Original Article

Access this article online
Quick Response Code:

Website:
https://eurasianjpulmonol.org
DOI:
10.14744/ejp.2026.67756

Automated Versus Manual PSG Scoring: Evaluation of Somnolyzer in a Clinical Cohort

Damla Azaklı Yazıcı, İpek Çalık, Ayşe Bahadır, Sibel Yurt, Mehmet Akif Özgül

ORCID:
Damla Azaklı Yazıcı: 0000-0002-5850-2076
İpek Çalık: 0000-0001-7373-8581
Ayşe Bahadır: 0000-0002-7006-5550
Sibel Yurt: 0000-0002-7703-110X
Mehmet Akif Özgül: 0000-0003-1110-6823

Abstract:

BACKGROUND AND AIM: Manual polysomnography (PSG) scoring is time-consuming and subject to inter-scorer variability. Automated systems such as *Somnolyzer24X7* (*The Siesta Group Schlafanalyse GmbH, Vienna, Austria*) have been developed to improve efficiency and standardization, but real-world validation data remain limited. This study aimed to evaluate the performance of Somnolyzer in a clinical PSG cohort by comparing automated outputs with manual scoring by two independent human scorers.

METHODS: This methodological comparison study included 42 full-night diagnostic PSG recordings. Each study was scored independently by two experienced human scorers according to AASM criteria, and by Somnolyzer using automated analysis. Twenty-two PSG parameters were extracted, including sleep architecture measures, respiratory indices, and arousal indices. Agreement was assessed using intraclass correlation coefficients (ICC), and differences in PSG parameters across scoring methods were analyzed using the Friedman test with post-hoc comparisons.

RESULTS: Measures of sleep continuity and REM-related parameters did not differ significantly between scoring approaches (all $p>0.05$). Significant differences were observed for N1 %, N2 %, and N3 % and for multiple respiratory indices (all $p<0.05$). Somnolyzer showed excellent agreement with the human consensus for key respiratory parameters, including obstructive apnea index (ICC 0.97), apnea index (ICC 0.96), and apnea–hypopnea index (ICC 0.93). Agreement was also high for total sleep time (ICC 0.94), sleep efficiency (ICC 0.95), and REM percentage (ICC 0.90). Moderate agreement was observed for sleep stage proportions, whereas lower agreement was observed for the mixed apnea index and the arousal index.

CONCLUSIONS: Somnolyzer provides automated PSG scoring results that are largely consistent with expert manual scoring and may serve as a reliable first-pass analysis tool in clinical practice.

Keywords:

Automated sleep scoring, polysomnography, sleep staging, sleep scoring, Somnolyzer

Department of Pulmonology, Başakşehir Çam and Sakura City Hospital, İstanbul, Türkiye

Address for correspondence:
Dr. Damla Azaklı Yazıcı, Department of Pulmonology, Başakşehir Çam and Sakura City Hospital, İstanbul, Türkiye.
E-mail: azakli.damla@gmail.com

Received: 27-03-2026
Revised: 22-04-2026
Accepted: 29-04-2026
Published: 03-08-2026

How to cite this article: Azaklı Yazıcı D, Çalık İ, Bahadır A, Yurt S, Özgül MA. Automated Versus Manual PSG Scoring: Evaluation of Somnolyzer in a Clinical Cohort. Eurasian J Pulmonol 0000;00:1-8.

Introduction

Sleep staging remains a fundamental component of polysomnography (PSG) analysis and is still largely performed through manual scoring by trained technologists. Manual sleep staging involves visual inspection of overnight PSG recordings and may require up to two hours to score an eight-hour PSG recording.[1] Despite standardized scoring criteria, sleep stage scoring remains susceptible to variability among scorers. Inter-scorer agreement between trained sleep technologists has been reported to range around 70–80%, while intra-scorer agreement is typically around 90%.[2–4] In a large dataset including more than 2,500 scorers and over 3 million scoring decisions, the overall agreement for sleep staging was approximately 83%, with particularly low agreement for stages N1 and N3.[2]

Recent advances in machine learning and deep learning have therefore aimed to automate sleep staging and provide more objective representations of sleep architecture.[5] Over the years, several automated or semi-automated sleep scoring systems have been developed, including commercial platforms such as *Somnolyzer24X7* (*The Siesta Group Schlafanalyse GmbH, Vienna, Austria*),[6–9] *Morpheus by WideMed, Ltd.* (*Israel*),[9,10] and the *Michele Sleep Scoring System* (*Cerebra, Winnipeg, Canada*),[11] to facilitate or partially automate the scoring process.

