Nature and Science of Sleep

Elevate Series

Open Access Full Text Article

ORIGINAL RESEARCH

Simulation-Based Evaluation of Large Language Models for Comorbidity Detection in Sleep Medicine – a Pilot Study on ChatGPT o1 Preview

Christopher Seifen[1,*], Katharina Bahr-Hamm[1,*], Haralampos Gouveris[1], Johannes Pordzik[1], Andrew Blaikie[2], Christoph Matthias[1], Sebastian Kuhn[3], Christoph Raphael Buhr[1,2]

[1]Sleep Medicine Center & Department of Otolaryngology, Head and Neck Surgery, University Medical Center Mainz, Mainz, Germany; [2]School of Medicine, University of St Andrews, St Andrews, UK; [3]Institute for Digital Medicine, Philipps University Marburg, University Hospital Giessen and Marburg, Marburg, Germany

*These authors contributed equally to this work

Correspondence: Christopher Seifen, Sleep Medicine Center & Department of Otolaryngology, Head and Neck Surgery, University Medical Center Mainz, Mainz, 55131, Germany, Email kim.seifen@unimedizin-mainz.de

Purpose: Timely identification of comorbidities is critical in sleep medicine, where large language models (LLMs) like ChatGPT are currently emerging as transformative tools. Here, we investigate whether the novel LLM ChatGPT o1 preview can identify individual health risks or potentially existing comorbidities from the medical data of fictitious sleep medicine patients.

Methods: We conducted a simulation-based study using 30 fictitious patients, designed to represent realistic variations in demographic and clinical parameters commonly seen in sleep medicine. Each profile included personal data (eg, body mass index, smoking status, drinking habits), blood pressure, and routine blood test results, along with a predefined sleep medicine diagnosis. Each patient profile was evaluated independently by the LLM and a sleep medicine specialist (SMS) for identification of potential comorbidities or individual health risks. Their recommendations were compared for concordance across lifestyle changes and further medical measures.

Results: The LLM achieved high concordance with the SMS for lifestyle modification recommendations, including 100% concordance on smoking cessation ($\kappa = 1$; $p < 0.001$), 97% on alcohol reduction ($\kappa = 0.92$; $p < 0.001$) and endocrinological examination ($\kappa = 0.92$; $p < 0.001$) or 93% on weight loss ($\kappa = 0.86$; $p < 0.001$). However, it exhibited a tendency to over-recommend further medical measures (particularly 57% concordance for cardiological examination ($\kappa = 0.08$; $p = 0.28$) and 33% for gastrointestinal examination ($\kappa = 0.1$; $p = 0.22$)) compared to the SMS.

Conclusion: Despite the obvious limitation of using fictitious data, the findings suggest that LLMs like ChatGPT have the potential to complement clinical workflows in sleep medicine by identifying individual health risks and comorbidities. As LLMs continue to evolve, their integration into healthcare could redefine the approach to patient evaluation and risk stratification. Future research should contextualize the findings within broader clinical applications ideally testing locally run LLMs meeting data protection requirements.

Keywords: ChatGPT, large language model, obstructive sleep apnea, comorbidities, health risk

Introduction

The rising prevalence of sleep-related breathing disorders, such as obstructive sleep apnea, and their strong association with serious comorbidities necessitate innovative diagnostic tools to support timely diagnosis and risk assessment. Obstructive sleep apnea is the most common type of sleep-related breathing disorder with its characteristic collapses of the upper airways during sleep, which usually lead to daytime sleepiness and fatigue.[1] Obesity, especially an elevated body mass index, is the strongest risk factor for obstructive sleep apnea.[2] Regardless of body habitus, adverse lifestyle habits such as smoking and alcohol consumption have also been associated with heightened risk for obstructive sleep apnea.[3,4] Conversely, the presence of obstructive sleep apnea has been linked to several serious and clinically relevant comorbidities: not only is there an association with an increased risk of hypertension,[5] but patients with obstructive sleep apnea are also at risk for coronary heart disease,[6,7] impaired glycemic control and type 2 diabetes,[8,9] as well as dyslipidemia[10] and non-alcoholic fatty liver disease.[11,12] Therefore, obstructive sleep apnea and

Received: 16 December 2024
Accepted: 11 April 2025
Published: 29 April 2025

its many adverse associated comorbidities have been recognized as a significant public health concern making early identification and intervention crucial for improving patient outcomes and reducing healthcare costs.

