting to a sleep medicine center. However, these general assumptions were necessary in order to achieve a stringent evaluation by ChatGPT and thus establish the comparability of the results to those of the board-certified sleep physician. Third, only those polysomnographic cases were designed that allowed little room for alternative therapy suggestions based on objective measurement data. In particular, subjective and anatomical factors must be taken into account for lower AHI ranges, as these are often decisive for a patient-specific therapy suggestion in such cases. Fourth, clinical symptoms, anatomical factors or sleep medicine questionnaires were not included in the design of the fictitious patients. This compromise was accepted in order to find out whether ChatGPT is fundamentally capable of making a correct diagnosis on the basis of specific, selected polysomnographic data and making a suitable therapy suggestion. Moreover, it should be noted that only cases of obstructive sleep apnea were evaluated in this study and other forms of sleep-related breathing disorders were not taken into account. In conclusion, the main limitation of this study is that only fictitious and basically very simplified sleep medicine cases were generated and subsequently used to demonstrate the potential of a ChatGPT-based evaluation. The transfer to clinical practice is therefore limited due to the often much greater complexity and individuality of sleep medicine cases.

Accepting these limitations, this is the first study to demonstrate the potential use of ChatGPT in the evaluation of polysomnographic data. Rather than the often discussed fear of replacing medical professionals, artificial

intelligence-based clinical decision support and LLMs are tools to augment doctoral intelligence. The integration of augmented intelligence in medicine has tremendous potential to enhance human cognition without replacing human labor. However, it is important to emphasize the need to address concerns about transparency, accountability, and data reliability to ensure successful implementation in clinical practice [34]. In the field of sleep medicine, ChatGPT has the potential to perform preliminary evaluation of polysomnographic results, e.g. when qualified medical personnel are scarce. Although ChatGPT has high potential to support sleep physicians, it may face difficulties in correctly weighting information from anamnestic discussions or physical examinations to form a diagnosis and make a suitable therapy suggestion. Further studies need to focus on these considerations and should include subjective factors, such as clinical symptoms, sleep medicine questionnaires or the results of the physical examination. In addition, further studies should address critical aspects of the use of LLMs, e.g. ChatGPT, such as the mandatory protection of the personal data used. In this context, it has already been pointed out by others that regulatory oversight of LLMs is essential [4, 35].

Conclusion

ChatGPT-4o and sleep physicians show a high level of agreement in terms of diagnosis and therapy suggestions based on fictitious polysomnographic results from simple patient cases. ChatGPT-4o shows interpretation shortcomings in cases of more complex polysomnographic result constellations. In conclusion, precise prompting of LLMs holds great potential to economize polysomnographic result interpretation in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00405-024-08985-3.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability All data used is accessible in the supplementary material.

References

1. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al (2023) Gpt-4 technical report. arXiv preprint arXiv:230308774
2. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
3. Dave T, Athaluri SA, Singh S (2023) ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 6:1169595. https://doi.org/10.3389/frai.2023.1169595. PMID: 37215063 ; PMCID: PMC10192861
4. Meskó B, Topol EJ (2023) The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 6(1):120. https://doi.org/10.1038/s41746-023-00873-0. PMID: 37414860; PMCID: PMC10326069
5. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS (2023) Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med 19(12):1989–1995
6. Martini A, Ielo S, Andreani M, Siciliano M (2024) ChatGPT: Friend or foe of patients with sleep-related breathing disorders?. Sleep Epidemiol 4:100076, ISSN 2667-3436. https://doi.org/10.1016/j.sleepe.2024.100076
7. Mira FA, Favier V, dos Santos Soberira Nunes H, de Castro JV, Carsuzaa F, Meccariello G et al (2023) Chat GPT for the management of obstructive sleep apnea: do we have a polar star? Eur Arch Oto-Rhino-Laryngol 281(4):2087–2093
8. 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
9. Heinzer R, Vat S, Marques-Vidal P, Marti-Soler H, Andries D, Tobback N et al (2015) Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study. Lancet Respir Med 3(4):310–318
10. Eckert DJ, White DP, Jordan AS, Malhotra A, Wellman A (2013) Defining phenotypic causes of obstructive sleep apnea identification of novel therapeutic targets. Am J Respir Crit Care Med 188(8):996–1004
11. Peppard PE, Young T, Palta M, Skatrud J (2000) Prospective study of the association between sleep-disordered breathing and hypertension. N Engl J Med 342(19):1378–1384
12. Loke YK, Brown JWL, Kwok CS, Niruban A, Myint PK (2012) Association of obstructive sleep apnea with risk of serious cardiovascular events. Circul Cardiovasc Qual Outcomes. 5(5):720–728
13. Marin JM, Carrizo SJ, Vicente E, Agusti AGN (2005) Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. The Lancet 365(9464):1046–1053
14. Epstein LJ, Kristo D, Strollo PJ Jr, Friedman N, Malhotra A, Patil SP et al (2009) Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. J Clin Sleep Med 5(3):263–276
15. Patil SP, Ayappa IA, Caples SM, Kimoff RJ, Patel SR, Harrod CG (2019) Treatment of adult obstructive sleep apnea with positive airway pressure: an american academy of sleep medicine clinical practice guideline. J Clin Sleep Med 15(2):335–343

