Under-Mattress Temporal Sensing for Next-Day Agitation Risk Scoring in Dementia Wards

Zhen Liu[1], Marta Bono[1], Robbe Decloedt[2], Ajda Flisar[2],
Maarten Van Den Bossche[2,3,5,†], Maarten De Vos[1,4,†]
[1]STADIUS Center, Department of Electrical Engineering, KU Leuven, Leuven, Belgium
[2]Geriatric Psychiatry, University Psychiatric Center KU Leuven, Leuven, Belgium
[3]Neuropsychiatry, Department of Neurosciences, Leuven Brain Institute, KU Leuven, Leuven, Belgium
[4]Woman and Child, Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[5]Department of Psychiatry, Maastricht University Medical Centre+ (MUMC+), Maastricht, The Netherlands
[†]Both senior authors contributed equally to this study

***Abstract*—*Objective:* Agitation fluctuates over short time horizons in people living with dementia, yet continuous physiological information for anticipating next-day risk is limited. We assessed whether contactless under-mattress signals from the preceding night inform next-day agitation risk and whether preserving minute-level temporal structure improves performance over conventional nightly summaries. *Methods:* We analyzed 423 patient-nights from 65 subjects in a specialized hospital dementia unit using two under-mattress sensing systems. A unified four-paradigm benchmark compared nightly handcrafted summaries, three-period handcrafted features, full-night sequence modeling, and sliding-window multiple-instance learning. Source-specific preprocessing and five-fold patient-grouped cross-validation were used, with performance estimated from pooled out-of-fold predictions. Evaluation included discrimination, calibration, fixed-threshold metrics, and a comparison of period–signal attribution patterns across two temporal models. *Results:* Full-night sequence modeling achieved the highest discrimination (AUROC, 0.692; AUPRC, 0.849) and balanced accuracy (0.658). Both minute-level pipelines had higher AUROC than nightly summaries, but differences from three-period handcrafted features were uncertain. Cross-model attribution prioritized activity, heart rate, and respiratory rate during the core overnight period. Calibration remained limited. *Conclusion:* The preceding night's signals supported modest next-day risk discrimination, with minute-level temporal modeling outperforming nightly summaries. Prospective calibration and external validation are needed before use in individual care decisions. *Significance:* This patient-grouped benchmark identifies contactless overnight sensing as a promising biomedical engineering direction for agitation-risk research in hospitalized dementia cohorts.**

***Index Terms*—Dementia, Alzheimer's disease, agitation, sleep physiological signals, ballistocardiography, machine learning, multiple-instance learning, temporal representation, risk scoring.**

I. Introduction

Agitation is one of the most prevalent and clinically burdensome neuropsychiatric symptoms in people living with dementia (PLWD) [1]. It includes excessive motor activity, verbal agitation, and aggressive behavior associated with emotional distress [2]. In care homes, clinically significant agitation affects approximately 40% of residents with dementia [3]. Because pharmacological management has limited efficacy and substantial safety concerns [4]–[6], earlier identification of elevated risk could support proactive non-pharmacological care. Agitation varies over short time horizons, while routine clinical observations provide limited continuous information for estimating risk on the following day.

Night-time physiology is a plausible source of short-horizon agitation information, but its predictive value remains underspecified. Sleep disturbances are common in dementia and have been associated with next-day behavioral symptoms [7]–[10]. More specific longitudinal evidence suggests that sleep–agitation coupling varies across individuals [11], that sleep patterns and behavioral symptoms show day-to-day temporal interdependence [12], and that long-term heart rate variability measures are associated with agitation risk in Alzheimer's disease [13]. These findings motivate testing whether nocturnal cardiorespiratory and activity signals recorded during the preceding night are associated with agitation risk on the following day in dementia care.

Contactless under-mattress sensing offers a practical route to measure these signals without adding burden to PLWD or ward staff. Commercially available devices, including EMFIT QS® (EMFIT) and Withings Sleep Analyzer® (WSA), have been used in sleep and digital health studies [14]–[18]. These two sensors use pressure-sensitive under-mattress sensing to capture body movement and respiratory signals and derive cardiac measures from ballistocardiographic (BCG) signals. Independent evaluations against polysomnography and actigraphy support their utility for low-burden longitudinal monitoring of sleep physiology, particularly cardiorespiratory and movement-derived signals [15], [17]. Unobtrusive mattress-based BCG has also been used to estimate nocturnal awakening and sleep efficiency [19], while contactless radar has recovered cardiorespiratory, movement, and sleep-stage information in older people and people with neurodegenerative disorders [20]. Their relevance to agitation risk scoring lies in the within-night physiological trajectories they provide during routine overnight care.

Existing sensor-based studies have not established whether physiological signals recorded during the preceding night can predict agitation on the following day in a specialized hospital

dementia unit. Prior studies have focused on home monitoring, multimodal sensing, or measurements aggregated over several days [21], [22], which represent different prediction settings and time horizons. Among the most closely related predictive studies, Ramesh et al. [23] selected LightGBM for next-day prediction from daily aggregated sleep and physiological variables, whereas Bafaloukou et al. [18] selected LightGBM for agitation monitoring over 8-day periods using multimodal in-home variables. These studies compress sensor streams into engineered summary features or aggregate exposure variables, including minimum respiratory rate, mean heart rate, respiratory-rate variability, and awake ratio [18], [23]. Their findings demonstrate the predictive utility of engineered summaries but leave unclear whether within-night temporal structure adds information beyond nightly summaries. Sequence-aware models have exploited neighboring and longer-range temporal context in automatic sleep staging [24], [25], while multiple-instance learning (MIL) has been applied to long-term physiological recordings [26], [27]. However, whether fine-grained nocturnal structure supports prediction of a distinct next-day behavioral endpoint remains unclear in small repeated-measures hospital cohorts, where evaluation must account for event prevalence, threshold-dependent operating characteristics, device-source heterogeneity, and between-patient differences.

A related explanatory study analyzed an earlier subset from the same hospital monitoring program using mixed-effects models to examine associations between handcrafted nocturnal features and agitation occurrence and severity while accounting for repeated observations [28]. The present study addresses a distinct predictive question by evaluating held-out-patient next-day risk scoring and comparing engineered and time-series representations.

Accordingly, this study evaluates whether under-mattress signals recorded during the preceding night support cross-patient next-day agitation risk estimation. We compared four pipelines that represented the preceding night as whole-night handcrafted summaries, period-resolved handcrafted features, a complete minute-level sequence, or overlapping windows with MIL aggregation. Performance was evaluated using patient-grouped out-of-fold discrimination, calibration, and fixed-threshold metrics. We further examined whether period-specific handcrafted features and sliding-window modeling emphasized similar overnight periods and physiological signals. Source-stratified and higher-burden endpoint analyses examined whether performance patterns varied by sensor source and outcome definition.

II. Data and Methods

Figure 1 summarizes the study workflow, from under-mattress sensing and night-to-day label pairing to patient-grouped evaluation, pooled out-of-fold (OOF) analysis, calibration assessment, and cross-paradigm attribution.

A. Study Cohort and Outcome

This study analyzed 423 patient-nights from 65 patients, pairing overnight recordings from two passive under-mattress sensing systems (EMFIT and WSA) with next-day agitation assessments. Under the hospital monitoring protocol described by Davidoff et al. [29], continued sensor monitoring and agitation-label collection expanded the interim dataset analyzed by Liu et al. [28] (55 participants and 333 patient-nights) to the current cohort. The study protocol was approved by the Ethics Committee Research of University Hospitals Leuven (reference ID S62882). Written informed consent was obtained from participants with capacity or, when cognitive impairment precluded consent, from a legally authorized representative, with study information explained to participants in a manner appropriate to their cognitive abilities.

The two sensor systems provided heart rate (HR), respiratory rate (RR), bed activity (ACT), and sleep-stage streams at different temporal resolutions. The preprocessing pipeline harmonized these streams to 1-minute resolution and derived a minute-level quality channel from device-specific signal-availability and missingness patterns. Supplementary Section S1 describes the device-specific cleaning of the 1-minute sequences, including overnight-record construction, quality harmonization, missingness marking, and source-aware normalization. Table I summarizes the analysis cohort.

TABLE I
COHORT CHARACTERISTICS AND SENSING CONFIGURATION FOR THE POOLED EMFIT AND WSA IN-HOSPITAL ANALYSIS.

Characteristic	Value
Cohort and outcome	
Patients	65
Patient-nights	423
Agitation-positive nights	306 (72.3%)
Demographic and cognitive characteristics	
Male	41 (63.1%)
Female	24 (36.9%)
Age, years	80.0 ± 8.3 (57–96)
MMSE score	11.6 ± 6.7 (0–27; $n = 48$)
Dementia diagnosis	
AD	34 (52.3%)
NOS	11 (16.9%)
DLB	6 (9.2%)
VaD	6 (9.2%)
Mixed	4 (6.2%)
PDD	2 (3.1%)
Other	2 (3.1%)
Monitoring setting and sensing configuration	
Setting	Specialized hospital dementia unit
Sensor placement	Under-mattress sensing
EMFIT patients / nights	40 / 214
WSA patients / nights	25 / 209
Temporal resolution	1 min
Channels	HR, RR, ACT, sleep stage, quality

Note: Values are mean $\pm$ standard deviation (range) or n (%). MMSE was available for 48 patients (17 missing). AD = Alzheimer's disease; DLB = dementia with Lewy bodies; MMSE = Mini-Mental State Examination; NOS = not otherwise specified; PDD = Parkinson's disease dementia; VaD = vascular dementia; Mixed = Alzheimer's + vascular dementia; HR = heart rate; RR=respiratory rate; ACT = bed activity; WSA = Withings Sleep Analyzer.

