Nature and Science of Sleep

ORIGINAL RESEARCH

Evaluating Locally Run Large Language Models for Obstructive Sleep Apnea Diagnosis and Treatment: A Real-World Polysomnography Study

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

[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, Philipps University Marburg, University Hospital Giessen and Marburg, Marburg, Germany

Correspondence: Christoph Raphael Buhr, Sleep Medicine Center & Department of Otolaryngology, Head and Neck Surgery, University Medical Center Mainz, Mainz, 55131, Germany, Email christophraphael.buhr@unimedizin-mainz.de

Purpose: Sleep medicine is a highly resource-intensive field where large language models (LLMs) could offer a promising solution by supporting diagnostic processes. As web-based LLMs have obvious data protection constraints, locally run LLMs are essential for clinical implementation. This study is the first to investigate the performance of locally run LLMs in the interpretation of real-world polysomnographic (PSG) results.

Methods: We randomly selected N=30 patients (18 male, 12 female, mean age 50.5 ± 11.1 years, mean body mass index 29.7 ± 5.5 kg/m², mean apnea hypopnea index 30.9 ± 23.8) from the clinical database of our sleep laboratory who underwent PSG due to clinical complaints typical of obstructive sleep apnea (OSA). The board-certified sleep physician's interpretations of diagnosis, suitable first-line therapy or alternative therapy were compared with those of three locally run LLMs (Gemma2, Llama3 and Mistral Nemo) assessing the level of concordance.

Results: Gemma2 showed the lowest concordance of 33% (10/30 patients) with the board-certified sleep physician regarding OSA severity, followed by Mistral Nemo at 47% (14/30 patients) and Llama3 at 50% (15/30 patients). For automatic positive airway pressure (aPAP) recommendations, Mistral Nemo showed the highest concordance at 90% (27/30 patients), followed by Gemma2 and Llama3 with 83% (25/30 patients) each.

Conclusion: Although locally run LLMs bypass data security constraints and show promising potential for clinical practice, their performance needs significant improvement prior to real-world implementation. Therefore, at present, the routine implementation of locally run LLMs in sleep medicine needs more refinement and fine tuning before they can be used for interpretation of real-world PSG results.

Keywords: large language model, LLM, polysomnography, PSG, obstructive sleep apnea, OSA, sleep medicine, digital health

Introduction

Sleep medicine is a resource intensive field, particularly as the prevalence of sleep-related breathing disorders (eg obstructive sleep apnea, OSA) increases worldwide.[1,2] Data obtained from standard diagnostic procedures in sleep medicine practice, such as polysomnography (PSG), require processing and analyzing to make a diagnosis and a management plan. Given ongoing shortages in clinical personnel, these demands are challenging. Use of large language models (LLMs) has recently been tested in sleep medicine and might provide a promising approach to interpreting PSG data.[3–5]

In contrast to other natural language processing models, LLMs are trained on large, heterogeneous data sets, creating generalists, rather than narrow-task specialists. Once trained, a general-purpose model can be used for a wide variety of different purposes in a cost-efficient manner.[6] Despite the lack of specialization, LLMs showed promising performance in processing and analysing medical data in various specialties: a recent study demonstrated the effectiveness of Chat Generative Pre-Trained Transformer (ChatGPT) in analyzing polysomnographic data and offering suitable diagnosis and

Received: 6 May 2025
Accepted: 23 June 2025
Published: 8 July 2025

therapy recommendations for synthetic cases. Here, the concordance between the tested version ChatGPT-4o and the sleep physician reached 97% for diagnosis and 100% for therapy recommendation of simple cases, whereas for more complex cases with automatic positive airway pressure (PAP)-intolerance a concordance of 70% for diagnosis and 22% for therapy suggestions was found.[7] Despite these promising findings, the use of synthetic cases due to obvious data protection constraints highlights a key limitation of that study. The clinical consequences of misclassification can be severe, leading to under-treatment of severe OSA or over-treatment (eg with PAP therapy) when unnecessary.

Particularly in environments with strict data protection regulations, the clinical use of LLMs is only conceivable using locally run LLMs without external data transfer. The present study thus focuses on the use of locally run LLMs for analysis of real-world PSG patient data instead of synthetic or simulated patient data used in previous studies to comply with local data protection guidelines. It is the first study to analyze real-world PSG results using three different locally run LLMs (Gemma 2, Llama 3, Mistral Nemo) while benchmarking their diagnosis and treatment recommendations against a board-certified sleep physician.

Materials and Methods

Study Design and Workflow

We investigated the clinical database of our sleep laboratory (which is part of a university medical center) for N=30 patients who underwent first-time full-night PSG. All patients were admitted due to clinical complaints typical of OSA (eg snoring, apneas, or daytime sleepiness). A specifically trained technician performed the PSG overnight. Each PSG was analyzed and interpreted by a board-certified sleep physician based on the American Academy of Sleep Medicine (AASM) standard guidelines.[8]

All polysomnographic results were presented in a structured tabular format as a pdf file providing the following details:

- Name and surname of the patient,
- Date of birth,
- Height (cm),
- Weight (kg),
- Patients sex,
- Body mass index (BMI) of the patient,
- Begin of the PSG,
- End of the PSG,
- Measurement duration of the PSG,
- Lights out and lights on,
- Evaluation period and evaluation duration,
- Artefact for flow and SpO2 in percentage,
- Apnea hypopnea index (AHI) per hour,
- Respiratory distress index (RDI) per hour,
- Apnea index (AI) per hour,
- Hypopnea index (HI) per hour,
- Total central sleep apnea time (CSA) in seconds,
- Number of CSA phases,
- AHI in rapid eye movement (REM) per hour,
- Number of apneas,
- Number of central apneas,
- Number of obstructive apneas,
- Number of mixed apneas,
- Number of hypopneas,
- Number of central hypopneas,

- Number obstructive hypopneas,
- Number of unclassified hypopneas,
- Number of respiratory effort-related arousals (RERA),
- Number of apneas / hypopneas in REM,
- Total apnea / hypopnea time in minutes,
- Longest apnea in minutes,
- Longest hypopnea in minutes,
- Snoring index (SI) per hour,
- SI in REM per hour,
- Irregular SI per hour,
- Total snoring time in hours,
- Maximum heart rate,
- Mean heart rate,
- Minimal heart rate,
- Number of bradycardias,
- Number of tachycardias,
- Number of extrasystoles,
- Oxygen desaturation index (ODI) per hour,
- ODI in REM per hour,
- Number of desaturations,
- Number of desaturations in REM,
- Number of desaturations under 90%,
- Total desaturation time in hours,
- Desaturation time per hour in minutes,
- Lowest desaturation in percentage,
- Longest desaturation in minutes,
- Mean desaturation duration in seconds,
- Mean desaturation in percentage,
- Mean saturation in percentage,
- Maximum saturation in percentage,
- Minimal saturation in percentage,
- Saturation under 90% (t90) in percentage,
- Minimal pulse per minute,
- Maximum pulse per minute,
- Mean pulse per minute,
- Number of pulse variances,
- Pulse variance index per hour,
- Number of limb movements (LM),
- LM index (LMI) per hour,
- Number of non-periodic limb movements (PLM),
- Non-PLM index per hour,
- Number of PLMs,
- PLM index (PLMI) per hour,
- Graphic representation of respiratory events,
- Graphic representation of pulse distribution,
- Graphic representation of SpO2 distribution,
- Graphic representation of sleep profile (hypnogram),
- Sleep onset (SO) latency in minutes,

