Nature and Science of Sleep

Elevate Series

Open Access Full Text Article

EDITORIAL

Beyond Algorithms: The Case for Standardized Reporting in AI Sleep Scoring

Ahmed S BaHammam [1,2], Malak Abdullah Almarshad [3]

[1]The University Sleep Disorders Center, Department of Medicine, College of Medicine, King Saud University, Riyadh, Saudi Arabia; [2]King Saud University Medical City, Riyadh, Saudi Arabia; [3]Department of Computer Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 11564, Saudi Arabia

Correspondence: Ahmed S BaHammam, The University Sleep Disorders Center, Department of Medicine, College of Medicine, King Saud University, Box 225503, Riyadh, 11324, Saudi Arabia, Email ashammam2@gmail.com

Introduction

Automated sleep scoring algorithms have matured substantially over the past five years, achieving Cohen's kappa values of 0.75–0.80 that approach the upper end of the range observed among experienced human scorers.[1–3] However, widespread clinical adoption remains constrained by unresolved challenges in data privacy, fairness and transparency, infrastructure requirements, and medical-legal accountability.[3] The American Academy of Sleep Medicine (AASM) emphasizes that responsible AI integration requires careful consideration to overcome clinical validation challenges, ensure ongoing accuracy after implementation, and incorporate clinically relevant and user-friendly tools into practice, while upholding standards of safety, appropriateness, and transparency.[3] Nonetheless, the rapid proliferation of AI applications has outpaced the development of standardized reporting frameworks specific to sleep medicine.[2,4] Despite these advances, fewer than 1% of AI sleep studies undergo rigorous external validation on independent datasets.[5] Most lack sufficient methodological transparency for independent replication, with training data and source code publicly unavailable in over 90% of studies.[6]

The absence of sleep medicine-specific reporting standards creates three critical challenges. First, heterogeneous reporting practices prevent meaningful comparison across studies and hinder meta-analyses. Second, inadequate documentation of algorithmic performance under real-world conditions limits clinical translation. Third, the unique characteristics of sleep data, including night-to-night variability, inherent interscorer disagreement in polysomnography, and the subjective nature of manual scoring, demand specialized reporting considerations beyond general AI guidelines.

Current Landscape: General AI Reporting Guidelines and Their Limitations for Sleep Medicine

Existing Frameworks

Recent years have witnessed important developments in AI reporting guidelines across medicine. The TRIPOD-AI statement (2024) provides updated guidance for clinical prediction models using machine learning, with 27 main items (comprising 52 checklist subitems) addressing model transparency, reproducibility, and validation rigor.[7] The CONSORT-AI extension (2020) established 14 new reporting items for randomized controlled trials evaluating AI interventions, emphasizing description of AI integration, input data handling, and human-AI interaction.[8] The companion SPIRIT-AI guideline addresses trial protocol reporting.[9] The STARD-AI extension for diagnostic accuracy studies is currently in development,[10] and the Checklist for AI in Medical Imaging (CLAIM) was initially published in 2020 and updated in 2024, providing specialty-specific guidance for radiology applications.[11,12]

Gap Analysis

These frameworks represent substantial progress but remain insufficient for sleep medicine applications. Spitschan et al assessed journal-level transparency policies in 28 sleep and chronobiology journals using the TOP Factor, finding median scores of only 2.5 out of 29 points, suggesting limited adoption of reporting standards.[13] More recently, Collins

https://doi.org/10.2147/NSS.S602114
Received: 8 February 2026
Accepted: 23 February 2026
Published: 26 February 2026

et al found that open science practices in prognostic model studies using machine learning need substantial improvement, demonstrating that transparency gaps persist even in newer research domains.[14] Sleep scoring poses unique challenges that current guidelines do not adequately address. The concept of "hypnodensity", quantifying sleep-stage ambiguity via probability distributions rather than discrete classifications, requires reporting standards for uncertainty quantification.[1]

Additionally, the acceptable performance threshold for automated sleep scoring differs from other medical AI applications due to substantial inter-rater variability among expert scorers. While overall agreement reaches 76%, this masks considerable disagreement at specific stages, with N1 showing only fair agreement ($\kappa < 0.30$) and N3 moderate agreement ($\kappa = 0.50–0.65$), establishing a performance ceiling that must be acknowledged when evaluating automated systems.[15,16]