Agitation was scored by nursing and research staff using the Pittsburgh Agitation Scale (PAS) [30] within the previously described signal- and event-contingent ecological momentary

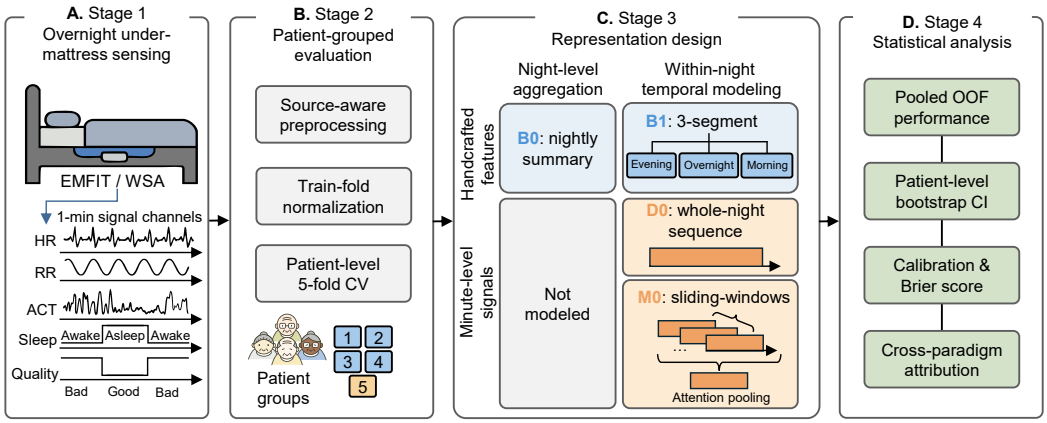


Fig. 1. Four-paradigm modeling pipeline for retrospective next-day agitation risk scoring. The design contrasts handcrafted features with minute-level signals and whole-night aggregation with within-night representations. B0 uses nightly handcrafted summaries. B1 uses handcrafted features from the evening period (17:00–22:00), core overnight period (22:00–06:00), and morning period (06:00–10:00). D0 encodes the full ordered minute-level sequence. M0 encodes overlapping minute-level windows followed by attention pooling. All paradigms use the same patient-grouped folds and are compared using pooled OOF predictions.

assessment protocol [29], [31]. Each overnight recording on night j was paired with agitation observations from the following daytime study period on day $j+1$. Full participant monitoring, annotation, and data-collection procedures are reported by Davidoff et al. [29]. PAS rates motor agitation, verbal agitation, aggression, and resistance to care as four domains scored from 0 to 4. The day-level PAS total score was calculated as the sum of the maximum daytime score in each domain. The primary endpoint was intentionally sensitive: a patient-day was classified as positive if any PAS domain score exceeded zero. It therefore includes minor or transient behaviors that may not prompt a change in care and should be interpreted as any recorded agitation rather than clinically significant agitation. The resulting prevalence was 72.3% (306/423 nights), reflecting a specialized hospital dementia-unit cohort with substantial behavioral and psychological symptoms of dementia (BPSD) and intensive behavioral monitoring rather than a general community-dwelling dementia population [1].

B. Four-Paradigm Modeling Pipeline

The four paradigms compare how alternative representations of one overnight sensor record are mapped to a next-day agitation probability (Fig. 1). B0 and B1 use handcrafted features at different levels of temporal aggregation, whereas D0 and M0 learn representations from minute-level signals.

To formalize this comparison, let $\mathcal{O}_i$ be patient-night i, $\mathbf{X}_i \in \mathbb{R}^{T_i \times 5}$ its cleaned sequence of HR, RR, ACT, binary sleep–wake, and quality, and $y_i \in \{0,1\}$ its following-day agitation outcome. Paradigm m estimates

$$\hat{p}_i^{(m)} = f_m(\mathcal{R}_m(\mathcal{O}_i)), \quad m \in \{\mathrm{B0}, \mathrm{B1}, \mathrm{D0}, \mathrm{M0}\}, \tag{1}$$

where $\mathcal{R}_m$ is the paradigm-specific representation and f_m is its selected prediction rule, including fold-local preprocessing and feature selection where applicable. The four representations can be summarized as

$$\begin{aligned}
\mathcal{R}_{\mathrm{B0}}(\mathcal{O}_i) &= \Phi(\mathcal{O}_i), \\
\mathcal{R}_{\mathrm{B1}}(\mathcal{O}_i) &= [\Phi(\mathcal{O}_i^{(E)}), \Phi(\mathcal{O}_i^{(C)}), \Phi(\mathcal{O}_i^{(M)}), \mathbf{\Delta}_i], \\
\mathcal{R}_{\mathrm{D0}}(\mathcal{O}_i) &= \mathbf{X}_i, \\
\mathcal{R}_{\mathrm{M0}}(\mathcal{O}_i) &= \sum_{j=1}^{J_i} a_{ij}\mathbf{h}_{ij}, \quad \mathbf{h}_{ij} = e_\theta(\mathbf{W}_{ij}).
\end{aligned} \tag{2}$$

Here, Φ denotes handcrafted feature extraction; E, C, and M denote the evening, core overnight, and morning periods; and $\mathbf{\Delta}_i$ contains cross-period descriptors. For M0, $\mathbf{W}_{ij}$ is the jth valid 90-minute window obtained with a 45-minute stride, e_θ is the window encoder, and J_i is the number of valid windows. The learned gated-attention score s_ψ produces normalized weights

$$a_{ij} = \frac{\exp\{s_\psi(\mathbf{h}_{ij})\}}{\sum_{k=1}^{J_i} \exp\{s_\psi(\mathbf{h}_{ik})\}}. \tag{3}$$

The weights satisfy $\sum_{j=1}^{J_i} a_{ij} = 1$ and quantify relative model-associated contribution to the pooled night representation. They should not be interpreted as causal or physiological effect estimates. A separate minute-level-input/night-level-summary pipeline was not evaluated because aggregating the minute-level streams over the full night would substantially overlap with B0.

B0 – Nightly summary (*handcrafted features* × *whole-night aggregation*). The operator Φ summarizes HR, RR, sleep architecture, activity, and signal quality over the full night [28]. Within each training fold, SHAP-based feature selection (SHAP-FS) ranks features by their mean absolute SHAP value for the positive class, reflecting their average attribution magnitude within the selector model [32]. The selected representative is a Random Forest using fold-wise SHAP-FS.

B1 – Period-resolved summary (*handcrafted features* × *period-resolved aggregation*). B1 applies Φ separately to the evening period (17:00–22:00), core overnight period (22:00–06:00), and morning period (06:00–10:00), then appends $\mathbf{\Delta}_i$

to capture cross-period changes. Using the same fold-wise SHAP-FS procedure as B0, the selected representative is Logistic Regression. To standardize tabular input dimensionality and limit feature-space complexity, B0 and B1 use the top 30 fold-specific SHAP-ranked features. Supplementary Section S2 details the period definitions and cross-period descriptors.

D0 – Whole-night raw-sequence modeling (*minute-level signals* × *full-sequence modeling*). D0 directly models $\mathbf{X}_i$ as the complete ordered minute-level sequence. The selected representative is InceptionTime [33].

M0 – Sliding-window MIL (*minute-level signals* × *window-based MIL*). M0 partitions $\mathbf{X}_i$ into overlapping 90-minute windows with a 45-minute stride. The selected representative uses a Transformer window encoder with gated-attention MIL pooling [25], [26]. Supplementary Section S6 reports the window/stride sensitivity analysis.

Candidate model families were screened according to held-out validation performance using the predefined patient-grouped folds, and one representative model was selected for each pipeline. For B0 and B1, selection was performed within the prespecified SHAP-FS configuration. M0 window and stride were selected using the same mean-fold criterion. Further implementation and candidate-model screening details are provided in Supplementary Sections S3–S6. Final performance is summarized from pooled OOF predictions.

C. Preprocessing and Cross-Validation

Each overnight record comprised four 1-minute data streams: heart rate, respiratory rate, bed activity, and a binary sleep–wake channel derived from the device-provided Deep, Light, REM, and Wake/Awake stages. Deep, Light, and rapid eye movement (REM) were encoded as sleep (1), whereas Wake/Awake was encoded as wake (0); missing or unrecognized stages were also encoded as 0, consistent with the implemented preprocessing. A fifth quality channel was derived from device-specific signal-availability and missingness rules. B0 and B1 used statistical and sleep-related features extracted over either the full night or three predefined overnight periods, whereas D0 and M0 received the cleaned 1-minute channels directly. EMFIT and WSA differ in activity scale, derived sleep metrics, and signal dropout characteristics [17], [28]. Derived tabular features and continuous sequence channels were therefore imputed and normalized separately by source using parameters estimated only from each training fold. The binary sleep–wake and quality sequence channels remained on their original scales. Detailed cleaning, quality construction, feature extraction, and normalization procedures are provided in Supplementary Section S1.

All four pipelines used the same predefined five-fold patient-grouped assignment, constructed to balance sensor source and the distribution of patient-level agitation rates across folds. All nights from a patient remained within one fold, and each patient-night contributed exactly one held-out OOF prediction. Class-imbalance weights were estimated exclusively from the corresponding training fold, with no imbalance parameter tuned on held-out predictions.