https://doi.org/10.2147/NSS.S536823

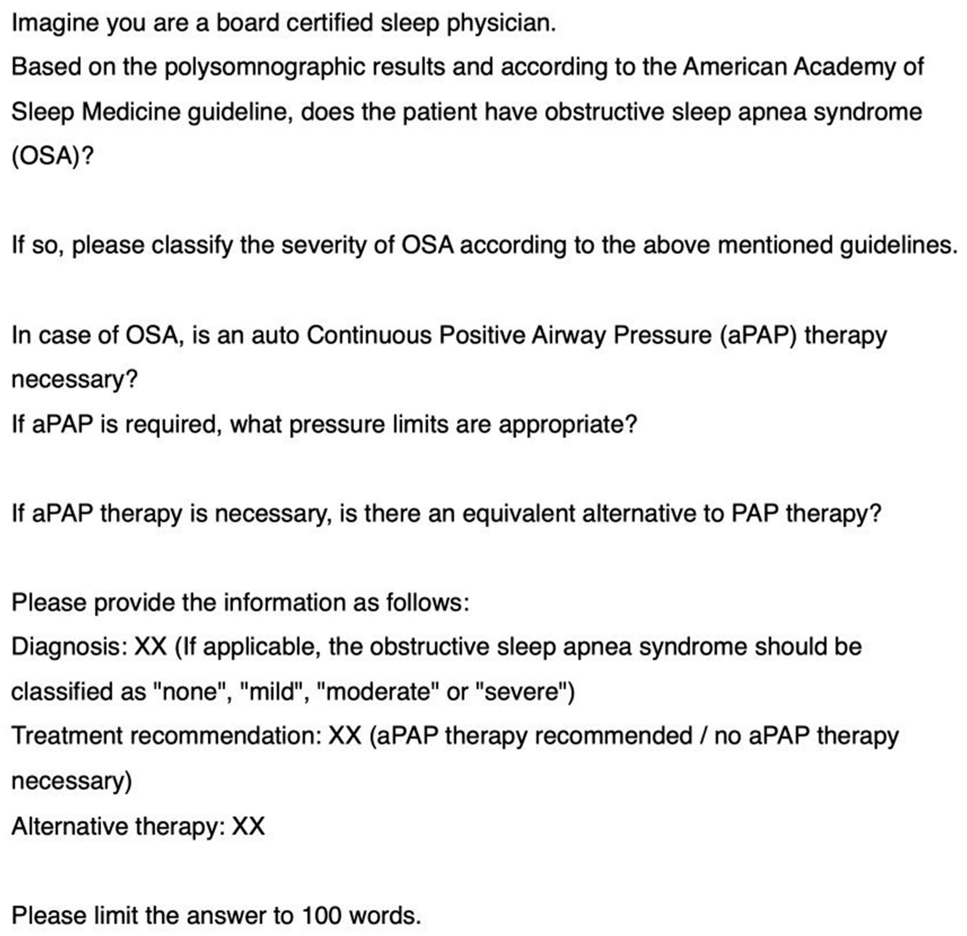
- N1/N2/N3-latency from lights off in minutes,
- REM-latency from SO in minutes,
- Latency to onset of persistent sleep in minutes,
- Time in bed (TIB) in minutes,
- Sleep period time (SPT) in minutes,
- Total sleep time (TST) in minutes,
- Sleep efficacy (TST/TIB) in percentage,
- Sleep efficacy (TST/SPT) in percentage,
- Distribution of sleep stages N1/N2/N3/REM/awake in TIB/SPT/TST in minutes and percentage,
- Duration of REM cycles in minutes,
- Number of arousals for TIB and TST,
- Number of respiratory/motoric/PLM-related/spontaneous arousals for TIB and TST,
- Arousal index per hour for TIB and TST,
- Respiratory/motoric/PLM-related/spontaneous arousal-index for TIB and TST,
- Number of events during sleep stages N1/N2/N3/non-REM/REM/awake/artefact, including apneas, hypopneas, the AHI, the RDI, the ODI, the PLMI, the LMI and the SI,
- Distribution of body positions upright/right/back/left/abdomen/total in seconds/minutes/hours,
- Number of apnea/hypopneas during above-mentioned body positions,
- AHI and RDI during above-mentioned body positions,
- Ratio of AHI and RDI during back position vs non-back position,
- Graphic representation of body positions and ultimately
- An interpretation of the polysomnographic result by the board-certified sleep physician including a diagnosis (no/mild/moderate/severe OSA) and a suitable form of therapy or alternative therapy.

To ensure an unbiased interpretation of all polysomnographic results by the locally run LLMs, the board-certified sleep physician's interpretation was obscured with preview software integrated in Mac OS. The edited pdf files were then gradually presented to the locally run LLMs. The LLMs were run on a standard Dell Laptop (Intel Core i5-1225U, 12 CPUs, 1,30GHz, 16GB RAM, Intel UHD Graphics, Windows 10 64 Bit, LMStudio, 0.3.5). The LLMs were selected based on their availability as open-source models and their size. Therefore, models of up to 5 GB in size were used. Specifically, the following models were evaluated within our study: Meta-Llama-3-8B-Instruct-GGUF, Gemma-2-9b-it-GGUF and Mistral-Nemo-Instruct-2407-GGUF (all published by lmstudio-community).

Figure 1 shows the prompt that was used in each locally run LLM.

After each of the three locally run LLMs analyzed all N=30 pdf files, a comparison was made with the diagnosis and therapy recommendations of the board-certified sleep physician to evaluate the level of concordance. The correctness of answers was assessed as correct/incorrect, ie binary. Partial correctness was not assessed. Moreover, sensitivity (SEN: = correctly recognized as sick / all sick people), specificity (SPE = correctly identified as healthy / all healthy people), positive predictive value (PPV = number of true positives/(number of true positives + number of false positives) and the negative predictive value (NPV = number of true negatives/(number of true negatives + number of false negatives)) were determined. Figure 2 summarizes the workflow of the study.

Since the concordance between the locally run LLMs and the board-certified sleep physician differed substantially from the performance of web-based LLMs known from the literature, we extracted a sample of n=10 patients (patients 1, 2, 4, 7, 9, 10, 15, 16, 19, 28) in order to improve the performance of the locally run LLMs using different strategies. First, the locally run LLMs were provided a definition of OSA severity within the prompt (AHI 0–5/h none; AHI >5-15/h mild; AHI >15-30/h moderate and AHI >30/h severe). As a second optimization strategy, the most important polysomnography information was extracted and presented within the prompt, rather than providing the full polysomnographic pdf file. In detail, the following data were provided in tabular shape at the beginning of each prompt: sex, age, body mass index, apnea hypopnea index, apnea index, hypopnea index, cumulative apnea and hypopnea duration, oxygen desaturation index, average oxygen saturation, t90, total sleep time, sleep efficacy and ratio of supine to non-supine apnea-hypopnea index.



Imagine you are a board certified sleep physician.
Based on the polysomnographic results and according to the American Academy of Sleep Medicine guideline, does the patient have obstructive sleep apnea syndrome (OSA)?

If so, please classify the severity of OSA according to the above mentioned guidelines.

In case of OSA, is an auto Continuous Positive Airway Pressure (aPAP) therapy necessary?
If aPAP is required, what pressure limits are appropriate?

If aPAP therapy is necessary, is there an equivalent alternative to PAP therapy?