The increasing adoption of electronic patient records (EPRs) has fundamentally changed the availability and accessibility of patient data.[13] This influx of detailed patient information, while valuable, can overwhelm healthcare providers and contribute to the demand for medical evaluations as more patients present with flagged conditions or detailed histories. In this context, large language models (LLMs) like ChatGPT have garnered significant attention for their potential to transform clinical workflows through rapid and reliable data interpretation.[14] Their ability to process vast amounts of data and generate human-like responses presents opportunities for applications in sleep medicine.[15] However, these models are not without limitations. One major concern is their tendency to "hallucinate"-confidently presenting misinformation in areas where they lack adequate knowledge. To address this limitation, a new form of ChatGPT called "o1 preview" provides a reasoning mechanism for more reliable responses.[16] Moreover, LLMs providing medical device like output need to meet subsequent regulations.[17] The Medical Device Regulation (MDR) (EU) 2017/745 defines requirements for risk classification, clinical evaluation, and post-market surveillance for technologies that influence patient management or outcome in the European Union.[18] At the same time, the Food and Drug Administration (FDA) applies its own framework for software as a medical device, requiring premarket submissions, ongoing monitoring, and labeling requirements for software performing clinical decision support in the United States (FDA 2022).[19] Further legislation as the recently passed EU AI Act considers healthcare settings as high-risk scenarios (EU AI Act).[20] Despite these challenges, the use of LLMs in medical practice has expanded rapidly, with over 4700 ChatGPT-related publications on PubMed since its launch in late 2022. Notably, their role in sleep medicine, particularly for identifying comorbidities and health risks, remains un (der) explored.

Evaluating LLMs as a diagnostic tool coincides with the urgent need for innovative approaches in sleep medicine, particularly as the global prevalence of sleep-related breathing disorders, such as obstructive sleep apnea, continues to rise.[21,22] Especially in areas where medical professionals (eg sleep medicine specialists, SMS) are scarce, LLMs have the potential to perform preliminary evaluation of polysomnographic results.[15] The application of artificial intelligence (AI) in the field of sleep medicine is not entirely new as automate sleep study scoring was one of the earliest and most promising use cases.[23] Beyond that, the potential applications of AI in the field of sleep medicine are versatile, primarily due to the wealth of structured digital polysomnographic data: AI has been proposed to enable early detection, personalized treatment and improved management of sleep disorders, AI provides opportunities for advanced research using big data, and AI offers potential to detect early markers of neurodegenerative disorders linked to sleep abnormalities.[24] In this context, LLMs like ChatGPT enhance these capabilities further through complex algorithms that learn and improve from the provided data over time.

Due to the high incidence of obstructive sleep apnea with serious comorbidities, SMSs screen for these and other individual health risks, usually at the time of first polysomnography. This approach is considered complex, as it requires multidisciplinary clinical expertise (eg a fundamental knowledge of obstructive sleep apnea-associated comorbidities) and the necessity to appropriately assess personalized data based on an individual patient (risk) profile. As a support, SMSs most commonly use standardized questionnaires, the results of routine blood tests and clinical measurements (eg blood pressure values) besides polysomnographic results to complete this important but time-consuming task. At the time of this study, the potential of LLMs to support SMSs in this context remained uncertain. Consequently, this gap was addressed in this study by systematically evaluating the ChatGPT o1 preview's ability to detect comorbidities and health risks from fictitious patient data. Due to the obvious data protection constraints of web-based LLMs, the use of real-world patient data was not an option, possibly limiting the generalizability of the study. By using a simulation-based approach with fictitious patient profiles, we compare ChatGPT o1 preview's performance with that of an SMS, offering insights into its potential and limitations as a diagnostic support tool in sleep medicine.

Materials and Methods

Patient Profile Creation

To address the objective of this study, a broad spectrum of fictitious profiles of patients (N = 30), who presented to a sleep laboratory for the first-time polysomnography, were constructed by experienced sleep medicine specialists. In an initial step, the following personal data were defined for each of the 30 fictitious sleep medicine patients:

- age,
- sex,
- weight,
- height,
- body mass index,
- waist circumference,
- smoking status,
- alcohol consumption,
- blood pressure,
- GPT (glutamate pyruvate transaminase) level,
- GOT (glutamate oxaloacetate transaminase) level,
- GGT (gamma-glutamyl transferase) level,
- triglyceride level,
- total cholesterol level,
- HDL (high-density lipoprotein) level,
- LDL (low-density lipoprotein) level,
- hemoglobin A1c (HbA1c),
- apnea hypopnea index, and
- sleep medicine diagnosis.

The parameters listed above correspond to the standard clinical practice of the institute's own sleep medicine center, applicable to all individual patients at the time of their first polysomnography. The interested reader is referred to a more profound review of this topic as further details are beyond the scope of this article.[25] Since a previous study showed that ChatGPT-4o enables the evaluation of polysomnographic data to make a correct sleep medicine diagnosis and suggest a therapy,[15] we did not list further detailed polysomnographic parameters in the present study. Instead, we provided the final sleep medicine diagnosis (no/mild/moderate/severe obstructive sleep apnea) for all patients. Figure 1 shows the template we used to create all fictitious patient profiles.