D. Performance Evaluation

The evaluation was designed to distinguish ranking performance, probability error, and fixed-threshold operating behavior, which can diverge for a high-prevalence outcome (72.3% agitation-positive patient-nights in this cohort). Performance was evaluated from pooled OOF predictions using area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), Brier score, sensitivity, specificity, and balanced accuracy at a fixed probability threshold of 0.5. Balanced accuracy is the mean of sensitivity and specificity. Brier score summarizes probability error, and calibration is visualized with reliability diagrams using 10 quantile-based probability bins.

Given this event prevalence, three reference baselines were included to show how the prevalence of agitation-positive nights affects probability-based and fixed-threshold metrics. For a held-out fold with training-fold prevalence π_{train}, the always-positive baseline assigned probability 1, the train-prevalence baseline assigned the constant probability π_{train}, and the stratified-random baseline sampled predictions from $\mathrm{Bernoulli}(\pi_{\mathrm{train}})$.

E. Statistical Analysis

Statistical analysis was designed to quantify uncertainty in model performance and paired between-pipeline differences while preserving the dependence among repeated nights from the same patient. Confidence intervals were estimated from 1000 patient-level bootstrap replicates, and paired model differences were computed within the same bootstrap samples. Paired DeLong tests [34] were additionally reported as unadjusted secondary AUROC comparisons for B1–B0, D0–B0, M0–B0, D0–B1, M0–B1, and M0–D0. Patient-level bootstrap intervals remained the primary uncertainty estimates because they account for clustering by patient.

F. Cross-Model Feature Attribution Analysis

We compared B1 and M0 to assess whether two contrasting within-night representations assigned relatively high model-associated importance to similar overnight periods and signals. The comparison focused on these pipelines because their importance scores could be summarized over the same three periods and five signals. The resulting analysis comprised 15 period–signal combinations defined by the evening, core overnight, and morning periods and by HR, RR, ACT, the binary sleep–wake channel, and quality. For B1, absolute fold-specific SHAP values for segment-derived features corresponding to a single period and signal were grouped and averaged across OOF nights. For M0, prediction sensitivity was measured by replacing one signal within one period with its preprocessing reference value and calculating the absolute change in predicted probability. Since SHAP magnitude and perturbation-based probability change are measured on different scales, the 15 combinations were converted to within-model relative ranks and displayed in a rank–rank scatter plot. Spearman correlation summarized descriptive agreement between the rankings. This descriptive rank comparison is

TABLE II
MAIN POOLED OOF RESULTS FOR THE EMFIT AND WSA IN-HOSPITAL COHORT (423 NIGHTS, PREVALENCE 0.723). METRICS ARE REPORTED AS POOLED OOF POINT ESTIMATES WITH 95% CIS FROM PATIENT-LEVEL BOOTSTRAP RESAMPLING. BOLDFACE INDICATES THE BEST PERFORMANCE AMONG B0/B1/D0/M0.

Model	AUROC	AUPRC	Sensitivity	Specificity	Brier	BalAcc
Always positive	0.500 [0.500, 0.500]	0.723 [0.648, 0.796]	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.277 [0.204, 0.352]	0.500 [0.500, 0.500]
Train prevalence	0.441 [0.333, 0.554]	0.693 [0.594, 0.795]	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.201 [0.169, 0.235]	0.500 [0.500, 0.500]
Stratified random	0.516 [0.463, 0.572]	0.730 [0.652, 0.806]	0.732 [0.681, 0.787]	0.299 [0.211, 0.400]	0.388 [0.334, 0.447]	0.516 [0.463, 0.572]
B0-RF	0.565 [0.472, 0.657]	0.785 [0.690, 0.866]	**0.948 [0.907, 0.980]**	0.077 [0.031, 0.126]	**0.206 [0.174, 0.240]**	0.512 [0.482, 0.543]
B1-LR	0.622 [0.547, 0.703]	0.816 [0.737, 0.881]	0.644 [0.569, 0.727]	0.538 [0.414, 0.658]	0.246 [0.212, 0.278]	0.591 [0.525, 0.662]
D0-InceptionTime	**0.692 [0.622, 0.759]**	**0.849 [0.778, 0.907]**	0.641 [0.541, 0.745]	**0.675 [0.562, 0.777]**	0.218 [0.193, 0.241]	**0.658 [0.598, 0.718]**
M0-Transformer	0.679 [0.592, 0.764]	0.812 [0.733, 0.889]	0.680 [0.582, 0.775]	0.632 [0.490, 0.770]	0.223 [0.197, 0.249]	0.656 [0.585, 0.720]

hypothesis-generating and does not establish physiological mechanisms or causal relationships.

G. Subgroup and Sensitivity Analyses

Since the pooled cohort incorporated recordings from two sleep sensors and the primary endpoint used a broad definition of agitation, these additional analyses examined performance variation across sensor sources and under a higher-burden agitation definition.

Source-stratified performance was assessed by calculating metrics separately for EMFIT and WSA from the pooled patient-grouped OOF predictions. Both sensor subsets were represented within the common patient-grouped cross-validation framework.

The primary endpoint classified a night as positive when any next-day PAS domain score exceeded zero. Endpoint sensitivity was examined using a day-level PAS total score greater than 4. Because an individual PAS domain has a maximum score of 4, a day-level PAS total score greater than 4 requires agitation in at least two domains and at least one domain score above 1. This endpoint therefore represents more extensive or more intense agitation than the primary endpoint. The sensitivity analysis tests performance under a lower-prevalence, higher-burden outcome.

III. RESULTS

A. Overall Predictive Performance

The pooled OOF analysis separated the four prespecified pipelines into two broad performance patterns (Table II). D0 and M0 had higher discrimination point estimates and more balanced fixed-threshold behavior than B0 and B1. D0 achieved the highest AUROC (0.692 [0.622, 0.759]), AUPRC, specificity, and balanced accuracy, while M0 achieved similar balanced accuracy. B1 had higher AUROC and specificity point estimates than B0.

B. Pairwise Statistical Comparisons

Paired patient-level bootstrap comparisons localized the clearest discrimination gains to the sequence- and window-based models relative to B0: the AUROC differences were +0.127 for D0 and +0.114 for M0, with both confidence intervals excluding zero (Table III). The remaining paired intervals crossed zero, including comparisons of D0 and M0 with B1 and with each other. DeLong tests are reported as secondary comparisons.

TABLE III
PAIRED AUROC COMPARISONS ON SHARED POOLED OOF PREDICTIONS. PATIENT-LEVEL BOOTSTRAP INTERVALS ARE THE PRIMARY UNCERTAINTY ESTIMATES.

Comparison	ΔAUROC	Bootstrap 95% CI	DeLong p
B1–B0	+0.057	[-0.019, +0.133]	0.077
D0–B0	+0.127	[+0.037, +0.219]	0.0005
M0–B0	+0.114	[+0.021, +0.210]	0.0015
D0–B1	+0.070	[-0.022, +0.163]	0.057
M0–B1	+0.057	[-0.037, +0.144]	0.113
M0–D0	-0.013	[-0.103, +0.077]	0.714

C. Calibration and Prevalence-based Baselines

The simple reference baselines contextualize the impact of the high prevalence of agitation-positive nights (72.3%). The always-positive and train-prevalence baselines had no useful discrimination or specificity, although the train-prevalence baseline achieved the lowest Brier score overall because its constant probability closely matched the pooled event rate (Table II). B0 produced a similar Brier score but classified most nights as positive, yielding high sensitivity and very low specificity. Its relatively low probability error therefore did not yield balanced classification at the prespecified 0.5 threshold. The reliability diagram indicates that calibration remained limited across the learned models (Fig. 2).

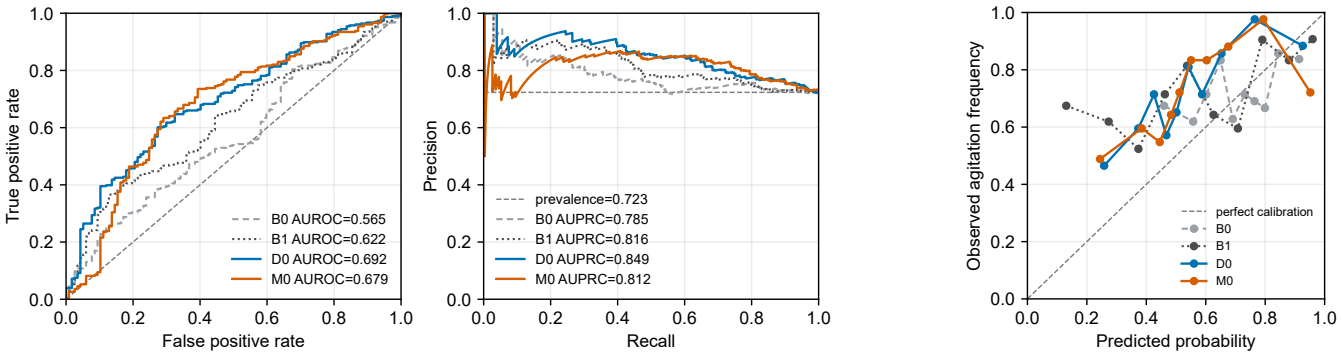
D. Cross-Model Feature Attribution

Feature attribution from the B1 Logistic Regression and M0 MIL Transformer was summarized across 15 overnight-period–signal combinations and converted to within-model relative ranks for qualitative comparison.

The final-model attribution summaries showed moderate descriptive rank agreement (Spearman $\rho = 0.57$, Fig. 3). ACT, HR, and RR during the core overnight period occupied the upper-right region, indicating relatively high attribution ranks in both models. Morning-period ACT and the binary sleep–wake channel during the core overnight period lay farther from the diagonal. Morning-period ACT ranked higher in B1, whereas the core-overnight sleep–wake channel ranked higher in M0.