Sleep-Specific Challenges Requiring Specialized Reporting Standards

Inter-Scorer Variability as a Fundamental Limitation

Unlike diagnostic imaging, where ground truth can be established through biopsy or surgical findings, sleep staging lacks an objective gold standard. Danker-Hopfe et al demonstrated overall inter-rater agreement of 80.6–82.0% (Cohen's kappa 0.68–0.76) depending on scoring rules,[17] while more recent studies involving 6–12 scorers report that unanimous agreement occurred in only 32–46% of epochs, decreasing systematically as the number of scorers increased.[1,18] This substantial variability creates a performance ceiling for automated systems trained using supervised learning, as training labels inherently contain scorer inconsistency. Reporting standards must therefore require explicit acknowledgment of this limitation and documentation of how training data were scored, including the number of scorers, their experience levels, and inter-scorer agreement statistics.

Beyond inter-scorer disagreement, the quality of polysomnographic recordings varies significantly across clinical settings. Preprocessing choices, such as filtering parameters, artifact reduction and rejection criteria, and handling of poor signal quality, are rarely reported despite profoundly affecting algorithm performance. No standardized preprocessing pipeline exists for sleep data, yet these decisions directly impact model reproducibility and clinical applicability.[19] Studies should also report whether preprocessing was performed as a standalone step or integrated within the model pipeline, as this distinction affects reproducibility and deployment behavior. Studies routinely fail to specify whether algorithms were trained on pristine laboratory recordings or real-world data with movement artifacts, electrode disconnections, and variable signal quality. This omission is particularly concerning because algorithms trained on high-quality data can experience substantial performance degradation when deployed in typical clinical environments where signal degradation is common.

The Hypnodensity Paradigm Shift

Traditional hypnograms assign a single definitive sleep stage to each 30-second epoch, oversimplifying the underlying ambiguity. Of note, the 30-second epoch has no real physiological basis; it is a historical remnant of paper-based polysomnography, where a standard 30-cm page recorded at a speed of 10 mm/s produced exactly 30 seconds of data.[20,21] AI systems are not bound by this convention and can analyze shorter or overlapping windows to capture brief events that the standard epoch may miss,[21] making it important for validation studies to report the epoch duration and overlap strategy used. Recent work demonstrates that AI systems can quantify prediction uncertainty, generating "hypnodensity" charts that display sleep-stage probabilities rather than categorical assignments.[1,22] Bechny et al showed that long short-term memory (LSTM) can identify uncertain predictions with AUROC (Area Under the Receiver Operating Characteristic Curve) values ranging from 82.5% to 85.7%,[22] enabling physicians to efficiently review only the most ambiguous epochs (approximately 26% for in-domain and up to 29% for out-of-domain data) to achieve 90% agreement. These approaches demonstrate the value of uncertainty quantification in sleep scoring, though further research is needed to establish standardized reporting frameworks for clinical implementation.

Algorithm Transparency and Clinical Integration

Published studies on AI-based sleep scoring provide insufficient detail regarding preprocessing steps, signal quality assessment, artifact handling, and clinical workflow integration. Xu et al documented widespread inconsistencies across AI sleep-scoring publications, with algorithms described using vastly different methodologies.[23] Studies routinely fail to specify algorithm versions, preprocessing protocols, or minimum signal quality thresholds, making independent replication nearly impossible.

Sleep medicine faces unique data challenges that general AI-reporting guidelines inadequately address. Polysomnography recordings exhibit night-to-night variability that may affect diagnostic classification, particularly in mild-to-moderate OSA, where single-night misclassification rates of 20–50% have been reported across both PSG-based and sensor-based multinight studies.[24,25] AI-based scoring systems must account for this source of diagnostic uncertainty, and validation studies should report how algorithm performance was assessed in the context of such inherent night-to-night variation. Manual scoring remains inherently subjective; inter-rater agreement for stage N1 falls as low as 63%, while REM sleep achieves 90.5% concordance.[26] Despite these challenges, no sleep-specific reporting standards exist. The CONSORT-AI and SPIRIT-AI extensions provide general frameworks[8] but do not address polysomnographic signal processing or sleep-stage ambiguity.

Dataset characteristics frequently remain undisclosed. Researchers often neglect to report demographic composition, sleep disorder prevalence, or the specifications of recording equipment. These details determine whether an algorithm can generalize beyond its training environment.[2] Among 230 AI sleep medicine studies, only 55% met high methodological rigor standards.[2] Reproducibility suffers most when implementation details go unreported.