Please provide the information as follows:
Diagnosis: XX (If applicable, the obstructive sleep apnea syndrome should be classified as "none", "mild", "moderate" or "severe")
Treatment recommendation: XX (aPAP therapy recommended / no aPAP therapy necessary)
Alternative therapy: XX

Please limit the answer to 100 words.

Figure 1 The standardized prompt used for each locally run LLM.

Ethical Statement

The study protocol was approved by the ethics commission of the Rhineland-Palatinate State Medical Association (reference no: 2023–17,391). The ethics commission did not require separate informed consent from the study participants because of the retrospective character of the study. In addition, the study protocol was discussed with the university medical center's local data protection officer. To protect patient privacy, all data was analyzed exclusively on an institutional computer with no external connections. No real-world patient data was transmitted to an internet-connected applications or third parties. The study was conducted according to the Declaration of Helsinki.

Results

Study Population Characteristics

We included N=30 patients in the present study of whom 18 (60%) were male and 12 (40%) were female. The mean age of the study population was 50.5 ± 11.1 years, the mean BMI was 29.7 ± 5.5 kg/m² and the mean AHI was 30.9 ± 23.8. After categorizing the study population according to OSA severity based on the AHI, n=1 (3%) had no OSA, n=7 (23%) had mild OSA, n=11 (37%) had moderate OSA and n=11 (37%) had severe OSA.

Diagnosis and Therapy Recommendations

Figure 3 shows the diagnosis and therapy recommendations as stated by the board-certified sleep physician and the locally run LLMs (Gemma 2, Llama 3 and Mistral Nemo). While the board-certified sleep physician classified the OSA severity as noted above, Gemma 2 diagnosed 3 patients (10%) with mild and 27 patients (90%) with moderate OSA. Llama 3 diagnosed 2

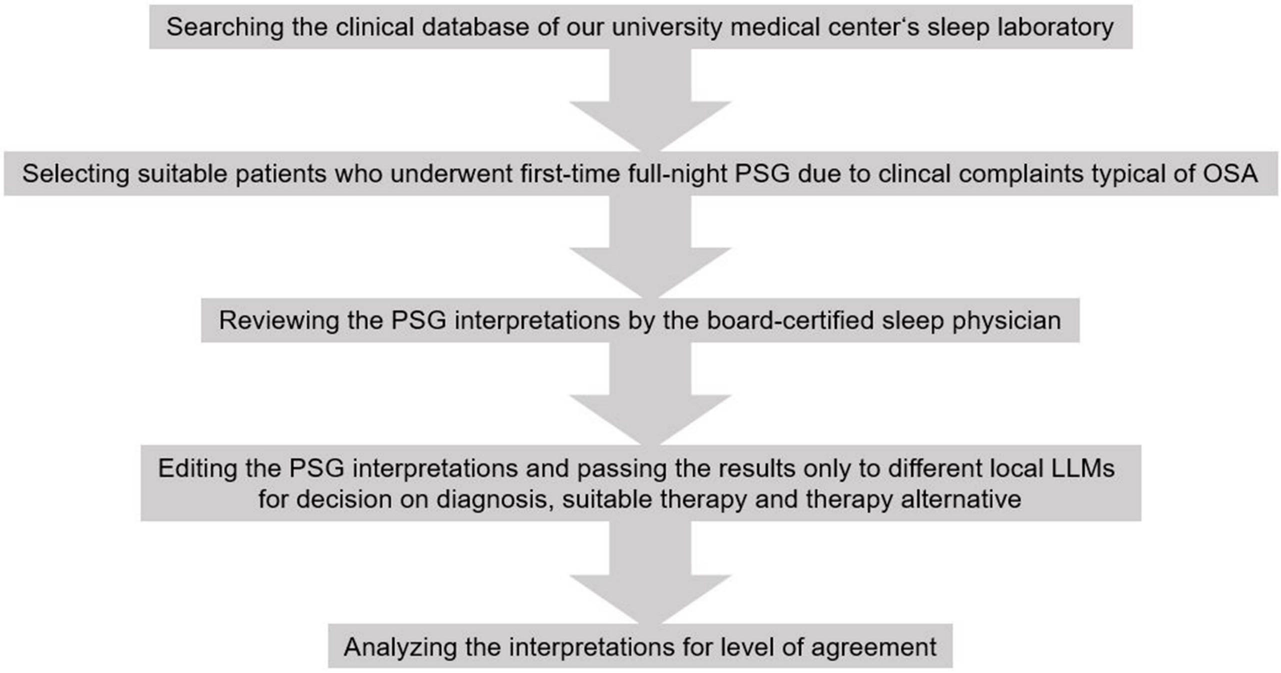

Searching the clinical database of our university medical center‘s sleep laboratory
Selecting suitable patients who underwent first-time full-night PSG due to clincal complaints typical of OSA
Reviewing the PSG interpretations by the board-certified sleep physician
Editing the PSG interpretations and passing the results only to different local LLMs for decision on diagnosis, suitable therapy and therapy alternative
Analyzing the interpretations for level of agreement

Figure 2 Workflow of the study.
Abbreviations: OSA, obstructive sleep apnea; PSG, polysomnography; LLM, large language model.

patients with mild (7%), 11 patients with moderate (37%) and 17 patients (57%) with severe OSA, whereas Mistral Nemo diagnosed 1 patient with no (3%), 3 patient with mild (10%), 5 patients with moderate (17%) and 21 patients (70%) with severe OSA. The board-certified sleep physician recommended aPAP therapy for 25 patients (83%) whereas Gemma 2 and Llama 3 recommended aPAP therapy for all 30 patients (100%) and Mistral Nemo for 28 patients (93%). Whereas the physician recommended weight loss as additional therapy to aPAP in 7 patients (23%), none of the LLMs stated this additional therapy.

Regarding alternative therapy, Gemma 2 recommended oral appliance for every patient (100%). Mistral Nemo showed a similar high recommendation ratio for oral appliance as alternative therapy, stating it for all 28 patients (93%) with aPAP recommendation. Llama 3 suggested oral appliance as alternative for 20 patients (67%). In contrast, the board-certified sleep physician recommended oral appliances for just a single patient (3%). While only Llama 3 and Mistral Nemo recommended lifestyle change as an alternative therapy, the board-certified sleep physician and Llama 3 represent the only entities, considering surgery for therapy as alternative therapy of OSA.