In a next step, we defined the following general assumptions for all patients to minimize important confounders and ensure better comparability in the evaluation:

- the admission to a sleep laboratory was for first-time polysomnography due to clinical symptoms typical of obstructive sleep apnea,
- the polysomnography was performed technically correct in an accredited sleep laboratory under the supervision of a licensed technician,
- the polysomnography was performed without technical problems and the recording was not interrupted,
- the polysomnography was interpreted by an SMS according to standard guidelines of the American Academy of Sleep Medicine (AASM),[26]
- the patient slept in all positions, with at least 60 minutes of total sleep time in supine position,
- the diagnosis of obstructive sleep apnea was made or the presence of a sleep-related breathing disorder was ruled out,
- a sleep-related breathing disorder other than obstructive sleep apnea was not present (eg central sleep apnea, obesity hypoventilation syndrome),

FICTITIOUS PATIENT NO.		
Personal data	Age (years)	
	Sex (male, female)	
	Weight (kg)	
	Height (cm)	
	Body mass index (kg/m²)	
	Waist circumference (cm)	
	Smoking status [answer: never smoked; active smoking measured in pack-years; quit smoking measured in pack-years]	
	Alcohol consumption [answer: no alcohol consumption; approximate quantity in g/week]	
	Blood pressure (mmHg)	
Blood sample results	GPT level (U/l)	
	GOT level (U/l)	
	GGT level (U/l)	
	Triglyceride level (mg/dl)	
	Total cholesterol level (mg/dl)	
	HDL level (mg/dl)	
	LDL level (mg/dl)	
	HbA1c (%)	
Polysomnography results (selection)	Apnea hypopnea index (n/h)	
	Sleep medical diagnosis based on complete polysomnography without listing further results [answer: no OSA; mild OSA; moderate OSA; severe OSA]	

Figure 1 The template used to create fictitious patient profiles.

- personal data (age, sex, weight, height, body mass index, waist circumference, smoking status, alcohol consumption, blood pressure) were collected on the day of admission prior to polysomnography or briefly (<3 months) during consultation hour,
- a routine blood sample was taken around 8 AM on an empty stomach the morning after the polysomnography was performed,
- there were no known/diagnosed/treated comorbidities in the individual medical history of the patient, and
- no permanent medication was taken.

Then, we predefined a pattern to identify an individual health risk or a potentially existing comorbidity from each of the fictitious patient profiles. The predefined pattern consisted of the decision for ("yes") or against ("no") the following lifestyle changes and further medical measures:

- weight loss,
- quit smoking,
- reduce alcohol consumption,
- cardiological examination (eg by primary care physician or cardiologist) due to one or the combination of the following: elevated blood pressure, elevated triglyceride level, elevated total cholesterol level, abnormal associated lipoprotein levels,

- endocrinological examination (eg by primary care physician or endocrinologist) due to elevated HbA1c, and
- gastroenterological examination (eg by primary care physician, gastroenterologist or hepatologist) due to abnormal liver enzyme levels.

Figure 2 shows the template we used to display the predefined pattern for the decision for or against a measure.

Evaluation Process

Subsequently, each fictitious patient profile was interpreted by an SMS according to the above-mentioned general assumptions. For every profile, the predefined pattern for the decision for or against a measure in order to identify an individual health risk or a potentially existing comorbidity was fully processed.

Then, each fictitious patient profile was passed to the LLM (ChatGPT o1 preview) in order to do the same interpretation. Therefore, the prompt as shown in Figure 3 was passed to the LLM together with the patient's personal data, blood sample results, polysomnography results (selection) as shown in Figure 1 and the empty result template as shown in Figure 2. In order to avoid bias, the same prompt was used for each patient and the SMSs' interpretation was withheld from the LLM.

Outcome Measures

Concordance between ChatGPT and SMSs was assessed across six predefined measures, including recommendations for lifestyle modifications and further medical measures. Agreement was quantified using percentage concordance and analyzed for trends in LLM decision-making compared to SMS decision-making. Interrater reliability (ChatGPT vs SMS) was proved by Cohen's Kappa (κ) with $p < 0.05$ being statistically significant.[27] Figure 4 provides a brief summary of the study workflow.

All patient profiles with corresponding recommendations as stated by the SMS and the LLM are provided within the Supplementary Material.

Rating instances → SMS: sleep medicine specialist → LLM: large language model		SMS	LLM
To prevent potential individual health risks based on the above-mentioned patient data, the following measures should be recommended to the patient (in addition to sleep medicine therapy, if applicable)	Weight loss [answer: yes; no]		
	Quit smoking [answer: yes; no]		
	Reduce alcohol consumption [answer: yes; no]		
	Cardiological examination (e.g. by primary care physician or cardiologist) due to one or the combination of the following: elevated blood pressure, elevated triglyceride level, elevated total cholesterol level, abnormal associated lipoprotein levels [answer: yes; no]		
	Endocrinological examination (e.g. by primary care physician or endocrinologist) due to elevated HbA1c [answer: yes; no]		
	Gastroenterological examination (e.g. by primary care physician, gastroenterologist or hepatologist) due to abnormal liver enzyme levels [answer: yes; no]		

Figure 2 The template used to display the predefined pattern for the decision for or against a measure in order to identify an individual health risk or a potentially existing comorbidity.
Abbreviations: SMS, sleep medicine specialist; LLM, large language model (ChatGPT o1 preview).

https://doi.org/10.2147/NSS.S510254

The following basic assumptions apply to all fictitious patients:

- The admission to the sleep laboratory was for first-time polysomnography (PSG) due to clinical symptoms of obstructive sleep apnea (OSA),
- the PSG was performed technically correctly in an accredited sleep laboratory under the supervision of a licensed technician,
- the PSG was performed without technical problems and the recording was not interrupted,
- the PSG was interpreted by a sleep medicine specialist according to AASM standard guidelines,
- the patient slept in all positions, with at least 60 minutes of total sleep time in supine position,
- the diagnosis of OSA was made or the presence of a sleep-related breathing disorder was ruled out,
- a sleep-related breathing disorder other than OSA was not present (e.g. central sleep apnea, obesity hypoventilation syndrome),
- personal data (age, sex, weight, height, body mass index, waist circumference, smoking status, alcohol consumption, blood pressure) was collected on the day of admission prior to PSG,
- a routine blood sample was taken around 8 AM on an empty stomach the morning after the PSG was performed,
- there were no known/diagnosed/treated comorbidities in the individual medical history of the patient,
- no permanent medication was taken.