Critical elements are frequently unreported: algorithm version control, handling of poor-quality input data, minimum signal quality requirements, procedures for human oversight, and performance documentation across diverse populations. Which predictions trigger technologist review? At what threshold does the algorithm fail to score an epoch? These questions remain unanswered despite growing awareness that algorithms may perform differently across demographic groups. Table 1 details essential reporting standards across seven domains, while Figure 1 translates these standards into a practical evaluation checklist for clinicians.

Proposed Framework: Essential Components for AI Sleep Reporting Standards

Algorithm Development and Training

Sleep-specific reporting must require comprehensive documentation of training data characteristics beyond standard demographic variables. Studies should report: (1) source datasets and their inter-scorer agreement statistics; (2) representation across sleep disorders, with performance stratified by diagnosis (Jahrami et al, 2025 demonstrate differential performance across insomnia, OSA, and narcolepsy)[2]; 3) sleep stage distribution in training data, with class imbalance ratios for underrepresented stages like N1 where agreement is lowest. Studies should report mitigation techniques such as reweighting or resampling; (4) handling of ambiguous epochs or "unknown" labels; and (5) whether training included multiple scoring per recording to capture variability.

Model architecture reporting should specify: the input signals used (EEG derivations, EOG, EMG, respiratory signals), sampling rates and preprocessing methods, model complexity (parameter count), training objective (loss function), whether models incorporate temporal context (bidirectional analysis), and how multi-channel information is aggregated. Studies should also document the use of data augmentation strategies (eg, time-shifting, noise injection) when employed. The concept of "development environment" requires particular attention; performance often degrades when models encounter test data from different acquisition systems, electrode montages, or clinical populations than those represented during training.[31]

Critically, studies must describe how they prevented data leakage by ensuring complete subject-level separation; all data from each patient must belong to only one set (training, validation, or test). In addition, studies must report how missing data and poor-quality epochs were handled; the common practice of simply excluding such epochs creates selection bias, potentially inflating reported accuracy while reducing real-world generalizability.[19]

Table 1 Essential Reporting Standards for AI-Based Automated Sleep Scoring

Domain	Current Gap	Proposed Reporting Requirement	Rationale	Citation
Preprocessing & Signal Quality	Preprocessing steps and artifact handling are inconsistently reported	Report all preprocessing steps: filtering parameters (high-pass/low-pass Hz), normalization method (min-max vs z-score), downsampling rates, epoch exclusion criteria (% missing data), artifact reduction/rejection methods (manual vs automated), and minimum signal quality threshold	Real-world recordings contain artifacts; preprocessing choices affect reproducibility and clinical applicability. No standardized pipeline exists; inconsistent practices prevent replication	[3,19]
Training Data & Inter-Scorer Reliability	Dataset characteristics and label quality are underreported	Report: dataset size (# patients, # epochs), population characteristics (age, sex, OSA severity), number of scorers, inter-scorer reliability (κ value overall and stage-specific), public vs private dataset	Ground truth labels vary by scorer (κ=0.76 overall; κ=0.24 for N1); population characteristics affect generalizability	[19,27]
Comparison Benchmark	Often compared to a single scorer	Compare to inter-scorer agreement range (not single "ground truth"); report κ = 0.68–0.76 (corresponding to 76–82% agreement); acknowledge that some epochs are inherently ambiguous (up to 28%)	Claims of "superhuman performance" are misleading when human agreement is 76–82%	[15,27,28]
Performance Reporting	Often report only overall accuracy	Report stage-specific metrics (accuracy, sensitivity, specificity, F1-score for W, N1, N2, N3, REM); stratify by age, sex, disease severity	Performance varies dramatically by stage (N1 worst: κ=0.24); accuracy deteriorates in elderly, male patients, and higher AHI	[15,21]
Uncertainty Quantification	Binary predictions only	Report hypnodensity/confidence scores; identify ambiguous epochs	Sleep staging is inherently ambiguous; probability distributions are more clinically useful than discrete labels	[1,18]
External Validation	Often validated only on single-center data	Require external validation on diverse populations (age, sex, ethnicity, OSA severity) from different institutions with different scorers and equipment; report performance stratified by subgroups	Single-center data may not generalize; only 0.85% of AI sleep studies meet rigorous validation criteria	[5,29,30]
Clinical Implementation Pathway	Integration workflow unclear	Specify epochs requiring human review, review time, human-AI agreement rates, error escalation procedures	Clinical utility depends on efficient human oversight; transparency needed for implementation	[3,22]