One of the early automated sleep staging systems developed to facilitate PSG scoring is Somnolyzer 24/7, which integrates automated signal processing and rule-based classification to emulate visual scoring according to the Rechtschaffen and Kales criteria.[6,12] It was developed and validated using the multicenter Siesta database, which includes nearly 600 PSG recordings from healthy individuals and patients with sleep disorders. In validation studies, Somnolyzer achieved an epoch-by-epoch agreement of approximately 80% (Cohen's κ=0.72), comparable to inter-scorer agreement between human experts.[6] These findings suggest that automated scoring can achieve a performance similar to that of manual scoring.

To the best of our knowledge, the performance of Somnolyzer has not yet been systematically evaluated in a Turkish PSG cohort. Given potential differences in patient characteristics, recording conditions, and scorer practices across sleep laboratories, external validation in diverse clinical populations is important to assess the generalizability of automated sleep staging systems. Therefore, the aim of the present study was to evaluate the performance of Somnolyzer in a Turkish clinical PSG dataset by comparing automated sleep stage classification with manual scoring performed by experienced human scorers.

Materials and Methods

Study design and participants

This methodological comparison study included 42 full-night diagnostic PSG recordings obtained in a sleep laboratory between 1 September and 15 October 2025. The number of recordings was determined by the availability of automated Somnolyzer analyses during the study period. Within this limit, all eligible full-night diagnostic PSG recordings were included consecutively without pre-selection. The final dataset comprised 42 PSG recordings from 42 individual patients, each contributing a single full-night study, with no repeated recordings. As the recordings were obtained during routine clinical practice, the study population represents a real-world cohort of patients referred for evaluation of suspected sleep disorders, including obstructive sleep apnea and related conditions. No additional inclusion or exclusion criteria were applied beyond the availability of a technically adequate full-night diagnostic PSG recording during the study period. All recordings were anonymized prior to analysis. Each PSG was scored independently by two human scorers blinded to each other's results and to the Somnolyzer output; Somnolyzer performed automated analysis.

The study protocol was approved by Başakşehir Çam and Sakura City Hospital Scientific Research Ethics Committee (Approval Number: 268, Date: 27.08.2025). The study was conducted in accordance with the principles of the Declaration of Helsinki. All PSG recordings were obtained as part of routine clinical care. Written informed consent for the use of clinical data for research purposes had been obtained from patients or their legal guardians at the time of the sleep study. No additional study-specific informed consent was required.

PSG recording and scoring

All studies consisted of standard attended overnight PSG recordings performed using routine clinical acquisition protocols. The recorded signals included electroencephalography (EEG: F3–M2, F4–M1, C3–M2, C4–M1, O1–M2, O2–M1), electrooculography (EOG), submental electromyography (EMG), electrocardiography (ECG), nasal airflow

(measured by a nasal pressure transducer and a thermistor), thoracic and abdominal respiratory effort (measured with belts), and pulse oximetry (SpO_2). Sleep staging and respiratory event scoring were performed according to the most recent version (Version 3.0) of the American Academy of Sleep Medicine (AASM) Scoring Manual. Respiratory events were defined as follows: Apnea was defined as a ≥90% reduction in airflow from baseline lasting at least 10 seconds. Obstructive apnea was scored when respiratory effort persisted during the airflow reduction. Central apnea was scored when both airflow and respiratory effort were absent. Mixed apnea was defined as an apnea event beginning without respiratory effort and ending with the resumption of respiratory effort. Hypopnea was defined as a ≥30% reduction in airflow lasting at least 10 seconds, associated with either ≥3% oxygen desaturation or an arousal. An arousal was defined as an abrupt shift in EEG frequency lasting at least 3 seconds, preceded by at least 10 seconds of stable sleep. Both human scorers manually evaluated the recordings according to these criteria. Somnolyzer automatically generated sleep staging and scored respiratory events using its internal algorithm. Sleep staging was performed using 30-second epochs.

Data extraction

From each scoring output (Human scorer 1, Human scorer 2, and Somnolyzer), PSG parameters were extracted. These included measures of sleep continuity (total sleep time [TST], sleep onset latency [SOL], sleep efficiency [SE], and wake after sleep onset [WASO]); sleep architecture (N1, N2, N3, and rapid eye movement [REM] sleep percentages); respiratory event indices (central apnea index [CAI], obstructive apnea index [OAI], mixed apnea index [MAI], apnea index [AI], hypopnea index [HI], and apnea–hypopnea index [AHI]); positional respiratory indices (supine and non-supine AHI); and the arousal index.

