Predicting Future Cognitive Impairment Using Whole-Night Sleep-Derived Outcomes

Franz Ehrlich[1,2], Sara Bäcker[1], Alex Schmidt[3], Nicolai Spicher[4], Dagmar Krefting[3], Miriam Goldammer[1], Philip Hempel[3]

[1] Institute for Medical Informatics and Biometry, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany
[2] Institute of Biomedical Engineering, TUD Dresden University of Technology, Dresden, Germany
[3] Department of Medical Informatics, University Medical Center Göttingen, Germany
[4] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

Abstract

Predicting cognitive impairment before clinical diagnosis may support earlier assessment, monitoring, and intervention. We, team OOM_finder, investigated whether overnight polysomnography (PSG) contains information predictive of cognitive impairment diagnosed 1–6 years later. We compared three levels of PSG representation: strongly regularized logistic regression using demographic and sleep-summary features, a hierarchical transformer operating on temporally resolved algorithmic sleep annotations, and a pretrained multimodal encoder of raw physiological signals. The raw-signal approach achieved the strongest internal discrimination with an age-conditioned AUROC of 0.804. However, this advantage did not transfer to the official hidden validation set, where the compact sleep-summary model achieved the highest age-conditioned AUROC of 0.703 and a reward of 0.157. The annotation-based and raw-signal models achieved hidden-validation age-conditioned AUROCs of 0.562 and 0.612, respectively. These results suggest that increased representational complexity does not necessarily improve cross-site generalization and that compact, clinically interpretable sleep features may help predict future cognitive impairment.

1. Introduction

Cognitive impairment can affect memory, attention, executive function, language, and other cognitive abilities that are essential for independent daily life. As impairment progresses, the need for assistance, supervision, and long-term care increases, with consequences extending beyond the affected individual to family members and other caregivers. Dementia represents one of the most severe manifestations of cognitive decline and is a major and increasing cause of disability and dependency worldwide, with approximately 57 million people affected in 2021 [1, 2]. Although Alzheimer's disease (AD) is the most common cause of dementia, cognitive decline can arise from a broad range of pathological processes or is the consequence of long-term alcohol abuse[1].

A major clinical challenge is that effective treatment options remain limited, while opportunities for intervention are generally greatest before substantial and irreversible functional decline has occurred. Recently introduced disease-modifying therapies for AD target selected patients in early symptomatic stages and slow rather than reverse disease progression [3, 4]. For many other causes of cognitive impairment, no disease-modifying therapy is currently available. However, early identification is essential even when an immediate causal treatment cannot be offered: Recognizing individuals at increased risk can motivate a more targeted diagnostic work-up, identify potentially modifiable or treatable contributors, enable closer longitudinal surveillance, and allow identifying patients who are eligible for etiology-specific therapies.

Achieving early risk stratification remains difficult because cognitive impairment is often recognized only after cognitive, behavioral, or functional changes have become clinically apparent, whereas the underlying disease processes may begin years earlier. Although disease-specific biomarkers can improve diagnostic precision, specialized neuroimaging, cerebrospinal-fluid analysis, and molecular testing are not equally accessible or appropriate for broad screening. This motivates the search for routinely obtainable physiological markers that may indicate increased vulnerability before overt clinical manifestation.

Sleep is particularly attractive in this context because it integrates neurological, respiratory, cardiovascular, and autonomic physiology. Alterations in sleep architecture, fragmentation, rapid-eye-movement (REM) sleep,

sleep-disordered breathing, and sleep microstructure have been associated with subsequent cognitive decline [5, 6]. Polysomnography (PSG) captures these processes simultaneously across an entire night, including sleep-stage dynamics, arousals, respiratory events, oxygen saturation, cardiac activity, and movement. The combination of these signals may therefore contain information about future cognitive risk.

The 2026 George B. Moody PhysioNet Challenge investigates whether future cognitive impairment can be predicted $1 - 6$ years before clinical diagnosis [7, 8].