Validation and Performance Reporting

Sleep medicine requires specialized performance metrics beyond accuracy. Comprehensive reporting should include: (1) epoch-wise and subject-wise metrics separately, as averaging can mask poor performance on individual recordings; (2) stage-specific sensitivity, precision, and F1-score for each sleep stage, particularly for clinically relevant stages like REM and N3, with a per-class confusion matrix to identify systematic misclassification patterns; (3) performance on ambiguous epochs where expert scorers disagree, quantified through hypnodensity or confidence trends[1,18]; 4) distance metrics showing alignment between predicted probabilities and actual accuracy; and (5) comparison against inter-scorer agreement benchmarks, acknowledging that most deep learning models approach but rarely exceed kappa values between experienced human scorers, making claims of "superhuman" performance misleading.[29]

External validation using independent cohorts remains uncommon but essential for robustness claims. Studies should explicitly report whether validation data derive from the same institution as training data, whether the same scorers

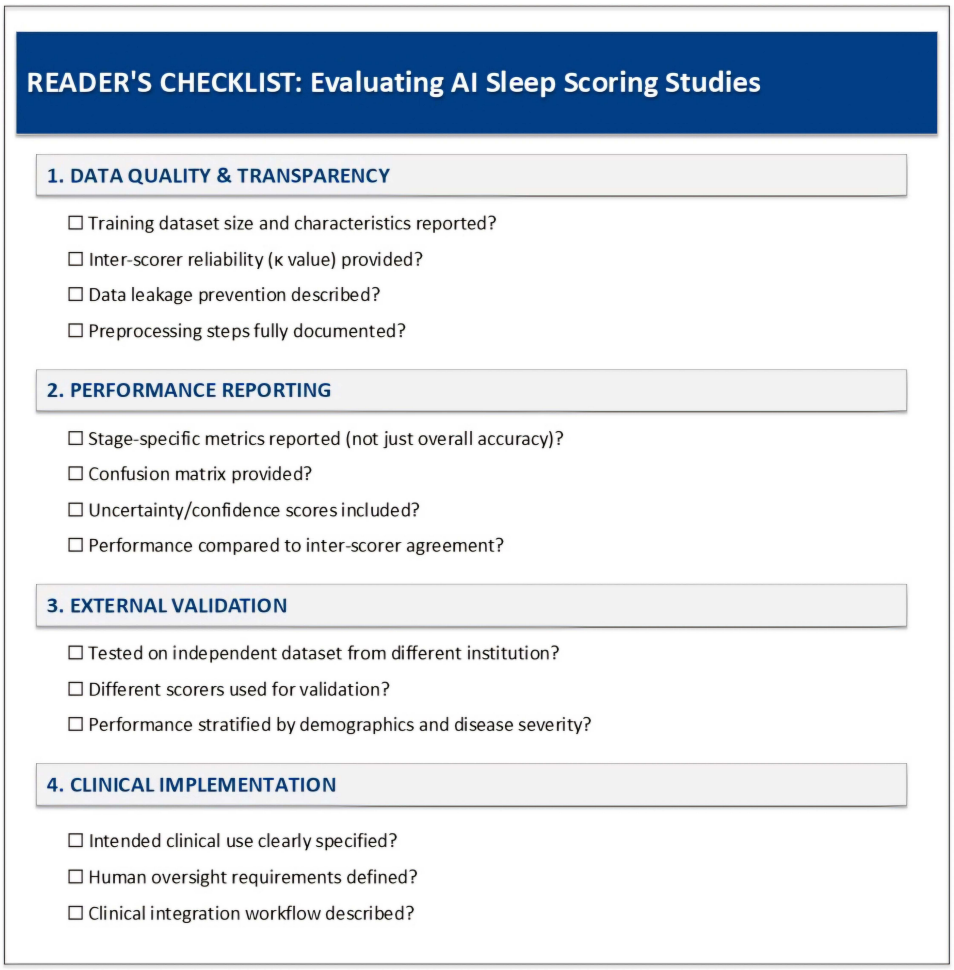

Figure 1 Reader's Checklist for Evaluating AI Sleep Scoring Studies.

annotated both sets, and whether performance is stratified by relevant demographic and clinical variables, as inter-scorer reliability and algorithmic accuracy have been shown to deteriorate in several clinically complex populations, including those with fragmented sleep and multiple comorbidities.[21,29]

Clinical Implementation and Human-AI Interaction

Reporting must address the integration pathway from algorithm to clinical decision. Essential elements include: (1) the intended clinical use, screening, diagnostic support, or autonomous scoring; (2) requirements for physician review, including specific criteria triggering manual verification; (3) time required for algorithm-assisted scoring versus traditional manual scoring; (4) error analysis detailing failure modes and their clinical implications; (5) procedures for handling edge cases such as pediatric recordings, unusual sleep disorders, or medication effects; and (6) mechanisms for ongoing quality assurance and algorithm monitoring post-deployment.