Somnolyzer automated analysis

Automated analysis was performed using Somnolyzer software integrated within the Philips sleep analysis platform (Philips Respironics, Murrysville, PA, USA). The system automatically derives standard PSG parameters, including sleep architecture metrics and respiratory event indices, from recorded physiological signals. For the present study, Somnolyzer outputs were extracted from each recording and compared with manual scoring, which was performed independently by two experienced scorers according to the AASM criteria. Human consensus values were defined as the mean of the two manual scores and used as the reference for agreement analyses.

Statistical analysis

Continuous variables were assessed for normality using the Shapiro–Wilk test and visual inspection of distribution plots. As most variables were not normally distributed, continuous data are presented as median (minimum–maximum), and non-parametric tests are used for group comparisons. Comparisons among the three scoring approaches (Human scorer 1, Human scorer 2, and Somnolyzer) were performed using the Friedman test for repeated measures. When a significant overall difference was detected, Durbin–Conover pairwise post-hoc comparisons were conducted. Inter-scorer agreement was assessed using the Intraclass Correlation Coefficient (ICC), specifically the two-way random-effects model with absolute agreement [ICC(2,k)], following the classification of Shrout and Fleiss. ICC values were interpreted according to commonly accepted reliability thresholds, with values below 0.50 indicating poor agreement, 0.50–0.75 indicating moderate agreement, 0.75–0.90 indicating good agreement, and above 0.90 indicating excellent agreement. ICC results were also visualized using forest plots. All analyses were performed using R version 4.4.2. A two-sided P value <0.05 was considered statistically significant.

Results

The analysis included 42 PSG recordings from 42 patients, of whom 33 (78.6%) were male; the mean age was 47.7 years. PSG parameters derived from the two human scorers and Somnolyzer are presented in Table 1. Measures reflecting overall sleep continuity and architecture, including TST, SE, SOL, WASO, REM latency, and REM %, did not differ significantly among the three scoring approaches (all pairwise comparisons, $p > 0.05$). In contrast, significant differences were observed for several sleep stage parameters. For N1%, N2%, and N3%, Human scorer 1 differed significantly from both Human scorer 2 and Somnolyzer, whereas no significant differences were observed between Human scorer 2 and Somnolyzer. Respiratory event indices differed consistently among scorers. CAI, OAI, AI, HI, and AHI differed significantly between scoring approaches (all $p<0.001$). In most respiratory parameters, Somnolyzer produced values closer to Human scorer 2 than to Human scorer 1. Similar patterns were observed for positional metrics, including both supine and non-supine AHI (both $p<0.001$).

The arousal index also differed significantly between scorers ($p<0.001$). Human scorer 2 reported the highest

Table 1: Comparison of PSG parameters between two human scorers and Somnolyzer

Parameter	Human scorer 1	Human scorer 2	Somnolyzer	p
TST (min)	341 (152–460)[a]	341 (155–449)[a]	343 (154–449)[a]	0.10
SE (%)	79.9 (36.3–96.1)[a]	79.7 (37.1–97.4)[a]	79.8 (37.1–97.4)[a]	0.20
SOL (min)	18.5 (2–170)[a]	18.5 (1.5–170)[a]	17.5 (1.5–170)[a]	0.20
WASO (min)	65.3 (3.5–243)[a]	65 (3.5–242)[a]	68.3 (3.5–242)[a]	0.80
REM latency (min)	123 (58.5–384)[a]	111 (58–389)[a]	111 (58–389)[a]	0.40
N1 sleep (%)	3.70 (1.4–14.9)[a]	4.10 (1.6–14.5)[b]	4.10 (1.6–17.8)[b]	<0.001
N2 sleep (%)	68.1 (6.4–94.4)[a]	70.7 (47.9–96.8)[b]	70.2 (13.2–96.8)[b]	0.03
N3 sleep (%)	13.1 (0–31.8)[a]	7.3 (0–26.6)[b]	7.3 (0–26.5)[b]	<0.001
REM sleep (%)	16 (0–40)[a]	17 (0–40.7)[a]	17.1 (0–40.7)[a]	0.20
CAI (events/h)	0 (0–13.8)[a]	0 (0–14.4)[a]	0.2 (0–16.9)[b]	<0.001
OAI (events/h)	0.7 (0–76.8)[a]	0.25 (0–63.7)[b]	0.7 (0–76.5)[a]	<0.001
MAI (events/h)	0 (0–0.8)[a]	0 (0–1.5)[a]	0 (0–0.6)[a]	0.10
AI (events/h)	0.8 (0–77.4)[a]	0.3 (0–65.4)[b]	1.2 (0–77)[a]	<0.001
HI (events/h)	4.3 (0.3–68.8)[a]	7.65 (0–82.6)[b]	8.2 (0.7–71)[c]	<0.001
AHI (events/h)	7.35 (0.5–86.4)[a]	9.85 (0–82.8)[b]	11.6 (0.8–89.4)[c]	<0.001
Supine AHI (events/h)	11.6 (0–107)[a]	16.8 (0–89)[b]	19.7 (0–111)[c]	<0.001
Non-supine AHI (events/h)	2.55 (0–87.2)[a]	3.57 (0–87)[b]	5.43 (0–94.4)[c]	<0.001
Arousal index (events/h)	4.75 (0.4–49.5)[a]	18.3 (2.4–44.5)[b]	14.6 (5.5–69.3)[b]	<0.001