2. Methods

We submitted three approaches with varying model complexity to predict future cognitive impairment from overnight PSG and basic demographic data. **Approach 1 (A1)** employs a lightweight logistic regression model using demographic variables and PSG features derived from algorithmic sleep annotations. **Approach 2 (A2)** uses a hierarchical transformer architecture on the derived algorithmic sleep annotations. **Approach 3 (A3)** leverages a pretrained multimodal PSG encoder that processes raw physiological signals, followed by a downstream classifier with demographic fusion via feature-wise linear modulation.

A1: Demographic + PSG-feature Logistic Regression

This entry takes a lightweight approach, relying on demographic variables and PSG features derived from the CAISR [9] calculated sleep annotations, rather than using the raw PSG signals.

Demographic features (13 dimensions): age, BMI, and one-hot encoded sex (female, male and other-unknown), self-reported race (Asian, Black, other, unavailable, white) and ethnicity (Hispanic, not Hispanic, unavailable).

PSG features (36 dimensions) derived from the CAISR annotation channels:

- Event indices (apnea-hypopnoea index, arousal index and limb movement index) as events per hour of sleep
- Sleep stage percentages (W, N1, N2, N3 and REM sleep) and sleep efficiency
- The mean and standard deviation of the per-epoch model probability traces for each sleep stage
- Total sleep time, REM latency and stage transition index as events per hour of sleep
- Arousal count, mean/maximum duration and percentage of sleep time with arousal
- Limb movement/PLM index and mean/maximum PLM duration
- Apnea index, AHI and RDI, computed from discrete respiratory events, restricted to sleep epochs.

Missing values (e.g. when a channel is unavailable) are imputed using the median of the training set. The features are standardised and then passed to an elastic-net-regularised logistic regression using the `scikit-learn` `saga` solver with an ℓ_1 ratio of 0.4 and a C value of 0.03. The class weights are balanced to correct for the low prevalence of cognitive impairment.

A2: Hierarchical Transformer with Algorithmic Sleep Annotations

Our second approach uses a hierarchical transformer architecture to predict cognitive impairment based on sleep annotations generated by the CAISR model [9].

The model ingests multiple annotation streams at various temporal resolutions:

- Sleep stage class and probabilities at a 30-second resolution
- Arousal class and probabilities at a 0.5-second resolution
- Respiratory event class at one-second resolution
- Limb movement class at 1-second resolution

The architecture comprises three micro-encoders that process each annotation stream in 30-second intervals. This is followed by a global transformer encoder that models the entire sequence of the night. Age and sex are explicitly supplied to the final cognitive impairment prediction head.

The training objective combines binary cross-entropy with an age-matched pairwise ranking loss. Positive/negative pairs contribute to the ranking term if they are no more than two years apart in age, in line with the comparison rule of the Challenge's primary age-conditioned AUROC metric. The final model is trained for a fixed four epochs on all labelled training data using the median best epoch identified during 5-fold cross-validation.

A3: Pretrained Multimodal PSG Encoder + Downstream Classifier

Our third approach builds on a pretrained multimodal PSG sleep scoring model that was trained to classify five sleep stages, binary arousal, binary leg movement, and five classes of respiratory events (obstructive, central and mixed apnea, hypopnea, respiratory effort related arousal) from raw signals on seven public PSG datasets (SHHS, MESA, MrOS, PhysioNet Challenge 2018, CFS, WSC, and SOF).

All available PSG channels are resampled to 64 Hz, and grouped by modality, with a maximum of four EEG channels, three EOG channels, six EMG channels, three respiratory effort channels, four airflow channels, two SpO_2 channels, two ECG channels, and one PPG channel per group. Named leads such as "C4-M1" are prioritised and

the remaining channels are dropped. Each channel group is downsampled by a shared convolutional feature extraction through a stack of eight layers (with the following feature dimensions: 64, 64, 64, 64, 64, 64, 64, 128). Each layer comprises three parallel convolutions with kernel sizes of 7, 15, and 35, which capture multiple temporal scales. The resulting embeddings from the pre-trained encoder are frozen and used as fixed input features for the downstream classifier.