E. Source-Stratified Performance

Source-stratified performance varied between EMFIT and WSA (Table IV). B0 showed the largest source-stratified



(a) ROC and precision-recall overlay. (b) Reliability diagram.

Fig. 2. Discrimination and calibration on shared pooled OOF predictions. D0 has the highest AUROC, AUPRC, and balanced accuracy, whereas B0 has the lowest learned-model Brier score despite weak specificity. Calibration remains limited relative to the train-prevalence probability baseline in Table II.

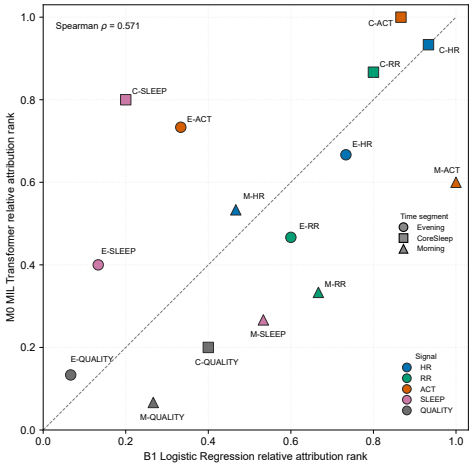


Fig. 3. Qualitative comparison of B1 segment-feature and M0 final-model attribution rankings across 15 segment–signal combinations. Each axis shows the relative attribution rank within one model, with higher values indicating higher rank. Combinations near the dashed diagonal have similar ranks, and those in the upper-right region rank highly in both models. Marker shape denotes the within-night segment and color denotes the signal. The descriptive Spearman rank correlation was $\rho = 0.57$.

TABLE IV
SENSOR-SOURCE-STRATIFIED PERFORMANCE FROM POOLED OOF PREDICTIONS. EMFIT INCLUDED 214 NIGHTS WITH A PREVALENCE OF 0.673, AND WSA INCLUDED 209 NIGHTS WITH A PREVALENCE OF 0.775.

Source	Model	AUROC	AUPRC	Sens.	Spec.	BalAcc
EMFIT	B0	0.506	0.679	0.938	0.057	0.497
	B1	0.606	0.749	0.611	0.557	0.584
	D0	0.688	0.802	0.625	0.743	0.684
	M0	0.648	0.758	0.597	0.643	0.620
WSA	B0	0.631	0.866	0.957	0.106	0.532
	B1	0.630	0.868	0.673	0.511	0.592
	D0	0.680	0.881	0.654	0.574	0.614
	M0	0.695	0.851	0.753	0.617	0.685

AUROC difference, while M0 also differed across the two subsets. B1 and D0 were comparatively stable in AUROC. Despite stable discrimination, D0's balanced accuracy was lower for WSA, primarily because of lower specificity.

F. Sensitivity to Agitation Endpoint Definition

Defining higher-burden agitation as a day-level PAS total score greater than 4 reduced outcome prevalence from 72.3% to 17.3% (Supplementary Table S3). D0 retained the highest AUROC (0.654), AUPRC (0.286), sensitivity (0.575), and balanced accuracy (0.638), whereas B1 achieved the lowest Brier score (0.164) and highest specificity (0.911). The analysis therefore showed partial stability in model ranking but substantial changes in operating characteristics under the lower-prevalence endpoint.

IV. Discussion

A. Temporal Representation and Model Behavior

The principal finding is that contactless under-mattress signals from the preceding night contained modest information for estimating next-day agitation risk among PLWD in a hospital setting. Representations retaining minute-level within-night structure achieved better discrimination than conventional whole-night summaries, and performance varied according to how that temporal structure was represented. Their AUROC gains over B0 were supported by patient-level bootstrap intervals, whereas differences relative to the period-specific B1 pipeline and between D0 and M0 remained uncertain. The overall predictive performance was modest, and the comparison does not isolate representation effects from classifier–representation interactions.

D0 and M0 retained temporal variation at different scales: D0 modeled the complete ordered minute-level sequence and achieved the highest AUROC, AUPRC, specificity, and balanced accuracy among the four evaluated pipelines, whereas M0 encoded local windows before learned aggregation and remained close to D0 in balanced accuracy. In contrast, B0 applies the greatest temporal compression, while B1 retains only a coarse predefined partition of the night. The four-pipeline benchmark therefore characterizes the consequences of different levels of temporal compression under a common patient-grouped design and identifies full-sequence and window-based approaches as stronger candidates than nightly summaries in this cohort. The data do not, however, establish that either learned temporal strategy is definitively superior.

This representation-centered comparison extends prior work in which LightGBM was selected for engineered daily or multi-day feature sets and complements the earlier mixed-effects analysis of feature-level associations [18], [23], [28]. Because the analytical tasks, input representations, and evaluation designs differed, performance values from the predictive studies should not be compared directly with those from the present patient-grouped benchmark.

The prevalence-based baselines demonstrate how the high event rate (72.3%) influences predictive metrics. The always-positive and train-prevalence baselines had zero specificity, while the train-prevalence baseline achieved the lowest Brier score despite having no useful discrimination. B0 showed a related pattern, combining high sensitivity with very low specificity and a Brier score close to the prevalence baseline. These findings demonstrate that sensitivity or probability error alone may overstate performance in a high-prevalence cohort and underscore the more balanced discrimination of D0 and M0. Calibration nevertheless remained limited across the learned models.

B. Physiological Context of Nocturnal Attribution Patterns

The attribution ranking comparison identified activity, heart rate, and respiratory rate during the core overnight period as common high-ranking period–signal combinations across B1 and M0. Their convergence across two different representations prioritizes these signals for targeted physiological validation beyond static nightly summaries. Movement-derived measures during this period may reflect sleep fragmentation or nocturnal restlessness [35], [36], both of which have been linked to daytime behavioral symptoms in PLWD [9].

The high attribution ranks of HR and respiratory rate during the core overnight period suggest that nocturnal cardiorespiratory variation may contain information relevant to next-day agitation. This interpretation is consistent with related contactless-sensing evidence associating nocturnal respiratory measures with next-day agitation occurrence [28], and with broader in-home monitoring work showing that sleep and physiological features can contribute to agitation prediction [18]. Previous work linking heart rate variability with agitation risk in Alzheimer's disease provides complementary evidence that autonomic physiology may be relevant, although at a substantially longer timescale [13]. These descriptive attribution patterns generate hypotheses about sleep disruption, autonomic physiology, and other potential mechanisms that require further validation.

The largest rank differences provide complementary insight into how temporal representation shapes model behavior. Activity during the morning period ranked higher in B1, suggesting that segment-level summary features may emphasize concentrated activity around the transition out of the overnight period. In contrast, the binary sleep–wake channel during the core overnight period ranked higher in M0, suggesting greater sensitivity to local sleep–wake composition than aggregated segment features capture. These representation-specific differences complement the shared emphasis on activity and cardiorespiratory signals during the core overnight period.

C. Robustness and Generalizability

Patient-grouped evaluation tests whether signals from the preceding night can support risk scoring for patients excluded from the corresponding training fold. Accordingly, the evidence pertains to cross-patient risk estimation and does not establish patient-specific day-to-day forecasting.

Source-stratified performance point estimates differed across EMFIT and WSA. Since the data were collected in the same ward during different periods, observed variation may reflect differences in device hardware, device-specific signal processing, labeling protocols, and patient cohorts. D0 rank discrimination was comparatively stable across the two acquisition streams, but the analysis does not establish transportability to other devices or clinical settings.

Model applicability also depends on how agitation is defined. The higher-burden endpoint preserved D0's leading rank on several metrics but changed prevalence and operating characteristics substantially. These findings show that model behavior and practical interpretation depend on the selected agitation definition.

D. Clinical Interpretation

The primary endpoint was intentionally sensitive and classified any nonzero PAS domain score as positive. It therefore includes minor or transient behaviors that may not prompt a change in care and should be interpreted as any recorded agitation rather than clinically significant agitation. A positive model output consequently represents an elevated probability of any observed agitated behavior under the study definition, not a diagnosis of clinically actionable agitation. The changes in prevalence and operating characteristics under the higher-burden endpoint show that performance depends on the intended behavioral target as well as the sensing representation.

Agitation in PLWD can reflect multiple and potentially modifiable contributors, including unmet needs, discomfort, and clinical or environmental stressors [37], [38]. A future sensor-based alert should therefore be treated as an adjunct to observation and structured assessment rather than as a standalone indication for treatment. Its most defensible role would be to prompt timely review of the patient and care context when staff judgment also indicates concern.

Limited calibration means that the predicted probabilities cannot yet be interpreted as reliable absolute risks; for example, a predicted risk of 70% does not necessarily correspond to a 70% observed probability of next-day agitation [39]. At the present performance level, a high-risk score should not trigger treatment. Before workflow use, calibration and the operating threshold would need prospective evaluation in the target setting because the consequences of missed events and unnecessary alerts depend on the intended response.

E. Strengths, Limitations, and Future Work

The study combines 423 repeated overnight observations acquired during routine hospital care, two passive under-mattress sensing systems, minute-level physiological and activity signals, and a common patient-grouped OOF evaluation

across four temporal representations. Source-aware preprocessing and patient-level bootstrap resampling address device heterogeneity and within-patient dependence. Together, these elements provide a dense short-term repeated-measures dataset under real-world acquisition conditions.