Overall p-values were calculated using the Friedman test. Values sharing the same superscript letter are not significantly different based on Durbin–Conover post-hoc pairwise comparisons. PSG: Polysomnography, TST: Total sleep time, SE: Sleep efficiency, SOL: Sleep onset latency, WASO: Wake after sleep onset, REM: Rapid eye movement, CAI: Central apnea index, OAI: Obstructive apnea index, MAI: Mixed apnea index, AI: Apnea index, HI: Hypopnea index, AHI: Apnea–hypopnea index.

Table 2: Inter-scorer agreement between human scorers and agreement between Somnolyzer and the human consensus across PSG parameters

Parameter	Human inter-scorer agreement			Somnolyzer vs. human consensus*		
	ICC	95% CI		ICC	95% CI	
		Lower	Upper		Lower	Upper
TST (min)	0.826	0.698	0.902	0.944	0.898	0.969
SE (%)	0.823	0.693	0.901	0.947	0.903	0.971
SOL (min)	0.991	0.982	0.995	0.841	0.724	0.911
WASO (min)	0.767	0.605	0.868	0.882	0.792	0.935
REM latency (min)	0.826	0.689	0.906	0.936	0.880	0.966
N1 sleep (%)	0.510	0.252	0.702	0.785	0.629	0.879
N2 sleep (%)	0.630	0.399	0.784	0.656	0.441	0.800
N3 sleep (%)	0.595	0.145	0.805	0.848	0.646	0.928
REM sleep (%)	0.714	0.528	0.835	0.904	0.829	0.947
CAI (events/h)	0.154	-0.159	0.437	0.655	0.444	0.798
OAI (events/h)	0.830	0.677	0.909	0.966	0.938	0.981
MAI (events/h)	0.279	-0.027	0.536	0.481	0.208	0.684
AI (events/h)	0.826	0.683	0.906	0.960	0.927	0.978
HI (events/h)	0.894	0.741	0.950	0.902	0.730	0.957
AHI (events/h)	0.971	0.948	0.984	0.925	0.796	0.967
Supine AHI (events/h)	0.769	0.604	0.871	0.841	0.690	0.917
Non-supine AHI (events/h)	0.965	0.936	0.981	0.932	0.853	0.966
Arousal index (events/h)	0.215	-0.077	0.488	0.493	0.209	0.696

*: Human consensus was defined as the mean score of the two independent human scorers. PSG: Polysomnography, ICC: Intraclass correlation coefficient, CI: Confidence interval, TST: Total sleep time, SE: Sleep efficiency, SOL: Sleep onset latency, WASO: Wake after sleep onset, REM: Rapid eye movement, CAI: Central apnea index, OAI: Obstructive apnea index, MAI: Mixed apnea index, AI: Apnea index, HI: Hypopnea index, AHI: Apnea–hypopnea index.

arousal index values, whereas Somnolyzer yielded intermediate values that were significantly higher than those of Human scorer 1 but not different from those of Human scorer 2 in post-hoc comparisons (Table 1).