The embedding sequence is processed by a downstream classifier, which is a four-layer, state-space, bidirectional temporal encoder with a state size of 16 and a dropout rate of 0.3, implemented as an FFT-based causal convolution. Demographic variables are fused with the pooled signal representation via feature-wise linear modulation (FiLM). A learned attention-pooling layer then reduces the sequence to a single embedding. This is followed by a linear classification head, which is trained using binary cross-entropy, AdamW and a cosine-annealed learning rate (with a peak of 0.004). The batch size is set to 32 and the training continues for up to 50 epochs.

We evaluated two data-splitting strategies for this entry:
1. **Patient-wise holdout**: 6581 patients, split into training data (75%) and validation data (25%).
2. **Leave-one-site-out cross-validation**: patients grouped by site into three folds, each fold holding out one site for validation. The final prediction is the average of the three fold-specific models. This was motivated by the observation that the patient-wise split did not transfer well to the unseen hidden validation site.

3. Results

3.1. Internal validation

Table 1 summarizes the internal validation performance of the three approaches. The patient-wise raw-signal model (A3-P), our most complex approach, achieved the highest age-conditioned AUROC (AUROC-age: 0.804) and conventional AUROC (0.878). In comparison, A1 and A2 achieved AUROC-age values of 0.662 and 0.673, respectively. F1 scores were similar across A1, A2, and A3-P.

Table 1. Internal validation performance. Values are averaged across folds where applicable. Best values are shown in bold.

Model	AUROC-age	AUROC	F1
A1	0.662	0.801	0.291
A2	0.673	0.793	0.289
A3-P	**0.804**	**0.878**	0.286
A3-S	0.634	–	–

These results indicate that the increased representational capacity of A3-P was beneficial during internal validation, particularly for the discrimination-based metrics used for model selection.

3.2. Official hidden-validation results

Performance on the official hidden validation set showed a different ordering of the approaches (Table 2). Despite its comparatively simple feature representation, A1 achieved the highest reward (0.157) and the highest AUROC-age (0.703). In contrast, the internal advantage of the more complex A3-P model did not transfer to the hidden data: its AUROC-age decreased from 0.804 internally to 0.612 on the official validation set. A2 similarly decreased from an internal AUROC-age of 0.673 to 0.562.

Table 2. Performance on the official hidden validation set. Best available values are shown in bold.

Model	Reward	AUROC-age	AUROC	AUPRC	F1
A1	**0.157**	**0.703**	TBD	TBD	TBD
A2	0.008	0.562	0.695	0.141	0.000
A3-P	0.047	0.612	0.785	0.208	0.226
A3-S	0.054	0.569	0.731	0.226	0.228

A3-P retained the highest conventional AUROC among the entries for which this metric was available, but was not superior on the Challenge's primary age-conditioned metric. The leave-one-site-out variant A3-S also did not improve AUROC-age over the patient-wise A3-P model (0.569 vs. 0.612), despite slightly higher reward, AUPRC, and F1 scores. Overall, the performance ranking on the hidden data did not follow model complexity: the strongest internal discrimination was obtained with A3-P, whereas the lightweight A1 model achieved the best hidden-validation AUROC-age.

4. Discussion

We compared representations of overnight PSG ranging from compact clinical sleep summaries to temporally resolved algorithmic annotations and learned raw-signal representations. The main finding was that greater representational complexity improved internal discrimination but did not translate into superior performance on the hidden validation set. The raw-signal model achieved the highest internal AUROC-age, whereas the strongly regularized summary-feature model achieved the highest hidden AUROC-age and reward.

This difference suggests that robustness across clinical sites may be more important than maximizing representation complexity. Compact sleep features provide relatively standardized physiological summaries and may be less sensitive to differences in recording systems, channel configurations, preprocessing, and patient populations.

Similar losses of performance under institutional distribution shift have been reported for machine-learning models in other clinical domains [10]. Algorithmic annotations provide a clinically interpretable intermediate representation that retains temporal sleep information while reducing direct dependence on raw waveforms. Nevertheless, neither the annotation-based model nor the raw-signal model generalized better than the summary-based approach. The leave-one-site-out strategy likewise did not improve AUROC-age, indicating that the tested form of site-wise training was insufficient to overcome the observed distribution shift.