The study contains a substantial number of repeated overnight observations, but cross-patient generalizability remains limited by the 65 independent participants and recruitment from a single specialized hospital dementia unit. Inter-rater reliability was not assessed, potentially introducing variability into the outcome labels due to differences between nursing and research staff ratings. The broad primary endpoint also captures minor or transient behaviors, limiting inference about clinically significant agitation. In addition, the representative classifiers and M0 windowing configuration were selected through exploratory comparisons on the same patient-grouped folds rather than through a separate model-selection cohort or nested cross-validation. Consequently, reported performance may be optimistic in future samples, and the bootstrap confidence intervals do not include uncertainty introduced by configuration selection.

Future work should use prespecified model selection, external validation across hospital and device settings, prospective and source-stratified calibration, and targeted physiological validation of the period–signal combinations that ranked highly in our cross-model attribution analysis. These steps are required to establish robustness beyond this single-site cohort and to determine whether the risk scores can support a defined care workflow.

V. Conclusion

Under-mattress signals recorded during the preceding night contained modest information about next-day agitation risk in this in-hospital dementia cohort. Full-sequence and sliding-window modeling achieved stronger discrimination than conventional nightly summaries, although their advantages over period-specific engineered features remained uncertain. Cross-model attribution consistently prioritized activity and cardiorespiratory signals during the core overnight period for further physiological validation. These findings support minute-level temporal modeling as a promising engineering direction, but external validation and prospective calibration are required before the resulting scores can guide individual care decisions.

Acknowledgments

The authors thank all participating patients and their families, as well as the nursing staff and psychiatrists in training at Cog K of UPC KU Leuven.

Statements and declarations

Ethical considerations

Ethical approval for this study was granted by the Ethics Committee Research of University Hospitals Leuven (reference ID S62882). Given the participants' diagnosis of dementia, written informed consent was primarily secured through legal representatives. However, to ensure patient inclusion, the study details were explained verbally in a manner commensurate with each patient's cognitive abilities. All patient data were pseudonymized and handled with strict confidentiality. Participation was voluntary and uncompensated, with the explicit understanding that the decision to enroll would not impact the standard of care received during admission.

Consent to participate

Informed consent was obtained from all individual participants included in the study. For participants who were unable to provide consent due to cognitive impairment, written informed consent was obtained from their legal representativesor primary caregivers. All procedures were performed in accordance with the study protocol approved by the Research Ethics Committee of University Hospitals Leuven.

Funding statement

This work was supported by the Flanders AI Research (FAIR) Program; the Research Foundation – Flanders (FWO) under grant number G046925N; the 'Funds Malou Malou, Perano, Georgette Paulus, JMJS Breugelmans and Gabrielle, François and Christian De Mesmaeker' managed by the King Baudouin Foundation of Belgium (2021-J1990130-222081); KU Leuven (grants C2M/23/053, CELSA/24/019, and the Global PhD partnerships with the University of Melbourne GPUM/22/022); the Paris Brain Institute and KU Leuven (Big Brain Theory 4 grant); the 'Klinische onderzoeks- en opleidingsraad (KOOR)' of the University Hospitals Leuven; and the AI-PROGNOSIS project under the European Union's Horizon Europe research and innovation program (Grant Agreement No. 101080581). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Data and code availability

The primary dataset analyzed in the current study is not publicly available due to patient privacy and ethical restrictions, but de-identified data are available from the corresponding author upon reasonable request.

The code for experiments presented in this study will be made available by the corresponding author upon reasonable request.

Declaration of generative AI use

During the preparation of this work, the authors used OpenAI's ChatGPT for English editing, code debugging, and optimization during manuscript preparation. All content was reviewed and verified by the authors, who retain full responsibility for the final work.

References

[1] J. Cerejeira, L. Lagarto, and E. B. Mukaetova-Ladinska, "Behavioral and psychological symptoms of dementia," *Frontiers in neurology*, vol. 3, p. 73, 2012.

[2] M. Sano, J. Cummings, S. Auer, S. Bergh, C. E. Fischer, D. Gerritsen, G. Grossberg, Z. Ismail, K. Lanctôt, M. I. Lapid *et al.*, "Agitation in cognitive disorders: progress in the international psychogeriatric association consensus clinical and research definition," *International psychogeriatrics*, vol. 36, no. 4, pp. 238–250, 2024.

[3] G. Livingston, J. Barber, L. Marston, P. Rapaport, D. Livingston, S. Cousins, S. Robertson, F. La Frenais, and C. Cooper, "Prevalence of and associations with agitation in residents with dementia living in care homes: MARQUE cross-sectional study," *BJPsych Open*, vol. 3, no. 4, pp. 171–178, 2017.

[4] C. Carrarini, M. Russo, F. Dono, F. Barbone, M. G. Rispoli, L. Ferri, M. Di Pietro, A. Digiovanni, P. Ajdinaj, R. Speranza *et al.*, "Agitation and dementia: prevention and treatment strategies in acute and chronic conditions," *Frontiers in neurology*, vol. 12, p. 644317, 2021.

[5] J. A. Watt, Z. Ismail, N. Lane, D. Seitz, and Z. Goodarzi, "Management of behavioral and psychological symptoms of dementia: Nonpharmacologic and pharmacologic interventions," *Annual Review of Clinical Psychology*, vol. 22, 2026.

[6] L. S. Schneider, K. S. Dagerman, and P. Insel, "Risk of death with atypical antipsychotic drug treatment for dementia: meta-analysis of randomized placebo-controlled trials," *Jama*, vol. 294, no. 15, pp. 1934–1943, 2005.

[7] L. Shi, S.-J. Chen, M.-Y. Ma, Y.-P. Bao, Y. Han, Y.-M. Wang, J. Shi, M. V. Vitiello, and L. Lu, "Sleep disturbances increase the risk of dementia: a systematic review and meta-analysis," *Sleep medicine reviews*, vol. 40, pp. 4–16, 2018.

[8] L. Webster, S. Costafreda Gonzalez, A. Stringer, A. Lineham, J. Budgett, S. Kyle, J. Barber, and G. Livingston, "Measuring the prevalence of sleep disturbances in people with dementia living in care homes: a systematic review and meta-analysis," *Sleep*, vol. 43, no. 4, p. zsz251, 2020.

[9] D. T. Brown, J. L. Westbury, and B. Schüz, "Sleep and agitation in nursing home residents with and without dementia," *International psychogeriatrics*, vol. 27, no. 12, pp. 1945–1955, 2015.

[10] M. F. Georgescu, "Agitation and neuropsychiatric symptoms in alzheimer's disease: Potential for reverse causal associations," *Journal of Alzheimer's Disease*, vol. 109, no. 2, pp. 624–626, 2026.

[11] R. A. Wijbenga, F. J. Blaauw, S. I. Janus, C. Tibben, A. E. Smits, R. C. Oude Voshaar, S. U. Zuidema, and M. Zuidersma, "Individual differences in the temporal relationship between sleep and agitation: a single-subject study in nursing home residents with dementia experiencing sleep disturbance and agitation," *Aging & Mental Health*, vol. 26, no. 8, pp. 1669–1677, 2022.

[12] E. Cho, S. Hwang, M. Yang, E. Kim, J. Cho, and C. Park, "Dynamic relationship between sleep patterns and behavioral and psychological symptoms of dementia: Longitudinal observational study," *JMIR Aging*, vol. 9, p. e80422, 2026.

[13] K. Y. Liu, E. A. Whitsel, G. Heiss, P. Palta, S. Reeves, F. V. Lin, M. Mather, J. P. Roiser, and R. Howard, "Heart rate variability and risk of agitation in alzheimer's disease: The atherosclerosis risk in communities study," *Brain Communications*, vol. 5, no. 6, p. fcad269, 2023.

[14] M. Kholghi, I. Szollosi, M. Hollamby, D. Bradford, and Q. Zhang, "A validation study of a ballistocardiograph sleep tracker against polysomnography," *Journal of Clinical Sleep Medicine*, vol. 18, no. 4, pp. 1203–1210, 2022.

[15] K. K. Ravindran, C. Della Monica, G. Atzori, D. Lambert, H. Hassanin, V. Revell, and D.-J. Dijk, "Contactless and longitudinal monitoring of nocturnal sleep and daytime naps in older men and women: a digital health technology evaluation study," *Sleep*, vol. 46, no. 10, p. zsad194, 2023.

[16] E. Soreq, M. A. Kolanko, C. group, K. K. Ravindran, C. della Monica, V. Revell, S. Daniels, A. Joffe, H. Lai, M. Golemme *et al.*, "Contactless longitudinal monitoring in the home characterizes aging and alzheimer's disease–related night-time behavior and physiology," *Alzheimer's & Dementia*, vol. 21, no. 10, p. e70758, 2025.

[17] K. K. G Ravindran, C. Della Monica, G. Atzori, D. Lambert, H. Hassanin, V. Revell, and D.-J. Dijk, "Reliable contactless monitoring of heart rate, breathing rate, and breathing disturbance during sleep in aging: digital health technology evaluation study," *JMIR mHealth and uHealth*, vol. 12, p. e53643, 2024.

[18] M. Bafaloukou, A.-K. Schalkamp, N. Fletcher-Lloyd, A. Capstick, C. Walsh, C. Sandor, S. Kouchaki, R. Nilforooshan, and P. Barnaghi, "An interpretable machine learning tool for in-home monitoring of agitation episodes in people living with dementia: a proof-of-concept study," *EClinicalMedicine*, vol. 80, 2025.

[19] D. W. Jung, S. H. Hwang, H. N. Yoon, Y.-J. G. Lee, D.-U. Jeong, and K. S. Park, "Nocturnal awakening and sleep efficiency estimation using unobtrusively measured ballistocardiogram," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 1, pp. 131–138, 2014.