Agreement between Somnolyzer outputs and the human consensus was generally high across most PSG parameters (Table 2; Fig. 1). For respiratory indices, Somnolyzer demonstrated excellent agreement with the human con-

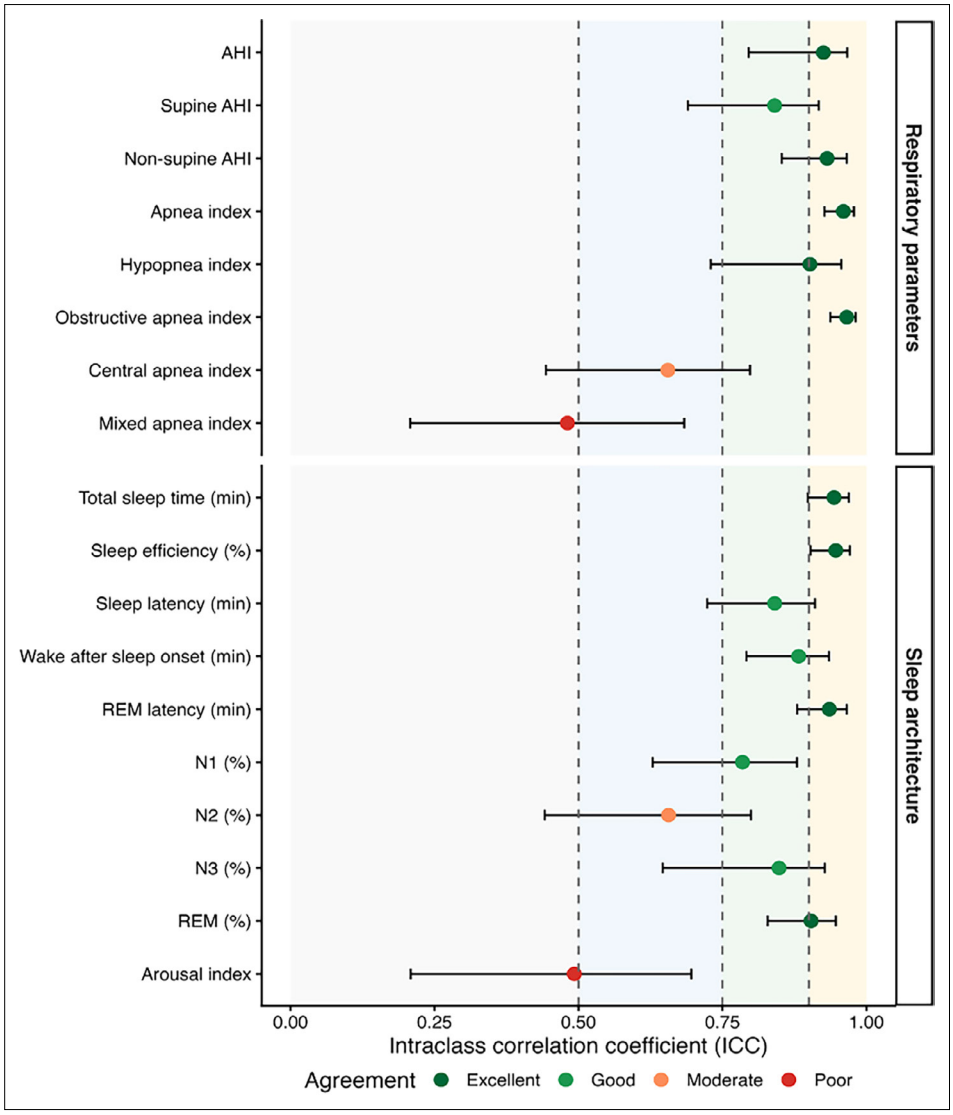

Figure 1: Agreement between Somnolyzer and the human consensus across PSG parameters. Forest plot of ICC and 95% CI. Dashed lines indicate ICC thresholds (0.50, 0.75, 0.90)

PSG: Polysomnography, ICC: Intraclass correlation coefficient, CI: Confidence interval

sensus, particularly for OAI (ICC 0.97, 95% CI 0.94–0.98), AI (ICC 0.96, 95% CI 0.93–0.98), AHI (ICC 0.93, 95% CI 0.80–0.97), and non-supine AHI (ICC 0.93, 95% CI 0.85–0.97). Agreement was also high for key sleep parameters, including TST (ICC 0.94), SE (ICC 0.95), REM latency (ICC 0.94), and REM% (ICC 0.90). Moderate agreement was observed for several sleep stage proportions, including N1% (ICC 0.78), N2% (ICC 0.66), and N3% (ICC 0.85). Lower agreement was observed for MAI (ICC 0.48) and arousal index (ICC 0.49) (Table 2).