Please provide the content column "LLM" for the following patient:

Figure 3 Prompt was passed to the large language model (LLM) ChatGPT o1 preview.

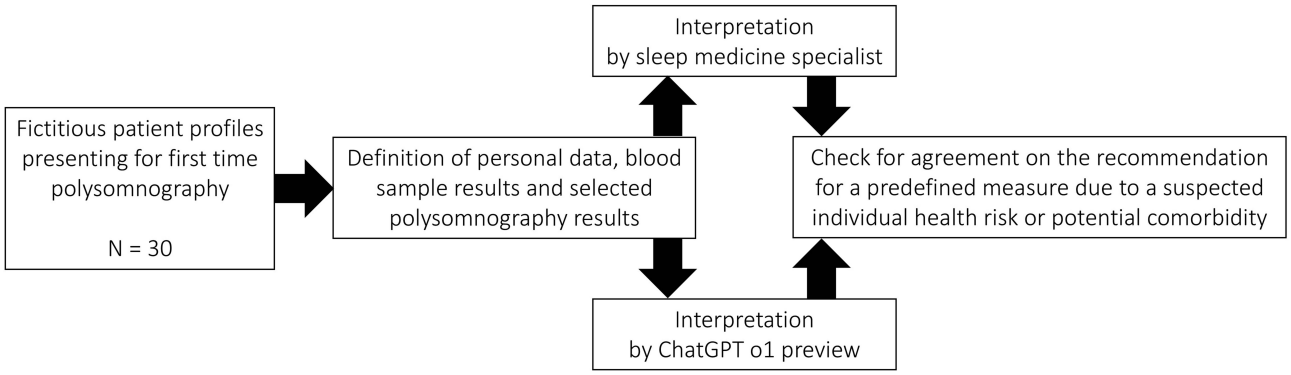

Figure 4 A brief summary of the study workflow.

All fictitious patient profiles were designed using Microsoft Word (Microsoft, Redmond, WA, USA). Graphical illustration was performed using Microsoft PowerPoint and Microsoft Excel (Microsoft, Redmond, WA, USA). Statistical analysis was performed using SPSS 27 (IBM, Armonk, NY, USA).

Ethical Statement

There is no need for any specific ethical approval in this kind of studies not involving patients, animals or cells. All patient data used in this study are fictitious. They do not correspond to the actual patient data of our or any third-party clinic. All right, title, and interest, if any, in and to output are assigned to the customer by ChatGPT (Open AI, San Francisco, United States) (see terms of use: ownership of content: https://openai.com/policies/row-terms-of-use/). All involved sleep medicine specialists are mentioned authors.

Results

The study evaluated concordance between the LLM and the SMS in identifying comorbidities and recommending further medical measures across 30 fictitious patient profiles. Overall, the LLM and the SMS showed 39.7% agreement (143/360 recommendations) in their evaluations.

Lifestyle Modifications

The LLM demonstrated high concordance with the SMS for lifestyle modification recommendations, achieving 100% agreement for smoking cessation (30/30 cases; $\kappa = 1$; $p < 0.001$) and 96.7% for alcohol reduction (29/30 cases; $\kappa = 0.92$; $p < 0.001$). Concordance for weight loss recommendations was similarly high at 93.3% (28/30 cases; $\kappa = 0.86$; $p < 0.001$).

Further Medical Examinations

Discrepancies were more pronounced in recommendations for further medical measures. The LLM recommended further cardiological examination in 96.7% (29/30 cases) compared to 53.3% by the SMS (16/30 cases). Similarly, further gastroenterological examination was suggested in 83.3% by the LLM (25/30 cases) versus 20.0% by the SMS (6/30 cases). Despite these differences, there was 96.7% agreement for further endocrinological examination (29/30 cases; $\kappa = 0.93$; $p < 0.01$), 56.7% agreement for further cardiological examination (17/30 cases; $\kappa = 0.08$; $p = 0.28$) and 33.3% for further gastroenterological examination (10/30 cases; $\kappa = 0.1$; $p = 0.22$). All individual recommendations provided for each of the 30 fictitious patients by the SMS and the LLM are shown in Figure 5.