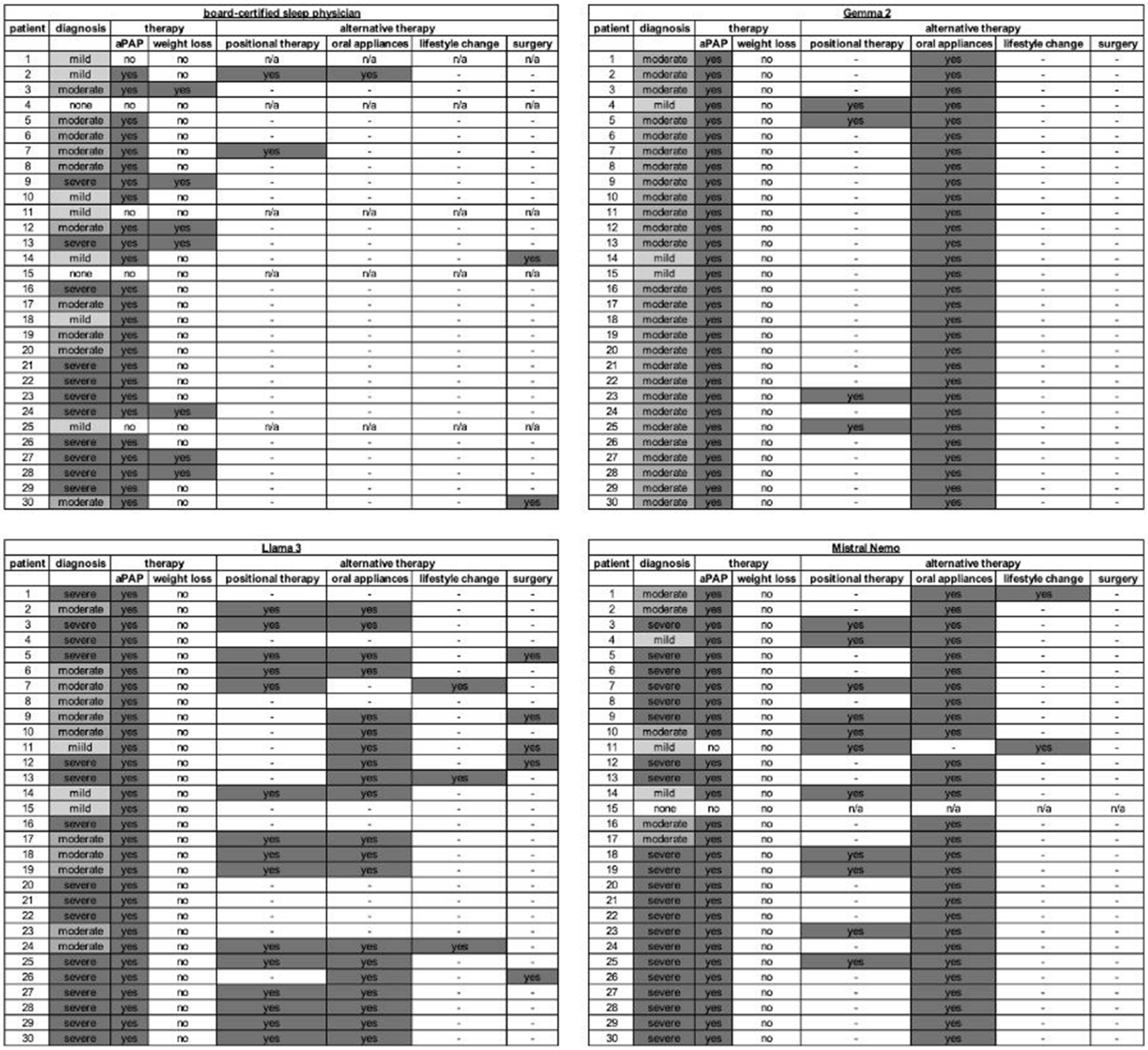
Concordance Between Board-Certified Sleep Physician and the Locally Run LLMs

The concordance between the board-certified sleep physician and the different locally run LLMs is visualized in Figure 4 for Gemma 2, Figure 5 for Llama and Figure 6 for Mistral Nemo. Gemma 2 showed the lowest concordance with the board-certified sleep physician for the severity of OSA (33%=10/30 patients), followed by Mistral Nemo (47%=14/30 patients) and Llama 3 (50%=15/30 patients). In contrast, regarding the recommendation for aPAP Mistral Nemo showed the highest concordance with the board-certified sleep physician achieving 90% (27/30 patients) and followed by Gemma 2 and Llama 3 with 83% (25/30 patients) each.

Discussion

Despite their promising performance, web-based LLMs have significant data protection concerns for implementation in clinical practice.[7,9] However, studies addressing the use of locally run LLMs that minimize risks of data sharing, for the interpretation of real-world polysomnographic patient data are lacking.

The data presented in this study reveals some shortcomings of locally run LLMs in classifying the OSA severity. Gemma 2 achieved only 33% (10/30 patients) concordance with the board-certified sleep physician, Mistral Nemo 47%



board-certified sleep physician

patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
1	mild	no	no	n/a	n/a	n/a	n/a
2	mild	yes	no	yes	yes	-	-
3	moderate	yes	yes	-	-	-	-
4	none	no	no	n/a	n/a	n/a	n/a
5	moderate	yes	no	-	-	-	-
6	moderate	yes	no	-	-	-	-
7	moderate	yes	no	yes	-	-	-
8	moderate	yes	no	-	-	-	-
9	severe	yes	yes	-	-	-	-
10	mild	yes	no	-	-	-	-
11	mild	no	no	n/a	n/a	n/a	n/a
12	moderate	yes	yes	-	-	-	-
13	severe	yes	yes	-	-	-	-
14	mild	yes	no	-	-	-	yes
15	none	no	no	n/a	n/a	n/a	n/a
16	severe	yes	no	-	-	-	-
17	moderate	yes	no	-	-	-	-
18	mild	yes	no	-	-	-	-
19	moderate	yes	no	-	-	-	-
20	moderate	yes	no	-	-	-	-
21	severe	yes	no	-	-	-	-
22	severe	yes	no	-	-	-	-
23	severe	yes	no	-	-	-	-
24	severe	yes	yes	-	-	-	-
25	mild	no	no	n/a	n/a	n/a	n/a
26	severe	yes	no	-	-	-	-
27	severe	yes	yes	-	-	-	-
28	severe	yes	yes	-	-	-	-
29	severe	yes	no	-	-	-	-
30	moderate	yes	no	-	-	-	yes

Gemma 2

patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
1	moderate	yes	no	-	yes	-	-
2	moderate	yes	no	-	yes	-	-
3	moderate	yes	no	-	yes	-	-
4	mild	yes	no	yes	yes	-	-
5	moderate	yes	no	yes	yes	-	-
6	moderate	yes	no	-	yes	-	-
7	moderate	yes	no	-	yes	-	-
8	moderate	yes	no	-	yes	-	-
9	moderate	yes	no	-	yes	-	-
10	moderate	yes	no	-	yes	-	-
11	moderate	yes	no	-	yes	-	-
12	moderate	yes	no	-	yes	-	-
13	moderate	yes	no	-	yes	-	-
14	mild	yes	no	-	yes	-	-
15	mild	yes	no	-	yes	-	-
16	moderate	yes	no	-	yes	-	-
17	moderate	yes	no	-	yes	-	-
18	moderate	yes	no	-	yes	-	-
19	moderate	yes	no	-	yes	-	-
20	moderate	yes	no	-	yes	-	-
21	moderate	yes	no	-	yes	-	-
22	moderate	yes	no	-	yes	-	-
23	moderate	yes	no	yes	yes	-	-
24	moderate	yes	no	-	yes	-	-
25	moderate	yes	no	yes	yes	-	-
26	moderate	yes	no	-	yes	-	-
27	moderate	yes	no	-	yes	-	-
28	moderate	yes	no	-	yes	-	-
29	moderate	yes	no	-	yes	-	-
30	moderate	yes	no	-	yes	-	-