Human inter-scorer agreement showed a broadly similar pattern, but was substantially lower for certain parameters, particularly for CAI (ICC 0.15, 95% CI -0.16–0.44) and the arousal index (ICC 0.22, 95% CI -0.08–0.49). Notably, for several parameters, such as CAI and N3%, agreement between Somnolyzer and the human consensus exceeded that observed among human scorers. The distribution of ICC values across parameters is illustrated in the forest plot [Fig. 1].

Discussion

In this methodological comparison study, we evaluated the performance of an automated sleep-scoring system (Somnolyzer) against manual scoring performed by two experienced human scorers, using a clinical PSG dataset. Overall, Somnolyzer demonstrated high agreement with

the human consensus for most clinically relevant parameters, particularly respiratory indices and global sleep measures. Agreement was moderate for sleep stage proportions but lower for the MAI and the arousal index. These findings indicate that automated PSG scoring yields results largely consistent with expert manual scoring.

In the original validation study by Anderer et al.,[6] Somnolyzer was developed and evaluated using the multicenter Siesta database, which included nearly 600 PSG recordings from both healthy individuals and patients with various sleep disorders. The system achieved an epoch-by-epoch agreement of approximately 80% with expert manual scoring (Cohen's κ=0.72), comparable to inter-scorer agreement between human experts. Our findings are consistent with this prior work, indicating that Somnolyzer yields results comparable to expert scoring and supporting its use in clinical sleep analysis. However, the studies differ in key aspects. While the original study focused on epoch-by-epoch agreement for sleep stage classification, our study assessed agreement across clinically derived PSG parameters, reflecting real-world clinical interpretation. This distinction is important because small epoch-level discrepancies may have limited impact on aggregated PSG variables used in clinical decision-making, such as sleep architecture percentages and respiratory indices. Sleep is a continuous neurophysiological process that is discretized into fixed 30-second epochs, making stage transitions inherently ambiguous, particularly between adjacent stages such as N1–N2 and N2–N3.[4] Inter-scorer variability is therefore most pronounced during these transition periods and reflects interpretation of borderline epochs rather than true physiological differences.[13] Consistent with this, N1 shows the lowest inter-scorer agreement.[14] In contrast, respiratory indices demonstrate higher reliability, suggesting that variability in epoch-level staging does not necessarily translate into clinically meaningful differences in aggregated PSG metrics.

Consistent with these considerations, variability in specific respiratory indices was more pronounced in the present study. Human inter-scorer agreement was particularly low for CAI (ICC 0.15) and MAI (ICC 0.28), whereas agreement between Somnolyzer and the human consensus was modest (ICC 0.65 and 0.48 for CAI and MAI, respectively). This variability likely reflects the low frequency of these events, in which small absolute differences in counts can substantially affect agreement metrics. In addition, classification of respiratory events depends on the interpretation of respiratory effort signals according to AASM criteria and may be influenced by signal quality and subtle variations in thoracoabdominal movement patterns.[15] Mixed apneas, representing transitional events with both central and obstructive components, may further increase variability. Similar variability in respiratory event scoring across scorers and laboratories has been reported in previous studies.[4,13] From a clinical perspective, these findings have direct implications for how differences between scoring approaches should be interpreted. Indices such as AHI, HI, and positional AHI differed systematically across scoring methods, with Somnolyzer generally yielding higher values, even when absolute differences were small, particularly in cases near diagnostic thresholds. Because AHI defines OSA severity, such differences may lead to reclassification across severity categories and influence treatment decisions, particularly in patients near diagnostic thresholds. These results indicate that observed differences reflect not only random variability but also systematic differences in event detection, highlighting the importance of careful interpretation in borderline clinical scenarios.

Lower agreement was also observed for the arousal index. Human inter-scorer agreement was low (ICC 0.22), and agreement between Somnolyzer and the human consensus remained moderate (ICC 0.49). This variability is consistent with previous studies showing that arousal scoring is inherently challenging. According to AASM criteria, cortical arousals are defined by transient EEG frequency shifts lasting at least 3 seconds, but their identification often requires interpretation of subtle and borderline EEG patterns.[4,14,15] Arousals frequently occur near stage transitions, where distinguishing among brief awakenings, stage changes, and short cortical activations is difficult. In addition, because sleep staging discretizes a continuous physiological process into fixed 30-second epochs, scorers may differ in whether an epoch is classified as wakefulness or sleep. As a result, brief awakenings may be inconsistently classified, contributing to variability in the arousal index.