Implementation Strategy and Call to Action

Developing robust reporting standards requires collaboration among sleep medicine professional societies (AASM, ESRS, WASM), AI researchers, journal editors, and regulatory bodies. The updated AASM statement on AI highlights the importance of transparent validation and harmonized methods. However, it still does not impose detailed, enforceable reporting requirements on developers or authors.[3] Sleep journals should therefore adopt structured reporting checklists as

https://doi.org/10.2147/NSS.S602114

a condition for submission and peer review, mirroring the experience in radiology and pathology where such frameworks have accelerated the uptake of reproducible AI research.

Phased Implementation

A practical path forward would be: (1) immediate adoption of existing frameworks (TRIPOD+AI, CONSORT-AI) with sleep-specific supplements; (2) convening an international working group to develop comprehensive reporting guidelines for sleep AI; (3) pilot testing these proposed standards in prospective studies; and (4) iterative refinement based on real-world experience. This phased approach allows the field to build on established methods while addressing the particular requirements of sleep medicine.

Regulatory Implications

Standardized reporting will facilitate regulatory review processes. The FDA has cleared multiple sleep-scoring algorithms (models), but the evaluation criteria remain opaque. Harmonized reporting standards would enable regulators to systematically assess validation rigor, potential biases, and real-world performance, ultimately accelerating safe clinical adoption.

Conclusion

The integration of AI into sleep medicine offers remarkable opportunities but demands equal responsibility. Automated sleep scoring has reached a level of maturity where clinical implementation is not only feasible but increasingly inevitable, given workforce constraints and the expanding demand for sleep services. However, the current documentation crisis threatens to undermine these advances through irreproducible research, premature adoption of inadequately validated systems, and erosion of trust in algorithmic tools.

Sleep medicine must choose: allow heterogeneous, incomplete reporting to continue or proactively establish comprehensive, field-specific standards for responsible innovation. Inter-scorer variability creates a performance ceiling for AI systems. Future progress, therefore, depends on transparency, reproducibility, and rigorous real-world validation rather than incremental algorithmic improvements, making standardized reporting essential before poor practices become entrenched. The moment to act is now, before poor practices become further entrenched. Developing these standards represents not a barrier to innovation but the foundation for sustainable, evidence-based integration of AI into sleep medicine practice.

Data Sharing Statement

Data sharing is not applicable as no new data was generated for this work.

Acknowledgments

The authors have no acknowledgments.

Author Contributions

Ahmed S. BaHammam: Conceptualization, Writing – original draft preparation, Writing – review and editing, and Supervision.

Malak A. Almarshad: Validation, Writing – original draft preparation, Writing – review and editing.

The authors give final approval of the version to be published; have agreed on the journal to which the article has been submitted; and have agreed to be accountable for all aspects of the work.

Funding

The Strategic Technologies Program of the National Plan for Sciences and Technology and Innovation in the Kingdom of Saudi Arabia, Riyadh, Saudi Arabia (MED511-02-08).

Disclosure

The author reports no conflicts of interest in this work. Grammarly assisted with grammar correction during the preparation of this manuscript.