16. Sateia MJ (2014) International classification of sleep disorders-third edition. Chest 146(5):1387–1394
17. Buhr CR, Smith H, Huppertz T, Bahr-Hamm K, Matthias C, Blaikie A, et al (2023) ChatGPT vs. consultants: a pilot study on answering otorhinolaryngology case-based questions. JMIR Med Educ (**forthcoming**)
18. Buhr CR, Smith H, Huppertz T, Bahr-Hamm K, Matthias C, Cuny C et al (2024) Assessing unknown potential—quality and limitations of different large language models in the field of otorhinolaryngology. Acta Otolaryngol 144(3):237–242
19. Dallari V, Sacchetto A, Saetti R, Calabrese L, Vittadello F, Gazzini L (2023) Is artificial intelligence ready to replace specialist doctors entirely? ENT specialists vs ChatGPT: 1–0, ball at the center. Eur Arch Otorhinolaryngol 281(2):995–1023
20. Chee J, Kwa ED, Goh X (2023) "Vertigo, likely peripheral": the dizzying rise of ChatGPT. Eur Arch Otorhinolaryngol 280(10):4687–4689
21. Hoch CC, Wollenberg B, Lüers J-C, Knoedler S, Knoedler L, Frank K et al (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
22. 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. PMID: 37614494 ; PMCID: PMC10442607
23. Nielsen JPS, von Buchwald C, Grønhøj C (2023) Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol 143(9):779–782
24. 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 170(6):1484-1491. https://doi.org/10.1002/ohn.465. PMID: 37529853
25. Warrier A, Singh R, Haleem A, Zaki H, Eloy JA (2024) The comparative diagnostic capability of large language models in otolaryngology. The Laryngoscope 134(9):3997-4002. https://doi.org/10.1002/lary.31434. PMID: 38563415
26. Long C, Lowe K, Zhang J, Santos Ad, Alanazi A, O'Brien D, et al (2024) A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology–Head and Neck Surgery Certification Examinations: Performance Study. JMIR Med Educ 10:e49970. https://doi.org/10.2196/49970. PMID: 38227351; PMCID: PMC10828939
27. Maniaci A, Saibene AM, Calvo-Henriquez C, Vaira L, Radulesco T, Michel J et al (2024) Is generative pre-trained transformer artificial intelligence (Chat-GPT) a reliable tool for guidelines synthesis? A preliminary evaluation for biologic CRSwNP therapy. Eur Arch Otorhinolaryngol 281(4):2167–2173
28. Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M et al (2023) Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol 281(4):2081–2086
29. Riestra-Ayora J, Vaduva C, Esteban-Sánchez J, Garrote-Garrote M, Fernández-Navarro C, Sánchez-Rodríguez C et al (2024) ChatGPT as an information tool in rhinology. Can we trust each other today? Eur Arch Oto-Rhino-Laryngol 281(6):3253–3259
30. Lechien JR, Chiesa-Estomba C-M, Baudouin R, Hans S (2023) Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. Eur Arch Otorhinolaryngol 281(4):2105–2114
31. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, et al (2024) Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. Eur Arch Oto-Rhino-Laryngol. https://doi.org/10.1007/s00405-024-08828-1. PMID: 39112556
32. Lechien JR, Briganti G, Vaira LA (2024) Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology–head and neck surgery. Eur Arch Otorhinolaryngol 281(4):2159–2165
33. Strollo PJ, Soose RJ, Maurer JT, de Vries N, Cornelius J, Froymovich O et al (2014) Upper-airway stimulation for obstructive sleep apnea. N Engl J Med 370(2):139–149
34. Bazoukis G, Hall J, Loscalzo J, Antman EM, Fuster V, Armoundas AA (2022) The inclusion of augmented intelligence in medicine: a framework for successful implementation. Cell Rep Med 3(1):100485
35. Ong JCL, Chang SY-H, William W, Butte AJ, Shah NH, Chew LST et al (2024) Ethical and regulatory challenges of large language models in medicine. The Lancet Digital Health. 6(6):e428–e432