Trends in Large Language Model Recommendations

The LLM tended to adopt a more conservative approach, frequently recommending a further medical measure even when the SMS deemed them unnecessary. This over-recommendation was observed in 35 instances (9.7%), particularly for

	Weight loss		Quit smoking		Reduce alcohol consumption		Cardiological examination		Endocrinological examination		Gastroenterological examination	
Patient	SMS	LLM	SMS	LLM	SMS	LLM	SMS	LLM	SMS	LLM	SMS	LLM
1	No	No	No	No	No	No	No	Yes	No	No	No	No
2	No	No	No	No	No	No	No	Yes	No	No	No	Yes
3	No	No	Yes	Yes	No	No	No	Yes	No	No	No	No
4	Yes	Yes	No	No	No	No	Yes	Yes	No	No	No	Yes
5	No	Yes	No	No	Yes	Yes	No	Yes	Yes	No	No	Yes
6	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	Yes
7	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes
8	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
9	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	Yes
10	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
11	No	No	No	No	No	No	Yes	Yes	No	No	No	Yes
12	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes
13	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	No	Yes
14	No	No	No	No	No	No	Yes	Yes	No	No	No	No
15	No	Yes	No	No	No	No	No	Yes	No	No	No	Yes
16	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
17	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes
18	Yes	Yes	No	No	No	Yes	No	Yes	Yes	Yes	No	Yes
19	No	No	No	No	No	No	No	Yes	No	No	No	No
20	No	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes
21	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	Yes
22	No	No	No	No	No	No	No	Yes	Yes	Yes	No	Yes
23	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
24	No	No	No	No	No	No	No	Yes	No	No	No	Yes
25	No	No	No	No	No	No	No	No	No	No	No	Yes
26	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes
27	No	No	No	No	Yes	Yes	No	Yes	No	No	Yes	Yes
28	No	No	Yes	Yes	No	No	Yes	Yes	No	No	No	Yes
29	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
30	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes	No	No
Yes vote	33% (10/30)	40% (12/30)	37% (11/30)	37% (11/30)	30% (9/30)	33% (10/30)	53% (16/30)	97% (29/39)	47% (14/30)	43% (13/30)	20% (6/30)	83% (25/30)

Figure 5 Recommendations as stated by the sleep medicine specialists (SMS) and the large language model (LLM) ChatGPT o1 preview for each of the 30 fictitious sleep medicine patients.

https://doi.org/10.2147/NSS.S510254

Patient	Weight loss	Quit smoking	Reduce alcohol consumption	Cardiological examination	Endocrinological examination	Gastroenterological examination
1	Yes	Yes	Yes	No	Yes	Yes
2	Yes	Yes	Yes	No	Yes	No
3	Yes	Yes	Yes	No	Yes	Yes
4	Yes	Yes	Yes	Yes	Yes	No
5	No	Yes	Yes	No	No	No
6	Yes	Yes	Yes	Yes	Yes	No
7	Yes	Yes	Yes	Yes	Yes	No
8	Yes	Yes	Yes	No	Yes	Yes
9	Yes	Yes	Yes	Yes	Yes	No
10	Yes	Yes	Yes	Yes	Yes	Yes
11	Yes	Yes	Yes	Yes	Yes	No
12	Yes	Yes	Yes	Yes	Yes	No
13	Yes	Yes	Yes	No	Yes	No
14	Yes	Yes	Yes	Yes	Yes	Yes
15	No	Yes	Yes	No	Yes	No
16	Yes	Yes	Yes	Yes	Yes	Yes
17	Yes	Yes	Yes	Yes	Yes	No
18	Yes	Yes	No	No	Yes	No
19	Yes	Yes	Yes	No	Yes	Yes
20	Yes	Yes	Yes	Yes	Yes	Yes
21	Yes	Yes	Yes	Yes	Yes	No
22	Yes	Yes	Yes	No	Yes	No
23	Yes	Yes	Yes	Yes	Yes	No
24	Yes	Yes	Yes	No	Yes	No
25	Yes	Yes	Yes	Yes	Yes	No
26	Yes	Yes	Yes	Yes	Yes	No
27	Yes	Yes	Yes	No	Yes	Yes
28	Yes	Yes	Yes	Yes	Yes	No
29	Yes	Yes	Yes	Yes	Yes	Yes
30	Yes	Yes	Yes	No	Yes	No
Concordance	93% (28/30)	100% (30/30)	97% (29/30)	57% (17/30)	97% (29/30)	33% (10/30)

Figure 6 Concordance between the recommendations of the sleep medicine specialist (SMS) and the large language model (LLM) ChatGPT o1 preview for further medical measures in each of the 30 fictitious sleep medicine patients.

cardiological and gastroenterological referrals. Conversely, there was only one instance where the SMS recommended a further medical measure but the LLM did not (endocrinological examination for patient 5). The concordance in recommendations for each patient between SMS and LLM is visualized in Figure 6.

Discussion

Studies addressing the use of LLMs as a tool to help identify individual health risks or potentially existing comorbidities in sleep medicine patients have been lacking. This study demonstrates the potential of the LLM ChatGPT o1 preview as a complementary tool for identifying comorbidities and health risks in sleep medicine patients. The model exhibited high concordance with sleep medicine specialists (SMSs) for lifestyle modification recommendations, including smoking cessation and alcohol reduction. However, its tendency to over-recommend further medical evaluations, particularly cardiological and gastroenterological examinations, highlights both opportunities and challenges for its integration into clinical workflows.