[20] M. Yin, C. Hadjipanayi, K. K. G. Ravindran, A. Bannon, A. Rapeaux, C. della Monica, T. S. Lande, D.-J. Dijk, and T. G. Constandinou, "Unobtrusive sleep health assessment using impulse radar: A pilot study in older people," *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 9, pp. 2663–2674, 2025.

[21] B. S. Husebo, H. L. Heintz, L. I. Berge, P. Owoyemi, A. T. Rahman, and I. V. Vahia, "Sensing technology to monitor behavioral and psychological symptoms and to assess treatment response in people with dementia. a systematic review," *Frontiers in pharmacology*, vol. 10, p. 1699, 2020.

[22] J. K. Deters, S. Janus, J. A. L. Silva, H. J. Wörtche, and S. U. Zuidema, "Sensor-based agitation prediction in institutionalized people with dementia a systematic review," *Pervasive and Mobile Computing*, vol. 98, p. 101876, 2024.

[23] K. Ramesh, A. Yakoub, Y. Ghoneim, R. Al Korabi, J. Ramesh, A. Sagahyroon, and F. Aloul, "Machine learning prediction of agitation in dementia patients using sleep and physiological data," *Applied Sciences*, vol. 15, no. 18, p. 9908, 2025.

[24] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2019.

[25] H. Phan, K. Mikkelsen, O. Y. Chen, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2456–2467, 2022.

[26] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the International Conference on Machine Learning*, vol. 80. PMLR, 2018, pp. 2127–2136. [Online]. Available: https://proceedings.mlr.press/v80/ilse18a.html

[27] H. Han, C. Lian, Z. Zeng, B. Xu, J. Zang, and C. Xue, "MAMIL: Multimodal multi-instance learning for long-term ECG classification," *Knowledge-Based Systems*, vol. 270, p. 110555, 2023.

[28] Z. Liu, M. Bono, A. Flisar, R. Decloedt, M. Van Den Bossche, and M. De Vos, "Dissociating the nocturnal physiological drivers of agitation occurrence and severity in dementia: An explanatory study using contactless sleep sensing," *medRxiv*, pp. 2026–02, 2026.

[29] H. Davidoff, L. Van den Bulcke, M. Vandenbulcke, M. De Vos, J. Van den Stock, N. Van Helleputte, C. Van Hoof, and M. J. Van Den Bossche, "Toward quantification of agitation in people with dementia using multimodal sensing," *Innovation in Aging*, vol. 6, no. 7, p. igac064, 2022.

[30] J. Rosen, L. Burgio, M. Kollar, M. Cain, M. Allison, M. Fogleman, M. Michael, and G. S. Zubenko, "The pittsburgh agitation scale: A user-friendly instrument for rating agitation in dementia patients," *The American Journal of Geriatric Psychiatry*, vol. 2, no. 1, pp. 52–59, 1994.

[31] H. Davidoff, A. Van Kraaij, L. Van den Bulcke, E. Lutin, M. Vandenbulcke, N. Van Helleputte, M. De Vos, C. Van Hoof, and M. Van Den Bossche, "Physiological profiling of agitation in dementia: insights from wearable sensor data," *Innovation in Aging*, vol. 8, no. 7, p. igae057, 2024.

[32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[34] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[35] A. S. Lim, M. Kowgier, L. Yu, A. S. Buchman, and D. A. Bennett, "Sleep fragmentation and the risk of incident alzheimer's disease and cognitive decline in older persons," *Sleep*, vol. 36, no. 7, pp. 1027–1032, 2013.

[36] D. Saleh, S. M. Bertisch, M. Reid, A. Lim, S. Purcell, and S. Redline, "Actigraphy-derived sleep fragmentation index: convergent validity and associations with clinical outcomes," *Journal of Clinical Sleep Medicine*, vol. 21, no. 9, pp. 1557–1565, 2025.

[37] H. C. Kales, L. N. Gitlin, and C. G. Lyketsos, "Assessment and management of behavioral and psychological symptoms of dementia," *Bmj*, vol. 350, 2015.

[38] J. Cohen-Mansfield, M. Dakheel-Ali, M. S. Marx, K. Thein, and N. G. Regier, "Which unmet needs contribute to behavior problems in persons

with advanced dementia?" *Psychiatry research*, vol. 228, no. 1, pp. 59–64, 2015.

[39] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, p. 230, 2019.

[40] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[41] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[42] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 1578–1585.

[43] M. Middlehurst, P. Schäfer, and A. Bagnall, "Bake off redux: a review and experimental evaluation of recent time series classification algorithms," *Data Mining and Knowledge Discovery*, vol. 38, no. 4, pp. 1958–2031, 2024.

[44] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[45] I. Oguiza, "tsai: A state-of-the-art deep learning library for time series and sequential data," GitHub, 2023. [Online]. Available: https://github.com/timeseriesAI/tsai

Supplementary Material

Under-Mattress Temporal Sensing for Next-Day Agitation Risk Scoring in Dementia Wards

This document provides preprocessing and model-configuration details, together with secondary analyses that qualify the four-paradigm comparison in the main manuscript. It documents the four input representations and representative-model selection, followed by analyses of window sensitivity, the higher-burden agitation endpoint, patient-level heterogeneity, and illustrative single-subject trajectories.

S1. Temporal Sequence Preprocessing Details

The temporal models are trained from cleaned 1-minute sequence files produced by device-specific raw-data processing. All device streams are restricted to an overnight window from 17:00 to the following 10:00 before modeling, and each resulting overnight record is paired with the subsequent daytime agitation label. Physiologically implausible HR and RR values are set unavailable, missingness indicators are included in the cleaned files, and severely unreliable minutes are marked by a 1-minute quality flag.

For EMFIT, raw vitals sampled at approximately 4-second resolution are range-checked using HR 35–180 bpm and RR 6–40 rpm, snapped to a complete 4-second grid, and aggregated to 1-minute resolution. HR and RR use median aggregation, while activity uses mean activity. The EMFIT quality flag is based on the 1-minute proportion of simultaneous HR/RR/ACT absence, with severe missingness marked when the all-signal missing proportion is at least 0.4 or the completed grid minute is empty. The EMFIT sleep-stage field is aligned from the sleep-class file to the vital stream and mapped to Deep, Light, REM, or Awake where vital data are present.

For WSA, the raw series is already at 1-minute resolution. HR and RR are range-checked with the same physiological bounds as EMFIT and then processed with a NaN-safe median filter with window size 3. The movement-score field is used as the activity channel and is not median-filtered. The WSA quality flag marks minutes in which HR, RR, and movement score are all missing, and the cleaned WSA state field provides the sleep-stage channel used by downstream feature extraction.

EMFIT and WSA are therefore treated as related but non-exchangeable measurement sources. Although HR and RR use the same physical units, the sensor streams differ in temporal resolution and device-specific processing. Activity is derived from device-specific activity or movement-score fields, sleep-stage fields originate from device-specific outputs, and the quality channel is constructed during preprocessing. Source-aware imputation and scaling are used to avoid forcing the resulting source-specific distributions onto a single pooled scale before model fitting.

The selected temporal channels are HR, RR, ACT, the binary sleep–wake channel, and quality. Device-provided Deep, Light, and REM stages are encoded as sleep (1), whereas Wake/Awake is encoded as wake (0). Missing or unrecognized stages are also encoded as 0, consistent with the implemented preprocessing. The quality channel is derived from the cleaned 1-minute severe-missingness flag. It is encoded as 0 for valid data and 1 for invalid or unknown data, and missing quality flags are treated conservatively as invalid. D0 loads each night as an ordered 1-minute sequence, center-cropped when longer than 800 minutes and right-padded when shorter. M0 splits each night into overlapping 90-minute windows with 45-minute stride before MIL pooling. For both D0 and M0, continuous channels are normalized within each training fold using source-specific robust median/MAD statistics from the training partition only, and held-out fold sequences are transformed with the corresponding training-fold statistics. Activity is winsorized at the within-night 1st and 99th percentiles, missing activity values are filled with zero before transformation, and $\log(1 + x)$ is applied. The sleep–wake and quality channels remain on their binary scales. D0 uses the validity mask during fold-specific normalization and represents padded timesteps as zeros during modeling, whereas M0 includes only windows meeting the minimum valid-data ratio (70%) and uses the minute-level quality flag as an input channel.

S2. Three-Segment Handcrafted Feature Construction

B1 uses a fixed three-segment representation designed to preserve coarse within-night timing while remaining compatible with tabular classifiers. Each overnight record is partitioned into the evening period (17:00–22:00), core overnight period (22:00–06:00), and morning period (06:00–10:00). The corresponding feature-name prefixes are Evening, CoreSleep, and Morning. Within each period, handcrafted statistics are computed for HR, RR, ACT, the binary sleep–wake channel, and quality-derived channels. These include central tendency, dispersion, extrema, interquartile range, coefficient of variation, and distribution-shape summaries where defined, together with sleep-related and presence/unknown-ratio descriptors.

The B1 table also includes cross-period descriptors that encode directional or stability changes across the night. These features include HR dipping from the evening to the core overnight period, HR morning surge from the core overnight to the morning period, RR dipping from the evening to the core overnight period, ACT escalation from the evening to the morning period, and HR/RR/ACT variability ratios across periods. These descriptors are not treated as separate outcomes or mechanistic claims. They are tabular summaries of within-night temporal contrast. All B1 imputation, scaling, SHAP-based feature selection, and classifier fitting are performed within the training fold before held-out OOF prediction.