Llama 3

patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
1	severe	yes	no	-	-	-	-
2	moderate	yes	no	yes	yes	-	-
3	severe	yes	no	yes	yes	-	-
4	severe	yes	no	-	-	-	-
5	severe	yes	no	yes	yes	-	yes
6	moderate	yes	no	yes	yes	-	-
7	moderate	yes	no	yes	-	yes	-
8	moderate	yes	no	-	-	-	-
9	moderate	yes	no	-	yes	-	yes
10	moderate	yes	no	-	yes	-	-
11	miild	yes	no	-	yes	-	yes
12	severe	yes	no	-	yes	-	yes
13	severe	yes	no	-	yes	yes	-
14	mild	yes	no	yes	yes	-	-
15	mild	yes	no	-	-	-	-
16	severe	yes	no	-	-	-	-
17	moderate	yes	no	yes	yes	-	-
18	moderate	yes	no	yes	yes	-	-
19	moderate	yes	no	yes	yes	-	-
20	severe	yes	no	-	-	-	-
21	severe	yes	no	-	-	-	-
22	severe	yes	no	-	-	-	-
23	moderate	yes	no	-	-	-	-
24	moderate	yes	no	yes	yes	yes	-
25	severe	yes	no	yes	yes	-	-
26	severe	yes	no	-	yes	-	yes
27	severe	yes	no	yes	yes	-	-
28	severe	yes	no	yes	yes	-	-
29	severe	yes	no	yes	yes	-	-
30	severe	yes	no	yes	yes	-	-

Mistral Nemo

patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
1	moderate	yes	no	-	yes	yes	-
2	moderate	yes	no	-	yes	-	-
3	severe	yes	no	yes	yes	-	-
4	mild	yes	no	yes	yes	-	-
5	severe	yes	no	-	yes	-	-
6	severe	yes	no	-	yes	-	-
7	severe	yes	no	yes	yes	-	-
8	severe	yes	no	-	yes	-	-
9	severe	yes	no	yes	yes	-	-
10	moderate	yes	no	yes	yes	-	-
11	mild	no	no	yes	-	yes	-
12	severe	yes	no	-	yes	-	-
13	severe	yes	no	-	yes	-	-
14	mild	yes	no	yes	yes	-	-
15	none	no	no	n/a	n/a	n/a	n/a
16	moderate	yes	no	-	yes	-	-
17	moderate	yes	no	-	yes	-	-
18	severe	yes	no	yes	yes	-	-
19	severe	yes	no	yes	yes	-	-
20	severe	yes	no	-	yes	-	-
21	severe	yes	no	-	yes	-	-
22	severe	yes	no	-	yes	-	-
23	severe	yes	no	yes	yes	-	-
24	severe	yes	no	-	yes	-	-
25	severe	yes	no	yes	yes	-	-
26	severe	yes	no	-	yes	-	-
27	severe	yes	no	-	yes	-	-
28	severe	yes	no	-	yes	-	-
29	severe	yes	no	-	yes	-	-
30	severe	yes	no	-	yes	-	-

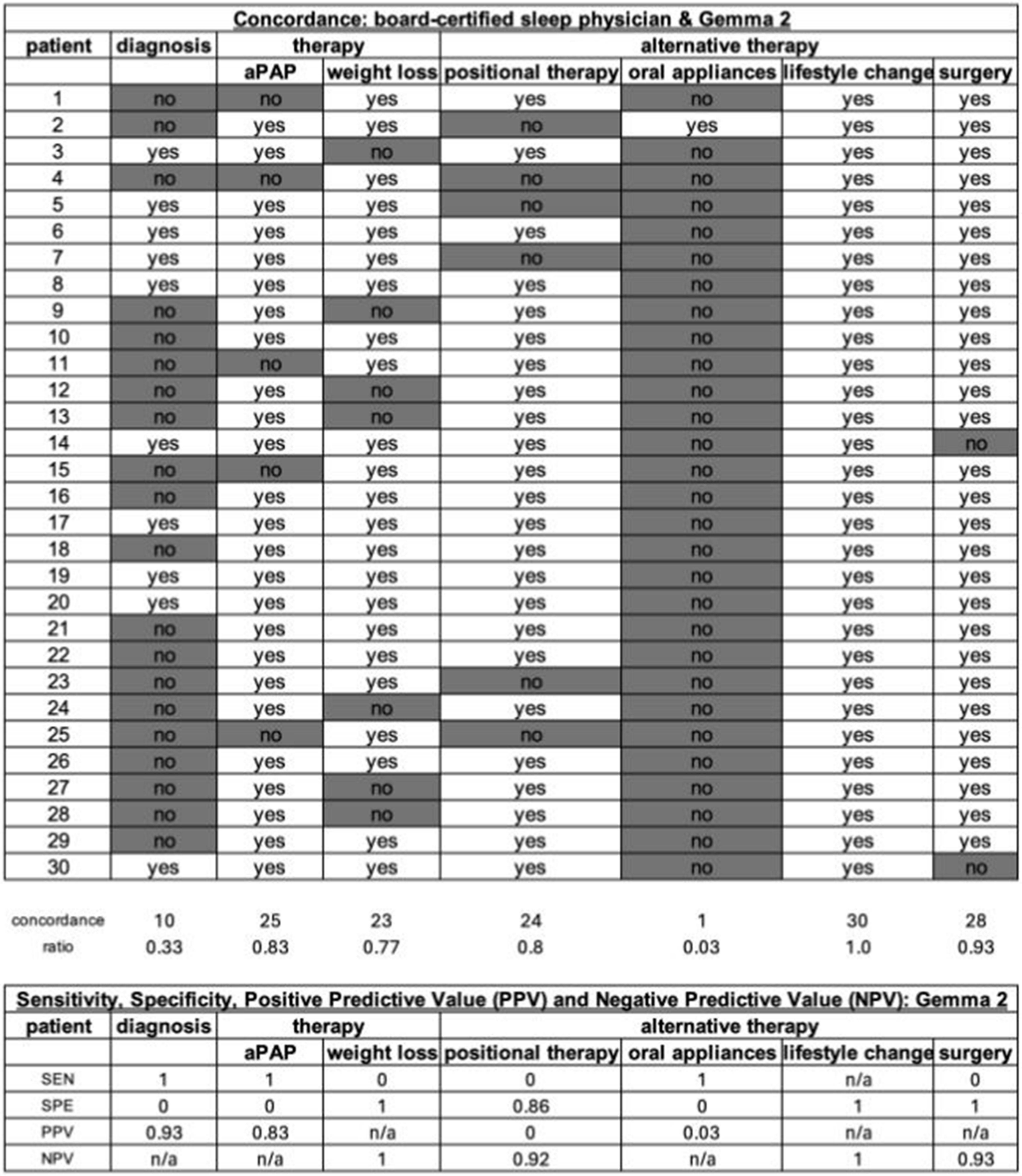
Figure 3 Diagnosis and therapy recommendations from the board-certified sleep physician and each locally run LLMs (Gemma 2, Llama 3 and Mistral Nemo).

(14/40 patients) and Llama 3 50% (15/30 patients). Nonetheless, the LLMs performance on indicating aPAP therapy was notably higher. Here, Mistral Nemo reached 90% (27/30 patients) concordance, Llama 3 and Gemma 2 83% (25/30 patients) each. Although Llama 3 and Gemma 2 recommended aPAP therapy for all patients, Mistral Nemo contra-indicated aPAP in one patient without OSA (AHI < 5/h) (Figure 6, patient 4). The performance of the LLM could be attributed to computational constraints and a lack of domain-specific fine-tuning associated with local deployment. However, the data demonstrates the general ability of locally run LLMs to reject treatment despite a strong tendency towards hedging of LLMs for security reasons, especially regarding health care queries.