The present data show a high agreement between the tested LLM and the SMS with the highest concordance for the recommendation to quit smoking and to lose weight. In contrast, the lowest level of agreement was detected for the recommendation for further cardiological examination and gastroenterological examination. This disagreement results from the fact that the LLM recommended further examination much more frequently than the SMS. For example, the

LLM advised further cardiological examination in 96.7% (29/30 cases) and gastroenterological examination in 83.3% (25/30 cases), while the SMS recommended these measures in 53.3% (16/30 cases) and 20.0% (6/30 cases), respectively. Furthermore, we found 35 (35/360=9.7%) constellations of disagreement in which the LLM recommended an additional measure, and the SMS did not. Yet, there was only one (1/360=0.3%) constellation in which the SMS recommended an additional measure, but the LLM did not. We initially assumed that this might be explained by different norm values used by the SMS and the LLM. Therefore, we added internal normal values to the prompt and repeated the entire interpretation for each patient. However, this procedure showed little impact on the results (data not shown). Thus, it can be assumed that the LLM does tend to be more conservative than the SMS in recommending more investigations. This might have been further reinforced by the "reasoning" function of the latest version of ChatGPT. The LLM therefore demonstrates significant limitations in this study due to their inherently protective and data-restricted nature. For both the patient and the healthcare system, over-treatment is a relevant burden. While unnecessary examinations may cause stress and set patients at undue risk, they are also expensive. However, more advanced and precise reasoning mechanisms might reduce LLMs' tendency to "over-recommend" in the future.

Despite the proneness of ChatGPT o1 preview to be conservative, the level of agreement still points to it being a useful tool in the assessment of patients. Growing evidence supports an independent association of obstructive sleep apnea with a wide range of comorbidities, including cardiovascular and metabolic.[28,29] Therefore, SMSs are faced with the complex task of correctly identifying individual health risks or undiagnosed comorbidities, naming them as such, communicating them and advising patients accordingly, where LLMs could assist. More specifically, special attention needs to be paid to those patients who are considered healthy or where no comorbidities are known until their first polysomnography. For the individual patient, the timely screening for sleep-related breathing disorders, such as obstructive sleep apnea, and associated comorbidities helps to improve and, in some cases, slow the progression of these disorders.[30] In addition to individual health benefits, it is in the economic interest of healthcare systems worldwide to diagnose and treat sleep-related breathing disorders and its associated comorbidities in a timely manner.[31]

There are several limitations to this study. First, the presented results are based on fictitious (and not real-world) patient data. This measure could have influenced the presented results individually but was necessary due to (local) data protection aspects to protect patient rights. Second, strict general assumptions were defined to enable the stringent interpretation by the LLM establishing the comparability of the results to those of the SMS. We recognize that such general assumptions only apply to a fraction of sleep medicine patients. Third, one of the general assumptions was that only a diagnosis of obstructive sleep apnea was considered unless a sleep-related breathing disorder has been ruled out. This neglects the relevant patient group with central sleep apnea, which also often has its own characteristic comorbidity profile.[32] Fourth, we considered only a predefined pattern for the decision for or against a further medical measure to identify an individual health risk or a potentially existing comorbidity. With no room for comments, the advising of individual patients is limited. Fifth, the data within this study was made available to the LLM in a pre-sorted format. In everyday clinical practice, LLMs could face the challenge of filtering appropriate data from a larger amount. This could limit the transfer of the results of this study to clinical reality. Sixth, as a rule, comorbidity profiles in sleep medicine patients are less "black or white" as presented in this study but typically with higher complexity. Seventh, while LLMs can access and process medical information, they fundamentally lack the capacity for direct personal interaction and nuanced clinical judgment. The experience-based, interpersonal interpretation by SMSs, which involves assessing individual health risks and potential comorbidities beyond raw data, currently remains irreplaceable by LLM technologies. Eighth, no subgroup analysis was conducted in order to enhance the understanding of the LLMs' generalizability. In addition, the current study did not apply interpretive methods such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) to shed light into understanding the decision-making of ChatGPT.[33]

Accepting these limitations, this is the first study that demonstrates the potential use of the LLM ChatGPT o1 preview as a helpful tool in the identification of health risks and potential existing comorbidities in sleep medicine patients. In the field of sleep medicine, the use of LLMs has significant potential for organizing and evaluating large amounts of data. LLMs show particular promise in addressing serious comorbidities associated with sleep-related breathing disorders. By systematically identifying and flagging individual health risks, early detection of unrecognized comorbidities and appropriate additional medical evaluations can be recommended.

As AI applications in sleep medicine evolve, it is critical to consider ethical and legal considerations, particularly to prevent the accentuation of inequalities in healthcare such as sex disparities in evaluating sleep disorders.[24] The AASM emphasizes the need for diverse training datasets, transparency from manufacturers and thorough testing of AI tools to ensure they are effective and equitable.[23] In addition, there is a need to establish robust guidelines for machine learning methods to ensure reliability and facilitate the responsible integration of AI into clinical practice for the management of sleep disorders.