S3. B0/B1 Tabular Model Configurations

B0 and B1 use the same patient-grouped folds, source-aware preprocessing, candidate classifier family, top-30 SHAP-FS procedure, and mean held-out-fold AUROC selection criterion. B0 uses whole-night summary features, while B1 uses the three-segment feature table plus cross-segment descriptors. This shared screening protocol supports comparison of the complete B0 and B1 pipelines, but the different selected classifiers prevent attribution of performance differences to temporal representation alone. For both paradigms, imputation, scaling, feature selection, and classifier fitting are restricted to the training fold before held-out OOF probabilities are generated.

Tuned models. The tuned tabular analysis evaluates the same classifier set for B0 and B1: Logistic Regression, Random Forest, AdaBoost, SVM with RBF kernel, linear SVM, multilayer perceptron (MLP), XGBoost [40], and LightGBM [41]. Source-specific robust scaling is applied within each training fold. Logistic Regression uses class weighting and a maximum of 1000 iterations. Random Forest uses 100 trees, maximum depth 10, minimum split size 5, and balanced class weights. AdaBoost uses 100 depth-1 decision stumps and a learning rate of 0.1. The SVM models use probability output and balanced class weights. The MLP uses hidden layers $(64, 32, 16)$, ReLU activation, the Adam optimizer, $\alpha = 0.001$, and a maximum of 500 iterations. XGBoost and LightGBM use 100 trees, maximum depth 6, learning rate 0.1, and fold-specific class-imbalance weighting. All imbalance parameters are estimated exclusively from training-fold labels. Mean held-out-fold AUROC is used to select the representative classifier within each prespecified tabular configuration. The primary input-dimension-standardized comparison selects B0 and B1 within the SHAP-FS configuration.

SHAP-based feature selection. Feature selection is performed independently within each training fold. A balanced Random Forest selector with 100 trees, maximum depth 10, and minimum split size 5 is trained on the fold-specific scaled training data, and features are ranked by mean absolute SHAP value for the positive class [32]. To standardize tabular input dimensionality and limit feature-space complexity, both the B0 summary-feature representation and the B1 three-segment representation use the top 30 fold-specific SHAP-ranked features. The candidate classifier set and principal hyperparameters then match the tuned-model analysis. For gradient-boosted classifiers, the positive class is weighted by the training-fold ratio of negative to positive observations. This fold-local procedure prevents held-out-fold information from influencing feature selection. The primary B0 model is SHAP-selected Random Forest on nightly summary features, and the primary B1 model is SHAP-selected Logistic Regression on three-segment features.

S4. D0/M0 Temporal Model Configurations

Both D0 and M0 use the same five model-input channels: HR, RR, ACT, the binary sleep–wake channel, and the derived quality channel. They are evaluated using the same patient-grouped folds as B0 and B1. For each selected temporal representative, each held-out fold is evaluated once using the final model obtained after the fixed number of training epochs.

D0 configuration. D0 models each overnight record as one five-channel 1-minute sequence. Sequences are center-cropped or right-padded to 800 minutes, and fold-specific sequence normalization is fit on the training partition only. The candidate family includes ResNet1D [42], InceptionTime [33], [43], TCN [44], Transformer, and LSTM, implemented with the tsai time-series library where applicable [45]. InceptionTime achieved the highest mean held-out-fold AUROC among the evaluated whole-night raw-sequence candidates and was selected as the D0 representative. The reported InceptionTime configuration uses a five-channel input, two output classes, batch size 8, Adam optimizer with a constant learning rate of 10^{-3}, and cross-entropy loss with fold-specific class weights derived exclusively from training-fold labels. After 20 training epochs, the final-epoch model produces the held-out OOF probabilities.

M0 configuration. M0 partitions each night into overlapping 90-minute windows with 45-minute stride and aggregates valid window embeddings with gated attention pooling. Supplementary Section S6 reports the window/stride sensitivity analysis that informed this windowing choice. The backbone family includes Transformer [25], LSTM, ResNet, TCN [44], and InceptionTime [33] under the five-channel input set with train-fold normalization. The primary M0 model uses a Transformer window encoder with input projection from five channels to hidden dimension 32, two Transformer encoder layers, four attention heads, GELU activation, feed-forward dimension 128, dropout 0.3, and output embedding dimension 64. The night-level classifier maps the pooled 64-dimensional embedding through a 32-unit hidden layer with ReLU and dropout to one logit. The reported Transformer MIL configuration uses batch size 8 bags, AdamW optimizer with a constant learning rate of 10^{-3} and weight decay 10^{-3}, and BCE-with-logits loss with fold-specific positive-class weighting computed exclusively from training-fold labels. After 20 training epochs, the final-epoch model produces the held-out OOF probabilities and attention weights.

S5. Model-Family and Configuration Screening

Exploratory candidate-model comparisons on the same patient-grouped folds informed the selection of one representative model for each paradigm before the shared pooled OOF analysis. Each comparison metric is computed separately on every held-out fold and summarized by its arithmetic mean and sample SD across the five folds. For B0 and B1, representative classifiers were selected within the prespecified SHAP-FS configuration. Tuned and default configurations without SHAP-FS are shown as sensitivity references. This differs from the main-manuscript results table, where held-out predictions are pooled

before calculating each point estimate and patient-level bootstrap resampling provides the uncertainty interval. The configuration comparisons are exploratory because candidate selection and performance estimation used the same patient-grouped folds.

TABLE S1

EXPLORATORY CONFIGURATION COMPARISONS ON THE SAME PATIENT-GROUPED FOLDS. VALUES ARE MEAN ± SAMPLE SD OF THE FIVE FOLD-SPECIFIC HELD-OUT METRICS AND ARE NOT POOLED OOF ESTIMATES. ONLY THE SELECTED SHAP-FS CLASSIFIER IS SHOWN FOR B0 AND B1. TUNED AND DEFAULT ROWS ARE NON-SHAP-FS SENSITIVITY REFERENCES. D0 AND M0 ROWS COMPARE NEURAL ENCODER FAMILIES UNDER THE SELECTED FIVE-CHANNEL INPUT SET.

Paradigm	Configuration / model	AUROC	AUPRC	F1	Accuracy	Sensitivity	Specificity
B0	SHAP-FS Random Forest	0.578 ± 0.103	0.798 ± 0.085	0.823 ± 0.061	0.709 ± 0.083	0.947 ± 0.058	0.080 ± 0.037
B0	Tuned Random Forest	0.589 ± 0.080	0.811 ± 0.062	0.827 ± 0.049	0.714 ± 0.072	0.953 ± 0.027	0.087 ± 0.078
B0	Default Random Forest	0.575 ± 0.083	0.806 ± 0.061	0.829 ± 0.048	0.714 ± 0.068	0.969 ± 0.026	0.040 ± 0.068
B1	SHAP-FS Logistic Regression	0.612 ± 0.056	0.819 ± 0.057	0.701 ± 0.091	0.617 ± 0.080	0.638 ± 0.118	0.535 ± 0.168
B1	Default XGBoost	0.601 ± 0.083	0.816 ± 0.070	0.777 ± 0.049	0.656 ± 0.059	0.832 ± 0.035	0.191 ± 0.051
B1	Tuned XGBoost	0.568 ± 0.058	0.798 ± 0.046	0.751 ± 0.066	0.625 ± 0.079	0.793 ± 0.066	0.174 ± 0.095
D0	InceptionTime	0.700 ± 0.075	0.859 ± 0.057	0.723 ± 0.092	0.655 ± 0.073	0.646 ± 0.133	0.671 ± 0.162
D0	TCN	0.694 ± 0.056	0.859 ± 0.063	0.797 ± 0.061	0.694 ± 0.058	0.855 ± 0.130	0.240 ± 0.236
D0	Transformer	0.664 ± 0.047	0.836 ± 0.035	0.706 ± 0.079	0.630 ± 0.060	0.630 ± 0.106	0.605 ± 0.152
D0	ResNet1D	0.680 ± 0.059	0.857 ± 0.045	0.744 ± 0.044	0.655 ± 0.047	0.698 ± 0.072	0.543 ± 0.103
D0	LSTM	0.543 ± 0.062	0.746 ± 0.029	0.700 ± 0.264	0.624 ± 0.161	0.787 ± 0.367	0.206 ± 0.387
M0	Transformer	0.698 ± 0.097	0.851 ± 0.072	0.745 ± 0.028	0.666 ± 0.025	0.681 ± 0.076	0.623 ± 0.222
M0	LSTM	0.677 ± 0.084	0.841 ± 0.056	0.676 ± 0.080	0.606 ± 0.066	0.582 ± 0.100	0.681 ± 0.132
M0	ResNet	0.669 ± 0.087	0.839 ± 0.059	0.580 ± 0.235	0.548 ± 0.149	0.501 ± 0.261	0.723 ± 0.305
M0	TCN	0.651 ± 0.066	0.834 ± 0.041	0.770 ± 0.097	0.676 ± 0.087	0.781 ± 0.155	0.397 ± 0.230
M0	InceptionTime	0.596 ± 0.057	0.803 ± 0.092	0.677 ± 0.144	0.593 ± 0.096	0.657 ± 0.261	0.450 ± 0.338

Table S1 provides the empirical basis for selecting one representative within each paradigm. Within the input-dimension-standardized SHAP-FS configurations, Random Forest and Logistic Regression had the highest mean AUROC for B0 and B1, respectively. The non-SHAP-FS tuned Random Forest had a marginally higher mean AUROC than B0 SHAP-FS Random Forest but similarly weak specificity, and is reported only as a configuration sensitivity reference. InceptionTime had the highest mean AUROC among the evaluated D0 candidates under the shared class-imbalance protocol and maintained substantially greater specificity than TCN, whose high sensitivity was accompanied by low and variable specificity. The 90-minute M0 backbone comparison favored the Transformer encoder, which was selected as the MIL representative.