Previous work evaluating the web-based ChatGPT-4o showed a concordance of 97% with a sleep physician for diagnosis and 100% for therapy recommendations in simple constructed cases. On more complex cases with PAP-intolerance the model still reached 70% concordance with the sleep physician for diagnosis and 22% for therapy suggestions.[7]

Due to the substantial performance gap compared to web-based LLM and the obvious problems of the locally run LLMs in classifying the OSA severity, we integrated a definition of OSA severity within the prompt. In a subset of 10 patients, Gemma 2's concordance remained at 30% (3/10), while Llama 3's was 50% (5/10). Mistral Nemo, however, improved to 70% (7/10). The recommendation regarding aPAP therapy remained unchanged across all three LLMs.

https://doi.org/10.2147/NSS.S536823



Concordance: board-certified sleep physician & Gemma 2

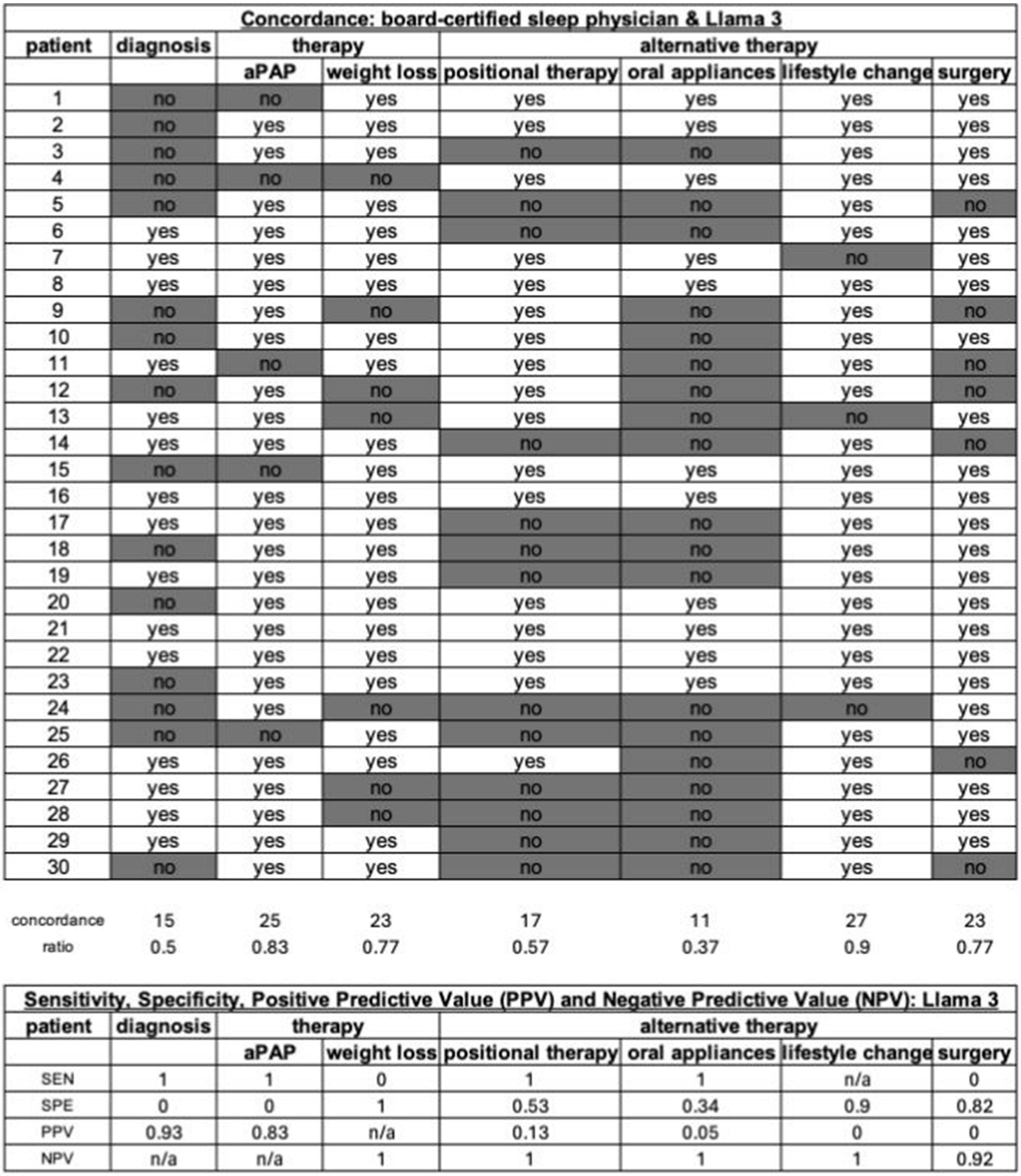
patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
1	no	no	yes	yes	no	yes	yes
2	no	yes	yes	no	yes	yes	yes
3	yes	yes	no	yes	no	yes	yes
4	no	no	yes	no	no	yes	yes
5	yes	yes	yes	no	no	yes	yes
6	yes	yes	yes	yes	no	yes	yes
7	yes	yes	yes	no	no	yes	yes
8	yes	yes	yes	yes	no	yes	yes
9	no	yes	no	yes	no	yes	yes
10	no	yes	yes	yes	no	yes	yes
11	no	no	yes	yes	no	yes	yes
12	no	yes	no	yes	no	yes	yes
13	no	yes	no	yes	no	yes	yes
14	yes	yes	yes	yes	no	yes	no
15	no	no	yes	yes	no	yes	yes
16	no	yes	yes	yes	no	yes	yes
17	yes	yes	yes	yes	no	yes	yes
18	no	yes	yes	yes	no	yes	yes
19	yes	yes	yes	yes	no	yes	yes
20	yes	yes	yes	yes	no	yes	yes
21	no	yes	yes	yes	no	yes	yes
22	no	yes	yes	yes	no	yes	yes
23	no	yes	yes	no	no	yes	yes
24	no	yes	no	yes	no	yes	yes
25	no	no	yes	no	no	yes	yes
26	no	yes	yes	yes	no	yes	yes
27	no	yes	no	yes	no	yes	yes
28	no	yes	no	yes	no	yes	yes
29	no	yes	yes	yes	no	yes	yes
30	yes	yes	yes	yes	no	yes	no
concordance	10	25	23	24	1	30	28
ratio	0.33	0.83	0.77	0.8	0.03	1.0	0.93

Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV): Gemma 2

patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
SEN	1	1	0	0	1	n/a	0
SPE	0	0	1	0.86	0	1	1
PPV	0.93	0.83	n/a	0	0.03	n/a	n/a
NPV	n/a	n/a	1	0.92	n/a	1	0.93

Figure 4 Concordance between the board-certified sleep physician and the locally run Gemma 2 LLM (absolute values/ratios per 30 patients). Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) as compared to the gold standard (board-certified sleep physician). n/a for values that cannot be calculated (division by zero) ie Gemma 2 diagnosed OSA in every patient.