Legal and safety aspects of AI are increasingly discussed.[17] New legislation and guidelines such as the EU AI Act or the FDAs "Policy for Device Software Functions and Mobile Medical Applications" show the legislator's efforts to regulate risks and liability issues in AI.[19,20] Beyond regulatory and security concerns ethical considerations mainly include patients' perception of AI in healthcare. Here, recent studies showed patients reservation towards AI in medical applications even when AI-generated content is medically supervised.[34,35] From our perspective, deploying AI technologies such as LLMs in medical practice necessitates patients' acceptance in the first place. In our view, the best way to encounter skepticism is through education and training. Therefore, a profound AI education and training for both medical staff and patients is crucial for clinical implementation. Understanding the capabilities, but also the limitations of AI, can promote critical questioning among medical staff, thereby enhancing safety and ultimately increasing acceptance among patients.[36] This aspect underscores the importance of the presented data providing a better understanding of LLMs potential for clinical application as well as the ultimate need for further investigation of LLMs performance and security.

Future studies should aim to confirm/validate the presented results, at best with real-world patient data from routine sleep medicine practice in larger collectives while respecting the confidentiality of personal data and taking into account the limitations mentioned above. In addition, an interesting prospective research approach would be to find out whether the recommendation for a further medical measure by the SMS or the LLM was correct. Further, detailed subgroup analysis across diverse patient characteristics should be a matter of future approaches to enhance the understanding of the LLMs' generalizability, ideally including interpretability methods such as SHAP or LIME.

Conclusion

In the interpretation of sleep medicine data, there is a high level of agreement between the LLM ChatGPT o1 preview and the SMS regarding the identification of individual health risks and potentially existing comorbidities. Despite the reliance on fictitious data and the lack of external validation within this pilot study, LLMs could offer potential as a useful tool in daily sleep medicine practice to prevent or recognize unknown comorbidities. However, its tendency to over-recommend evaluations highlights the need for further validation with real-world data and clinician oversight to improve generalizability before clinical integration. As LLMs continue to evolve, their clinical integration into healthcare could redefine the approach to patient evaluation and risk stratification.

Abbreviations

LLM, Large language model; SMS, Sleep medicine specialist; AASM, American Academy of Sleep Medicine.

Data Sharing Statement

All data can be obtained from the Supplementary Material.

Ethics Approval and Consent to Participate

There is no need for any specific ethical approval in this kind of studies not involving patients, animals or cells.

Author Contributions

CS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing.

KB-H: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review and editing.

HG: Formal analysis, Investigation, Writing – original draft, Writing – review and editing.

JP: Formal analysis, Investigation, Writing – original draft, Writing – review and editing.

AB: Formal analysis, Investigation, Writing – original draft, Writing – review and editing.
CM: Formal analysis, Investigation, Writing – original draft, Writing – review and editing.
SK: Formal analysis, Investigation, Writing – original draft, Writing – review and editing.
CRB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing.
All authors agreed on the journal to which the article will be submitted, reviewed and agreed on all versions of the article before submission, during revision, the final version accepted for publication, and any significant changes introduced at the proofing stage, agrees to take responsibility and be accountable for the contents of the article.

Funding

No funding was received for this study.

Disclosure

SK is the founder and shareholder of MED digital. The other authors have no conflict of interest to declare.