Taken together, these observations suggest that perfect epoch-by-epoch agreement may not be essential for reliable clinical interpretation of PSG recordings. This is consistent with the concept that sleep is a continuously evolving neurophysiological process, whereas conventional sleep staging represents a simplified framework imposed on its underlying dynamics.[16,17] In the present study, Somnolyzer demonstrated good agreement with human scoring

for major sleep-stage proportions and excellent agreement for key respiratory indices; however, variability was primarily observed in scoring of central and mixed apneas.

Beyond earlier validation studies of rule-based automated systems such as Somnolyzer, more recent deep learning models have advanced automated sleep staging by learning directly from raw multichannel PSG signals and incorporating temporal context into classification.[18] At the same time, growing evidence indicates that AI-based sleep research increasingly relies on multivariate and multimodal physiological data, reflecting a broader shift toward automated and data-driven PSG analysis,[19] within which systems like Somnolyzer may represent an intermediate step between rule-based and fully data-driven approaches.

A key consideration in interpreting these findings is that manual PSG scoring, even when based on consensus between expert scorers, does not represent a true gold standard. Sleep staging and respiratory event scoring are subject to inherent inter-scorer variability, particularly in borderline or transitional epochs. Therefore, the reference used in this study should be viewed as an approximation of expert interpretation rather than an absolute ground truth. However, manual scoring remains the current clinical standard and the basis for diagnosis and treatment decisions in routine practice; thus, agreement with expert scoring serves as a clinically meaningful benchmark. In this context, the variability observed between human scorers further underscores the potential value of automated systems. Rather than reproducing an idealized gold standard, automated scoring may provide greater consistency and reproducibility, particularly for parameters prone to subjective interpretation.

Taken together, these findings support the implementation of automated scoring systems as first-pass analysis tools within clinical workflows. In this approach, the automated system performs an initial scoring of the PSG recording, which is subsequently reviewed and refined by an experienced clinician. This targeted review strategy can reduce scoring time while maintaining clinical accuracy by directing human attention to segments that are more prone to variability or uncertainty. In this context, automated systems such as Somnolyzer can enhance efficiency and standardization in routine sleep laboratory practice, supporting their integration into a hybrid human–AI scoring workflow.

Limitations

Several limitations of this study should be acknowledged. First, the sample size was relatively modest and derived from a single sleep laboratory, which may limit the generalizability of the findings to other clinical populations or recording environments. Second, the study evaluated agreement at the level of aggregated PSG parameters rather than performing an epoch-by-epoch comparison of sleep stage classification. While parameter-level agreement is highly relevant for clinical interpretation and reporting, epoch-level analyses may provide additional insights into the specific types of staging discrepancies between automated and manual scoring. Third, no subgroup analyses were performed based on patient characteristics or OSA severity. As agreement metrics may vary across different patient groups, this may have influenced the observed results. Fourth, the human consensus used as the reference standard was defined as the mean of two independent human scorers. Although this approach is commonly used in methodological validation studies, manual PSG scoring is subject to inter-scorer variability, as demonstrated by the relatively low agreement observed for certain parameters such as CAI and the arousal index in the present study. Consequently, the reference standard cannot be considered an absolute ground truth, but rather represents an approximation of expert consensus. Finally, the study focused on standard PSG parameters derived from routine clinical scoring and did not evaluate additional aspects of automated analysis such as signal quality detection, artifact handling, or performance in specific patient subgroups. Future studies incorporating larger datasets and more detailed epoch-level analyses may further clarify the strengths and limitations of automated PSG scoring systems in clinical sleep medicine.

Conclusion

In this clinical PSG dataset, the Somnolyzer automated scoring system demonstrated high agreement with expert manual scoring across most clinically relevant sleep and respiratory parameters. Agreement was particularly strong for respiratory indices and global sleep metrics, whereas lower agreement was observed for parameters known to show higher inter-scorer variability, such as arousal index and MAI. These findings suggest that automated scoring may provide a reliable first-pass analysis of PSG recordings and reduce scoring workload while maintaining clinical consistency.

Ethics Committee Approval

The study was approved by the Başakşehir Çam and Sakura City Hospital Scientific Research Ethics Committee (No: 268, Date: 27/08/2025).

Informed Consent

Written informed consent for the use of clinical data for research purposes had been obtained from patients or their legal guardians at the time of the sleep study.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Funding

The authors declared that this study received no financial support.

Use of AI for Writing Assistance

AI-assisted tools were used for language editing and minor revisions. The authors take full responsibility for the content of the manuscript.