In order to verify that the data presentation in the shape of a pdf file does not cause an inferior performance, we extracted the most important data from the PSG results and provided them in a tabular shape at the beginning of each prompt. However, this maneuver did not influence the results (data not shown). Accordingly, the inferior performance



Concordance: board-certified sleep physician & Llama 3

patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
1	no	no	yes	yes	yes	yes	yes
2	no	yes	yes	yes	yes	yes	yes
3	no	yes	yes	no	no	yes	yes
4	no	no	no	yes	yes	yes	yes
5	no	yes	yes	no	no	yes	no
6	yes	yes	yes	no	no	yes	yes
7	yes	yes	yes	yes	yes	no	yes
8	yes	yes	yes	yes	yes	yes	yes
9	no	yes	no	yes	no	yes	no
10	no	yes	yes	yes	no	yes	yes
11	yes	no	yes	yes	no	yes	no
12	no	yes	no	yes	no	yes	no
13	yes	yes	no	yes	no	no	yes
14	yes	yes	yes	no	no	yes	no
15	no	no	yes	yes	yes	yes	yes
16	yes	yes	yes	yes	yes	yes	yes
17	yes	yes	yes	no	no	yes	yes
18	no	yes	yes	no	no	yes	yes
19	yes	yes	yes	no	no	yes	yes
20	no	yes	yes	yes	yes	yes	yes
21	yes	yes	yes	yes	yes	yes	yes
22	yes	yes	yes	yes	yes	yes	yes
23	no	yes	yes	yes	yes	yes	yes
24	no	yes	no	no	no	no	yes
25	no	no	yes	no	no	yes	yes
26	yes	yes	yes	yes	no	yes	no
27	yes	yes	no	no	no	yes	yes
28	yes	yes	no	no	no	yes	yes
29	yes	yes	yes	no	no	yes	yes
30	no	yes	yes	no	no	yes	no
concordance	15	25	23	17	11	27	23
ratio	0.5	0.83	0.77	0.57	0.37	0.9	0.77

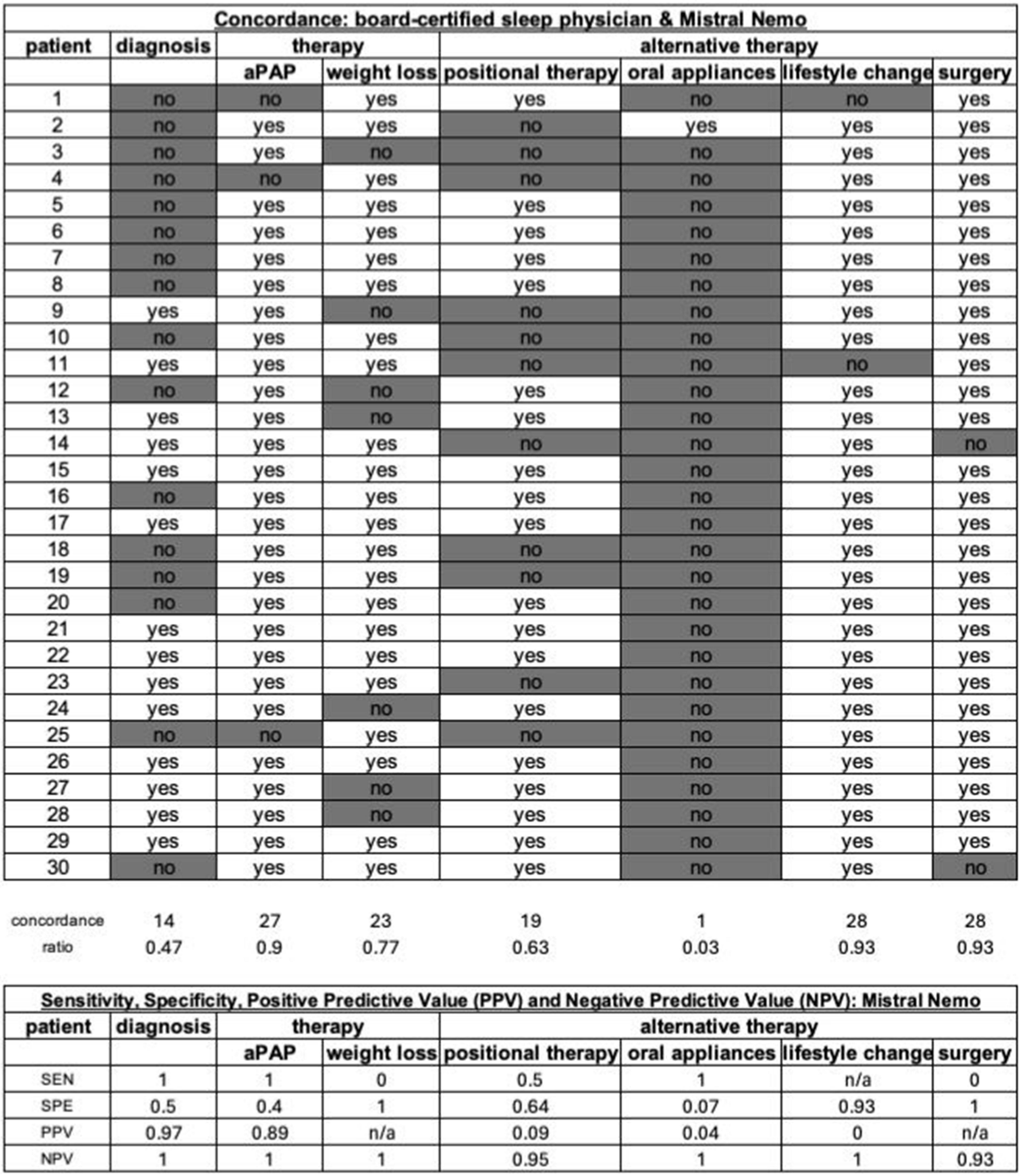
Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV): Llama 3

patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
SEN	1	1	0	1	1	n/a	0
SPE	0	0	1	0.53	0.34	0.9	0.82
PPV	0.93	0.83	n/a	0.13	0.05	0	0
NPV	n/a	n/a	1	1	1	1	0.92

Figure 5 Concordance between the board-certified sleep physician and the locally run Llama 3 LLM (absolute values/ratios per 30 patients). Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) as compared to the gold standard (board-certified sleep physician). n/a for values that cannot be calculated (division by zero) ie Llama 3 diagnosed OSA in every patient.

might be caused by insufficient information regarding PSG within the training data of the applied locally run LLMs. In this study, the locally run LLMs stated frequently the correct AHI from the pdf file indicating respectable performance in processing presented pdf files. When mentioning the AHI in the response, Gemma 2 stated the correct AHI in 100% (24/24), Mistral Nemo in 93% (13/14) and Llama 3 in 90% (20/22).

https://doi.org/10.2147/NSS.S536823



Concordance: board-certified sleep physician & Mistral Nemo							
patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
1	no	no	yes	yes	no	no	yes
2	no	yes	yes	no	yes	yes	yes
3	no	yes	no	no	no	yes	yes
4	no	no	yes	no	no	yes	yes
5	no	yes	yes	yes	no	yes	yes
6	no	yes	yes	yes	no	yes	yes
7	no	yes	yes	yes	no	yes	yes
8	no	yes	yes	yes	no	yes	yes
9	yes	yes	no	no	no	yes	yes
10	no	yes	yes	no	no	yes	yes
11	yes	yes	yes	no	no	no	yes
12	no	yes	no	yes	no	yes	yes
13	yes	yes	no	yes	no	yes	yes
14	yes	yes	yes	no	no	yes	no
15	yes	yes	yes	yes	no	yes	yes
16	no	yes	yes	yes	no	yes	yes
17	yes	yes	yes	yes	no	yes	yes
18	no	yes	yes	no	no	yes	yes
19	no	yes	yes	no	no	yes	yes
20	no	yes	yes	yes	no	yes	yes
21	yes	yes	yes	yes	no	yes	yes
22	yes	yes	yes	yes	no	yes	yes
23	yes	yes	yes	no	no	yes	yes
24	yes	yes	no	yes	no	yes	yes
25	no	no	yes	no	no	yes	yes
26	yes	yes	yes	yes	no	yes	yes
27	yes	yes	no	yes	no	yes	yes
28	yes	yes	no	yes	no	yes	yes
29	yes	yes	yes	yes	no	yes	yes
30	no	yes	yes	yes	no	yes	no
concordance	14	27	23	19	1	28	28
ratio	0.47	0.9	0.77	0.63	0.03	0.93	0.93

Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV): Mistral Nemo							
patient	diagnosis	therapy		alternative therapy			
		aPAP	weight loss	positional therapy	oral appliances	lifestyle change	surgery
SEN	1	1	0	0.5	1	n/a	0
SPE	0.5	0.4	1	0.64	0.07	0.93	1
PPV	0.97	0.89	n/a	0.09	0.04	0	n/a
NPV	1	1	1	0.95	1	1	0.93

Figure 6 Concordance between the board-certified sleep physician and the locally run Mistral Nemo LLM (absolute values/ratios per 30 patients). Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) as compared to the gold standard (board-certified sleep physician). n/a for values that cannot be calculated (division by zero) ie Mistral Nemo did not recommend weight loss for any patient.