Demographic information also contributed complementary predictive information, including sex in addition to the strong association between age and future cognitive impairment. This is consistent with reported sex-related differences in cognitive decline and neurodegenerative disease phenotypes [11]. At the same time, the heterogeneous etiologies underlying the prediction target may limit the existence of a single transferable sleep phenotype. Together, these factors may hinder generalization and emphasize the importance of the age-conditioned evaluation used in the Challenge [7, 8]. Overall, our results indicate that more detailed physiological representations do not necessarily yield more robust predictions in the investigated setting. For future cognitive-risk prediction from heterogeneous clinical PSG, strong regularization and clinically meaningful abstraction may be as important as model capacity.

Code Availability

All Challenge submissions were made under the team name `OOM_finder`. The code of all approaches are available on GitLab `https://gitlab.com/sleep-is-all-you-need/challenges/physionet2026-oomfinder` under MIT license. The pretrained models are also available in the repository under GNU General Public License.

Acknowledgments

This research was partly funded by the Federal Ministry of Research, Technology and Space under the funding code 01ZZ2324F. The authors gratefully acknowledge the computing time made available to them on the high-performance computer at the NHR Center of TU Dresden and NHR-Nord@Göttingen.

References

[1] World Health Organization. Dementia, July 2026. WHO Fact Sheet, updated 3 July 2026, accessed 25 August 2026.

[2] Livingston G, Huntley J, Liu KY, Costafreda SG, Selbæk G, et al. Dementia prevention, intervention, and care: 2024 report of the lancet standing commission. Lancet 2024; 404(10452):572–628.

[3] van Dyck CH, Swanson CJ, Aisen P, Bateman RJ, Chen C, Gee M, Kanekiyo M, Li D, Reyderman L, Cohen S, et al. Lecanemab in early alzheimer's disease. N Engl J Med 2023;388(1):9–21.

[4] Sims JR, Zimmer JA, Evans CD, Lu M, Ardayfio P, Sparks J, Wessels AM, Shcherbinin S, Wang H, Monkul Nery ES, et al. Donanemab in early symptomatic alzheimer disease: The trailblazer-alz 2 randomized clinical trial. JAMA 2023; 330(6):512–527.

[5] Carpi M, Fernandes M, Mercuri NB, Liguori C. Sleep biomarkers for predicting cognitive decline and alzheimer's disease: A systematic review of longitudinal studies. J Alzheimers Dis 2024;97(1):121–143.

[6] Zhang Y, Ren R, Yang L, Zhang H, Shi Y, Okhravi HR, Vitiello MV, Sanford LD, Tang X. Sleep in alzheimer's disease: a systematic review and meta-analysis of polysomnographic findings. Transl Psychiatry 2022;12(1):136.

[7] Reyna MA, Weigle AT, Li Q, Koscova Z, Sun H, Wen S, Katwa UA, Thomas RJ, Hwang D, Trotti LM, Mignot E, Sameni R, Nasiri S, Westover MB, Clifford GD. Screening for cognitive impairment during sleep studies: The george b. moody PhysioNet challenge 2026. In Computing in Cardiology 2026. 2026; .

[8] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[9] Nasiri S, Ganglberger W, Nassi T, Meulenbrugge EJ, Moura Junior V, Ghanta M, Gupta A, Stone KL, Kjaer MR, Sum-Ping O, Mignot E, Hwang D, Trotti LM, Clifford GD, Katwa U, Sun H, Thomas RJ, Westover MB. Caisr: achieving human-level performance in automated sleep analysis across all clinical sleep metrics. Sleep 2025;48(8):zsaf134.

[10] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Medicine 2018; 15(11):e1002683.

[11] Ferretti MT, Iulita MF, Cavedo E, Chiesa PA, Schumacher Dimech A, Santuccione Chadha A, Baracchi F, Girouard H, Misoch S, Giacobini E, Depypere H, Hampel H. Sex differences in alzheimer disease—the gateway to precision medicine. Nature Reviews Neurology 2018; 14(8):457–469.

Address for correspondence:

Franz Ehrlich
Institute for Medical Informatics and Biometry
Fetscherstraße 74, 01307 Dresden
franz.ehrlich@tu-dresden.de