References

1. Eckert DJ, Jordan AS, Merchia P, et al. Central sleep apnea: pathophysiology and treatment. *Chest*. 2007;131(2):595–607. doi:10.1378/chest.06.2287
2. Malhotra A, White DP. Obstructive sleep apnoea. *Lancet*. 2002;360(9328):237–245. doi:10.1016/S0140-6736(02)09464-3
3. Krishnan V, Dixon-Williams S, Thornton JD. Where there is smoke…there is sleep apnea: exploring the relationship between smoking and sleep apnea. *Chest*. 2014;146(6):1673–1680. doi:10.1378/chest.14-0772
4. Yang S, Guo X, Liu W, et al. Alcohol as an independent risk factor for obstructive sleep apnea. *Ir J Med Sci*. 2022;191(3):1325–1330. doi:10.1007/s11845-021-02671-7
5. Peppard PE, Young T, Palta M, et al. Prospective study of the association between sleep-disordered breathing and hypertension. *N Engl J Med*. 2000;342(19):1378–1384. doi:10.1056/NEJM200005113421901
6. Loke YK, Brown JW, Kwok CS, Niruban A, Myint PK. Association of obstructive sleep apnea with risk of serious cardiovascular events: a systematic review and meta-analysis. *Circ Cardiovasc Qual Outcomes*. 2012;5(5):720–728. doi:10.1161/CIRCOUTCOMES.111.964783
7. Marin JM, Carrizo SJ, Vicente E, et al. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. *Lancet*. 2005;365(9464):1046–1053. doi:10.1016/S0140-6736(05)71141-7
8. Seifen C, Pordzik J, Ludwig K, et al. Obstructive sleep apnea disrupts glycemic control in obese individuals. *Medicina*. 2022;58(11):1602. doi:10.3390/medicina58111602
9. Reutrakul S, Mokhlesi B. Obstructive sleep apnea and diabetes: a state of the art review. *Chest*. 2017;152(5):1070–1086. doi:10.1016/j.chest.2017.05.009
10. Börgel J, Sanner BM, Bittlinsky A, et al. Obstructive sleep apnoea and its therapy influence high-density lipoprotein cholesterol serum levels. *Eur Respir J*. 2006;27(1):121–127. doi:10.1183/09031936.06.00131304
11. Ahmed MH, Byrne CD. Obstructive sleep apnea syndrome and fatty liver: association or causal link? *World J Gastroenterol*. 2010;16(34):4243–4252. doi:10.3748/wjg.v16.i34.4243
12. Bahr K, Simon P, Leggewie B, et al. The snoring index identifies risk of non-alcoholic fatty liver disease in patients with obstructive sleep apnea syndrome. *Biology*. 2021;11(1):10. doi:10.3390/biology11010010
13. Gillum RF. From papyrus to the electronic tablet: a brief history of the clinical medical record with lessons for the digital age. *Am J Med*. 2013;126(10):853–857. doi:10.1016/j.amjmed.2013.03.024
14. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940. doi:10.1038/s41591-023-02448-8
15. Seifen C, Huppertz T, Gouveris H, et al. Chasing sleep physicians: chatGPT-4o on the interpretation of polysomnographic results. *Eur Arch Otorhinolaryngol*. 2024;282(3):1631–1639. doi:10.1007/s00405-024-08985-3
16. openai.com. Available from: https://openai.com/index/introducing-openai-o1-preview/. Accessed April 23, 2024.
17. Weissman GE, Mankowitz T, Kanter GP. Unregulated large language models produce medical device-like output. *NPJ Digit Med*. 2025;8(1):148. doi:10.1038/s41746-025-01544-y
18. European Commission. Regulation (EU) 2017/745 of the European parliament and of the council on medical devices. *Off J Eur Union*. 2017.
19. FDA. Policy for device software functions and mobile medical applications. Guidance for Industry and Food and Drug Administration Staff. 2022.
20. EU AI Act. Available from: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Accessed April 23, 2025.
21. Franklin KA, Lindberg E. Obstructive sleep apnea is a common disorder in the population-a review on the epidemiology of sleep apnea. *J Thorac Dis*. 2015;7(8):1311–1322. doi:10.3978/j.issn.2072-1439.2015.06.11
22. Heinzer R, Vat S, Marques-Vidal P, et al. Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study. *Lancet Respir Med*. 2015;3(4):310–318. doi:10.1016/S2213-2600(15)00043-0
23. Goldstein CA, Berry RB, Kent DT, et al. Artificial intelligence in sleep medicine: an American academy of sleep medicine position statement. *J Clin Sleep Med*. 2020;16(4):605–607. doi:10.5664/jcsm.8288
24. BaHammam AS. Artificial intelligence in sleep medicine: the dawn of a new era. *Nat Sci Sleep*. 2024;16:445–450. doi:10.2147/NSS.S474510

25. Bonsignore MR, Baiamonte P, Mazzuca E, Castrogiovanni A, Marrone O. Obstructive sleep apnea and comorbidities: a dangerous liaison. *Multidiscip Respir Med*. 2019;14:8.
26. Sateia MJ. International classification of sleep disorders-third edition. *Chest*. 2014;146(5):1387–1394. doi:10.1378/chest.14-0970
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174. (). doi:10.2307/2529310
28. Gleeson M, McNicholas WT. Bidirectional relationships of comorbidity with obstructive sleep apnoea. *Eur Respir Rev*. 2022;31(164):210256. doi:10.1183/16000617.0256-2021
29. McNicholas WT. Obstructive sleep apnoea and comorbidity – an overview of the association and impact of continuous positive airway pressure therapy. *Expert Rev Respir Med*. 2019;13(3):251–261. doi:10.1080/17476348.2019.1575204
30. Sircu V, Colesnic S-I, Covantsev S, et al. The burden of comorbidities in obstructive sleep apnea and the pathophysiologic mechanisms and effects of CPAP. *Clocks Sleep*. 2023;5(2):333–349. doi:10.3390/clockssleep5020025
31. Nieden PBZ. The economic burden of (obstructive) sleep apnea. costs and implications for Germany based on the results of an international systematic review. *J Public Health*. 2024. doi:10.1007/s10389-024-02269-0
32. Eckert DJ, White DP, Jordan AS, et al. Defining phenotypic causes of obstructive sleep apnea identification of novel therapeutic targets. *Am J Respir Crit Care Med*. 2013;188(8):996–1004. doi:10.1164/rccm.201303-0448OC
33. Ennab M, Mcheick H. Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. *Front Robot AI*. 2024;11:1444763. doi:10.3389/frobt.2024.1444763
34. Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' perceptions toward human-artificial intelligence interaction in health care: experimental study. *J Med Internet Res*. 2021;23(11):e25856. PMID: 34842535; PMCID: PMC8663518. doi:10.2196/25856
35. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. *Nat Med*. 2024;30(11):3098–3100. doi:10.1038/s41591-024-03180-7
36. Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. In: The Proceedings of the ACM on Human Computer Interaction (HCI); 2021.