Interpretation of the candidate-model comparisons is constrained by the small cohort, reuse of the same grouped folds for candidate selection and final evaluation, and high event prevalence, under which thresholded metrics can appear favorable despite weak specificity. Variants beyond the four selected models are therefore presented as exploratory configuration evidence rather than as additional primary comparisons.

S6. M0 WINDOW AND STRIDE SENSITIVITY

Table S2 summarizes the Transformer MIL window/stride sensitivity analysis under the five-channel, train-fold-normalized input set. These values are simple five-fold means used to select the M0 windowing configuration. The 90-minute window with a 45-minute stride had the highest mean AUROC and AUPRC among the evaluated settings, while the 180-minute setting was competitive but provided coarser temporal resolution.

TABLE S2

M0 TRANSFORMER WINDOW/STRIDE SENSITIVITY UNDER FIVE-CHANNEL TRAIN-FOLD NORMALIZATION. VALUES ARE SIMPLE FIVE-FOLD MEANS FOR SENSITIVITY SCREENING.

Window / stride	AUROC	AUPRC	F1	Accuracy	Sensitivity	Specificity
30 / 15 min	0.622	0.805	0.718	0.624	0.668	0.518
60 / 30 min	0.645	0.826	0.719	0.630	0.668	0.541
90 / 45 min	0.698	0.851	0.745	0.666	0.681	0.623
180 / 90 min	0.677	0.837	0.703	0.624	0.635	0.621

S7. HIGHER-BURDEN AGITATION ENDPOINT SENSITIVITY

The primary endpoint treats any PAS domain score above zero as next-day agitation. PAS rates four agitation domains individually from 0 to 4 [30]. As a stricter composite endpoint, Table S3 defines higher-burden agitation as a day-level PAS total score greater than 4, calculated as the sum of the maximum recorded motor, verbal, aggression, and resistance-to-care domain scores. This threshold requires agitation in at least two domains and at least one domain score above 1. The same 423 patient-nights were included, but the number of positive patient-nights decreased to 73 nights (17.3%). Under this lower-prevalence endpoint, D0 retained the highest AUROC, AUPRC, sensitivity, and balanced accuracy, whereas B1 achieved the

lowest Brier score and highest specificity. The sensitivity analysis shows partial robustness of the D0 ranking under a lower-prevalence, higher-burden endpoint, while demonstrating that model operating characteristics and the practical interpretation of predictions depend strongly on the agitation definition.

TABLE S3
HIGHER-BURDEN AGITATION ENDPOINT SENSITIVITY USING A DAY-LEVEL PAS TOTAL SCORE > 4 AS THE POSITIVE CLASS. METRICS ARE POOLED OOF POINT ESTIMATES ON 423 PATIENT-NIGHTS WITH 17.3% PREVALENCE.

Model	AUROC	AUPRC	Brier	Sens.	Spec.	BalAcc
B0	0.566	0.255	0.178	0.178	0.900	0.539
B1	0.636	0.280	0.164	0.260	0.911	0.586
D0	0.654	0.286	0.239	0.575	0.700	0.638
M0	0.635	0.267	0.179	0.274	0.877	0.576

S8. PER-PATIENT HETEROGENEITY

Within-patient discrimination was characterized for all four paradigms to distinguish pooled cross-patient performance from variation in patient-level ranking. Per-patient AUROC is defined only when a patient has at least one agitated and one non-agitated night. In this cohort, 37 of 65 patients met that condition, while 28 patients had single-class outcomes and therefore undefined within-patient AUROC. Among the 37 evaluable patients, patient-level time series remained short, with a median of 7 nights and a range of 3–10 nights per patient.

Figure S1 summarizes the resulting distribution. The horizontal line inside each box denotes the median patient-level AUROC, the green triangle denotes the mean, and the dashed horizontal line marks chance-level within-patient discrimination (AUROC=0.50). D0 had the highest median within-patient AUROC (0.60), followed by B0 (0.58), while B1 and M0 had medians of 0.50. The broad interquartile ranges and outlying low-AUROC patients indicate substantial heterogeneity, and within-patient discrimination remains uncertain over the available short follow-up.

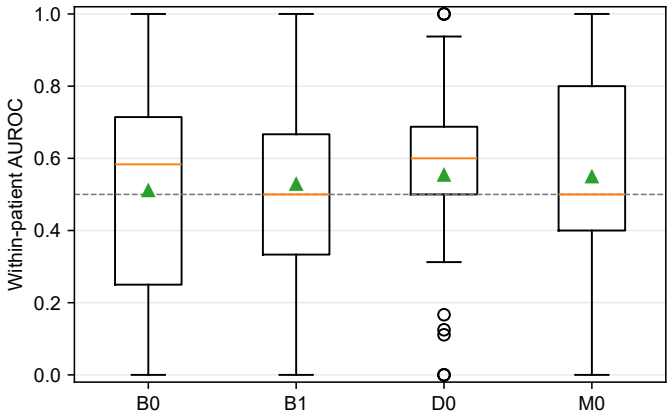


Fig. S1. Within-patient AUROC by paradigm. AUROC is computed only for patients with at least one agitated and one non-agitated night. Of the 65 patients, 37 were evaluable and 28 with single-class outcomes were excluded. Lines within boxes denote medians, green triangles denote means, and the dashed line marks chance-level AUROC=0.50.

S9. SINGLE-SUBJECT PREDICTION TRAJECTORIES

Patients with at least seven observed nights were eligible for illustration. The outcome-varying examples were selected deterministically as the patients with the largest numbers of non-agitated and agitated nights, respectively. The single-class examples were selected by longest follow-up, with ties resolved by greater cross-model prediction variability. These panels illustrate predefined patient-level behavior patterns and are not used for inferential comparisons. Subject-level trajectories use OOF probabilities to remain consistent with the held-out-patient primary evaluation and provide qualitative context for the heterogeneity summarized in Fig. S1. For the two outcome-varying patients, M0 generally assigned higher probabilities to agitation-positive than agitation-negative nights, whereas D0 showed limited separation. The single-class examples further demonstrate differences in absolute risk assignment: B0 consistently assigned high probabilities to the all-negative patient, while D0 frequently assigned probabilities below 0.5 to the all-positive patient. These examples illustrate that pooled discrimination can coexist with an inability to rank agitated above non-agitated nights within some patients, systematic risk-level differences, and unstable fixed-threshold classifications. The selected panels characterize possible model behaviors rather than estimate their prevalence across the cohort. The colored date labels mark the next-day agitation outcome associated with each preceding sensor night.

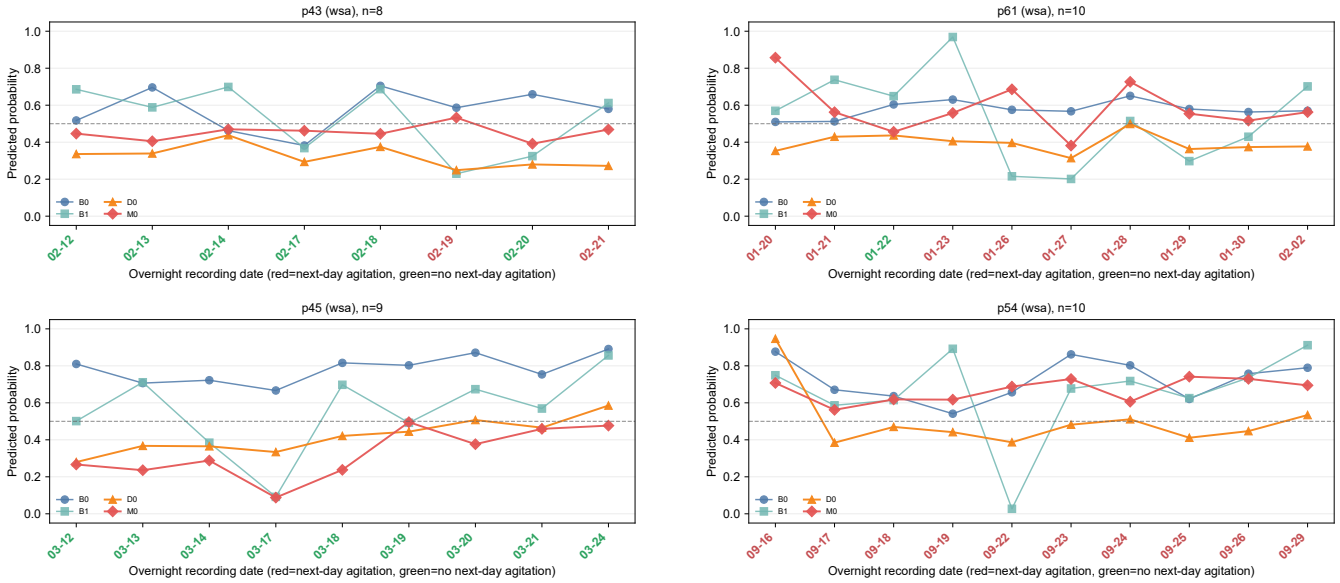


Fig. S2. Selected single-subject prediction trajectories from pooled OOF predictions. Patients with at least seven observed nights were eligible. The outcome-varying panels were selected by the largest numbers of non-agitated and agitated nights, respectively. The single-class panels were selected by longest follow-up, with ties resolved by greater cross-model prediction variability. Red and green date labels indicate agitated and non-agitated next-day outcomes, respectively. Each point is predicted from the preceding overnight sensor record.