References

1. Bakker JP, Ross M, Cerny A, et al. Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring. *Sleep*. 2023;46(2):zsac154.
2. Jahrami H, Husain W, Trabelsi K, et al. Artificial intelligence and sleep medicine II: a scoping review of applications, advancements, and future directions. *Sleep Med Rev*. 2025;85:102212.
3. Oks M, Sachdeva R, Davenport MA, et al. Artificial intelligence in sleep medicine: an updated American academy of sleep medicine position statement. *J Clin Sleep Med*. 2025;21(11):1953–1955. doi:10.5664/jcsm.11832
4. 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
5. Alattar M, Govind A, Mainali S. Artificial intelligence models for the automation of standard diagnostics in sleep medicine-a systematic review. *Bioengineering*. 2024;11(3). doi:10.3390/bioengineering11030206
6. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689.
7. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
8. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*. 2020;370:m3164. doi:10.1136/bmj.m3164
9. Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Med*. 2020;26(9):1351–1363. doi:10.1038/s41591-020-1037-7
10. Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11(6):e047709. doi:10.1136/bmjopen-2020-047709
11. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029. doi:10.1148/ryai.2020200029
12. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300. doi:10.1148/ryai.240300
13. Spitschan M, Schmidt MH, Blume C. Transparency and open science reporting guidelines in sleep research and chronobiology journals. *bioRxiv*. 2020. doi:10.1101/2020.09.26.314658
14. Collins GS, Whittle R, Bullock GS, et al. Open science practices need substantial improvement in prognostic model studies in oncology using machine learning. *J Clin Epidemiol*. 2024;165:111199. doi:10.1016/j.jclinepi.2023.10.015
15. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: a meta-analysis. *J Clin Sleep Med*. 2022;18(1):193–202. doi:10.5664/jcsm.9538
16. Younes M, Kuna ST, Pack AI, et al. Reliability of the American academy of sleep medicine rules for assessing sleep depth in clinical practice. *J Clin Sleep Med*. 2018;14(2):205–213. doi:10.5664/jcsm.6934
17. Danker-Hopfe H, Anderer P, Zeitlhofer J, 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. doi:10.1111/j.1365-2869.2008.00700.x
18. Stephansen JB, Olesen AN, Olsen M, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun*. 2018;9(1):5229. doi:10.1038/s41467-018-07229-3
19. Almarshad MA, Islam S, Bahammam S, Al-Ahmadi S, BaHammam AS. Polysomnography raw data extraction, exploration, and preprocessing. In: Berry RB, Pardalos P, Xiaochen X, editors. *Handbook of AI and Data Sciences for Sleep Disorders*. Cham: Springer Nature Switzerland AG; 2024:45–65.
20. de Chazal P, Mazzotti DR, Cistulli PA. Automated sleep staging algorithms: have we reached the performance limit due to manual scoring? *Sleep*. 2022;45(9). doi:10.1093/sleep/zsac159
21. Korkalainen H, Leppanen T, Duce B, et al. Detailed assessment of sleep architecture with deep learning and shorter epoch-to-epoch duration reveals sleep fragmentation of patients with obstructive sleep apnea. *IEEE J Biomed Health Inform*. 2021;25(7):2567–2574. doi:10.1109/JBHI.2020.3043507
22. Bechny M, Monachino G, Fiorillo L, et al. Bridging AI and clinical practice: integrating automated sleep scoring algorithm with uncertainty-guided physician review. *Nat Sci Sleep*. 2024;16:555–572. doi:10.2147/NSS.S455649
23. Xu S, Faust O, Seoni S, et al. A review of automated sleep disorder detection. *Comput Biol Med*. 2022;150:106100.
24. Lechat B, Naik G, Reynolds A, et al. Multinight prevalence, variability, and diagnostic misclassification of obstructive sleep apnea. *Am J Respir Crit Care Med*. 2022;205(5):563–569. doi:10.1164/rccm.202107-1761OC
25. Punjabi NM, Patil S, Crainiceanu C, Aurora RN. Variability and misclassification of sleep apnea severity based on multi-night testing. *Chest*. 2020;158(1):365–373. doi:10.1016/j.chest.2020.01.039
26. Alsolai H, Qureshi S, Iqbal SMZ, et al. A systematic review of literature on automated sleep scoring. *IEEE Access*. 2022;10:79419–79443. doi:10.1109/ACCESS.2022.3194145
27. 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):81–87. doi:10.5664/jcsm.2350
28. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. *J Clin Sleep Med*. 2016;12(6):885–894. doi:10.5664/jcsm.5894
29. Cesari M, Stefani A, Penzel T, et al. Interrater sleep stage scoring reliability between manual scoring from two European sleep centers and automatic scoring performed by the artificial intelligence-based Stanford-STAGES algorithm. *J Clin Sleep Med*. 2021;17(6):1237–1247.

30. Anderson AW, Marinovich ML, Houssami N, et al. Independent external validation of artificial intelligence algorithms for automated interpretation of screening mammography: a systematic review. *J Am Coll Radiol*. 2022;19(2 Pt A):259–273. doi:10.1016/j.jacr.2021.11.008
31. Alvarez-Estevez D, Rijsman RM. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLoS One*. 2021;16(8): e0256111. doi:10.1371/journal.pone.0256111