Despite the current limitations documented in the present study, locally run LLMs hold significant promise for clinical deployment, particularly when combined with traditional human expertise in a hybrid approach. Clinicians could supervise the model's diagnostic suggestions, thereby providing a second level of scrutiny that reduces the likelihood of patient harm caused by misclassification or insufficiently robust therapy recommendations. This form of human-in-theloop oversight has been shown to improve both safety and user trust in other healthcare AI applications. Additionally, explainable AI frameworks can foster transparency by enabling clinicians to trace model reasoning and better understand why certain recommendations are made.

From a regulatory standpoint, software that provide diagnostic or therapeutic guidance fall under the scope of the medical device regulation. In the European Union, the Medical Device Regulation (MDR) (EU) 2017/745 stipulates stringent requirements regarding risk classification, clinical evaluation, and post-market surveillance for technologies that influence patient management or outcome.[10] Similarly, the International Medical Device Regulators Forum (IMDRF) has issued guidelines on SaMD, emphasizing the need for transparency, evidence-based validation, and a lifecycle approach to software.[11] In the United States, the Food and Drug Administration (FDA) applies its own framework for SaMD, requiring premarket submissions, ongoing monitoring, and labeling requirements when software performs clinical decision support.[12] Moreover, the recently passed EU AI Act may introduce additional obligations, particularly since AI systems utilized in healthcare settings are generally considered high-risk.[13] Future iterations of locally run LLMs intended for direct clinical decision-making should therefore be developed in close alignment with these regulatory frameworks to ensure patient safety, data integrity, and legal compliance.

There are four major limitations of our study. First, the locally run LLMs tested here (Gemma 2, Llama 3 and Mistral Nemo) are less well known and used than web-based LLMs such as ChatGPT. Due to promising results in the application of ChatGPT in recent studies,[3–5,7] a review of the analysis of real-world polysomnographic data with ChatGPT is of great interest. However, current data protection concerns prohibit such an approach. Second, the selected and analyzed data set is rather small, and this circumstance could have influenced the results independently. Third, the diagnosis and treatment recommendations by the locally run LLMs were based exclusively on polysomnographic data. The personal discussion with patients and the consideration of patient-specific adapted evaluation of the measurement can only be achieved by sleep physicians. The lack of inter-individual experience of the sleep medicine patient “behind polysomnographic values” could have independently influenced the presented results. Fourth, the polysomnographic data was passed to the locally run LLMs as an edited pdf file including graphics and tables. In these pdf files, all parameters were displayed in German. The use of another language such as English could influence the generated results.

Despite these limitations, this study using real-world polysomnographic patient data processed by locally run LLMs is the first of its kind. Our study demonstrates the potential of locally run LLMs in the field of sleep medicine. Considering further LLM performance improvements in the future in conjunction with the option of model fine tuning with sleep medicine data, multilingual training or better input standardization LLMs have high potential to assist physicians, streamline sleep medicine and alleviate the shortage of specialists. However, meaningful improvement will require significant data resources. Thus, innovative approaches are required for the problem of making relevant clinical real-world data accessible without affecting individual data protection rights. While locally running LLMs avoid data protection constraints in their application, the problem of sufficient training or fine tuning on sleep medicine data remains, as locally available computing power may not be sufficient for adequate modification of the models. The integration of LLMs in sleep medicine continues to be a highly dynamic area with a variety of potential applications and potential to provide improved and supportive effects for sleep physicians in clinical practice.

Conclusion

This is the first study to demonstrate the general ability of locally run LLMs to interpret real-world polysomnographic data, providing diagnosis and therapy recommendations. Nevertheless, locally run LLMs still need significant improvement before they can reliably assist sleep physicians. As time and research progress, future studies should evaluate more powerful locally run LLMs on a larger scale. While the technical infrastructure for locally run LLMs already exists, their implementation in clinical practice cannot yet be fully recommended.

Abbreviations

LLM, Large language model; PSG, Polysomnography; OSA, Obstructive sleep apnea;

PAP, Positive airway pressure; AASM, American Academy of Sleep Medicine; EU, European Union; MDR, Medical Device Regulation; IMDRF, International Medical Device Regulators Forum; FDA, Food and Drug Administration.

Ethics Approval and Consent to Participate

The study protocol was approved by the ethics commission of the Rhineland-Palatinate State Medical Association (reference no: 2023-17391). In addition, the study protocol was discussed with the university medical center's local data protection officer. To protect patient privacy, all data was analyzed exclusively on an institutional computer with no external connections. No real-world patient data was transmitted to an internet-connected applications or third parties.

Data Sharing Statement

Only data shared within the paper can be provided due to the use of real patient data.

Author Contributions

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

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

KB-H: Formal analysis, Investigation, Validation, Writing – original draft, Writing – review and editing.

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

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

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

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

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

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

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

SK: Formal analysis, Investigation, Validation, 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. HS joined Manual.co a Health Company on April 28, 2025. All the work on the project was completed prior to his employment. HG reports grants from Inspire Medical Systems, Inc., outside the submitted work. The other authors have no conflict of interest to declare.

References

1. 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
2. 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
3. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. *J Clin Sleep Med*. 2023;19(12):1989–1995. doi:10.5664/jcsm.10728
4. Martini A, Ielo S, Andreani M, Siciliano M. ChatGPT: friend or foe of patients with sleep-related breathing disorders? *Sleep Epidemiology*. 2024;4:100076. doi:10.1016/j.sleepe.2024.100076
5. Mira FA, Favier V, Dos Santos Sobreira Nunes H, et al. Chat GPT for the management of obstructive sleep apnea: do we have a polar star? *Eur Arch Otorhinolaryngol*. 2024;281(4):2087–2093. doi:10.1007/s00405-023-08270-9

6. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. *Communicat Med*. 2025;5(1). doi:10.1038/s43856-024-00717-2
7. Seifen C, Huppertz T, Gouveris H, et al. Chasing sleep physicians: ChatGPT-4o on the interpretation of polysomnographic results. *Europ Archiv Oto-Rhino-Laryngol*. 2024;282(3):1631–1639. doi:10.1007/s00405-024-08985-3
8. Medicine AAoS. *International Classification of Sleep Disorders*. 3rd ed. American Academy of Sleep Medicine; 2014.
9. Mesko B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6 (1):120. doi:10.1038/s41746-023-00873-0
10. Commission E. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/E. Official Journal of the European Union; 2017.
11. MDCG 2020-1 Guidance on Clinical Evaluation (MDR) / Performance Evaluation (IVDR) of Medical Device Software. 2020.
12. FDA. Policy for Device Software Functions and Mobile Medical Applications Guidance for Industry and Food and Drug Administration Staff. 2022.
13. Commission E. AI Act; 2024.