Author Contributions

Concept – D.A.Y., İ.Ç., A.B., S.Y., M.A.Ö.; Design – D.A.Y., İ.Ç., A.B., S.Y., M.A.Ö.; Supervision – D.A.Y., İ.Ç., A.B., S.Y., M.A.Ö.; Resource – D.A.Y., İ.Ç.; Materials – D.A.Y., İ.Ç.; Data Collection and/or Processing – İ.Ç., A.B.; Analysis and/or Interpretation – D.A.Y., A.B.; Literature Review – D.A.Y., İ.Ç., A.B., S.Y., M.A.Ö.; Writing – D.A.Y., İ.Ç., A.B.; Critical Review – S.Y., M.A.Ö.

Acknowledgments

The authors thank Didem Akgün, Özge Karataş, and Yeşim Yetişgin for their support in the sleep laboratory.

Peer-review

Externally peer-reviewed.

References

1. Fiorillo L, Puiatti A, Papandrea M, Ratti PL, Favaro P, Roth C, et al. Automated sleep scoring: A review of the latest approaches. Sleep Med Rev 2019;48:101204. [CrossRef]
2. Younes M, Raneri J, Hanly P. Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability. J Clin Sleep Med 2016;12(6):885–94. [CrossRef]
3. Muto V, Berthomier C, Schmidt C, Vandewalle G, Jaspar M, Devillers J, et al. 0315 Inter- And Intra-expert Variability In Sleep Scoring: Comparison Between Visual And Automatic Analysis. 2018;41(Suppl 1):A121. [CrossRef]
4. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9(1):81–7. [CrossRef]
5. Stephansen JB, Olesen AN, Olsen M, Ambati A, Leary EB, Moore HE, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat Commun 2018;9(1):5229. [CrossRef]
6. Anderer P, Gruber G, Parapatics S, Woertz M, Miazhynskaia T, Klosch G, et al. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 x 7 utilizing the Siesta database. Neuropsychobiology 2005;51(3):115–33. [CrossRef]
7. Anderer P, Moreau A, Woertz M, Ross M, Gruber G, Parapatics S, et al. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 × 7. Neuropsychobiology 2010;62(4):250–64. [CrossRef]
8. Punjabi NM, Shifa N, Dorffner G, Patil S, Pien G, Aurora RN. Computer-Assisted Automated Scoring of Polysomnograms Using the Somnolyzer System. Sleep 2015;38(10):1555–66. [CrossRef]
9. Svetnik V, Ma J, Soper KA, Doran S, Renger JJ, Deacon S, Koblan KS. Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. Sleep 2007;30(11):1562–74. [CrossRef]
10. Pittman SD, MacDonald MM, Fogel RB, Malhotra A, Todros K, Levy B, et al. Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. Sleep 2004;27(7):1394–403. [CrossRef]
11. Gerardy B, Kuna ST, Pack A, Kushida CA, Walsh JK, Staley B, et al. An approach for determining the reliability of manual and digital scoring of sleep stages. Sleep 2023;46(11):zsad248. [CrossRef]
12. Rechtschaffen A, Kales A. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Public Health Service, U.S. Government Printing Office; 1968.
13. Danker-Hopfe H, Anderer P, Zeitlhofer J, Boeck M, Dorn H, Gruber G, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 2009;18(1):74–84. [CrossRef]
14. Whitney CW, Gottlieb DJ, Redline S, Norman RG, Dodge RR, Shahar E, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998;21(7):749–57. [CrossRef]
15. Berry RB, Brooks R, Gamaldo C, Harding SM, Lloyd RM, Quan SF, et al. AASM Scoring Manual Updates for 2017 (Version 2.4). J Clin Sleep Med 2017;13(5):665–66. [CrossRef]
16. Halász P. The microstructure of sleep. Suppl Clin Neurophysiol 2004;57:521–33. [CrossRef]
17. Parrino L, Ferri R, Bruni O, Terzano MG. Cyclic alternating pattern (CAP): the marker of sleep instability. Sleep Med Rev 2012;16(1):27–45. [CrossRef]
18. Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. IEEE Trans Neural Syst Rehabil Eng 2018;26(4):758–69. [CrossRef]
19. Fathima S, Ahmed M. Sleep Apnea Detection Using EEG: A Systematic Review of Datasets, Methods, Challenges, and Future Directions. Ann Biomed Eng 2025;53(5):1043–67. [CrossRef]