SLEEPJ, 2022, 1–15

https://doi.org/10.1093/sleep/zsac001
Advance Access Publication Date: 4 January 2022
Original Article

ORIGINAL ARTICLE

Optimal spindle detection parameters for predicting cognitive performance

Noor Adra[1,2,3], Haoqi Sun[1,2,3,4], Wolfgang Ganglberger[1,2,3], Elissa M. Ye[1,2,3], Lisa W. Dümmer[1,2,3,5], Ryan A. Tesh[1,2,3], Mike Westmeijer[1,2], Madalena Da Silva Cardoso[1,2,3], Erin Kitchener[1,2,3,4], An Ouyang[1,2,3,4], Joel Salinas[4,6], Jonathan Rosand[1,2,3,4], Sydney S. Cash[1,4], Robert J. Thomas[4,7,†] and M. Brandon Westover[1,2,3,4,*,†]

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA, [2]Clinical Data Animation Center (CDAC), Boston, MA, USA, [3]Henry and Allison McCance Center for Brain Health at Mass General, Boston, MA, USA, [4]Harvard Medical School, Boston, MA, USA, [5]University of Groningen, Groningen, The Netherlands, [6]Department of Neurology, Center for Cognitive Neurology, New York University Grossman School of Medicine, New York, NY, USA and [7]Department of Medicine, Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

[†]Co-senior authors.

*Corresponding author. M. Brandon Westover, Department of Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA. Email: mwestover@mgh.harvard.edu.

Abstract

Study Objectives: Alterations in sleep spindles have been linked to cognitive impairment. This finding has contributed to a growing interest in identifying sleep-based biomarkers of cognition and neurodegeneration, including sleep spindles. However, flexibility surrounding spindle definitions and algorithm parameter settings present a methodological challenge. The aim of this study was to characterize how spindle detection parameter settings influence the association between spindle features and cognition and to identify parameters with the strongest association with cognition.

Methods: Adult patients (n = 167, 49 ± 18 years) completed the NIH Toolbox Cognition Battery after undergoing overnight diagnostic polysomnography recordings for suspected sleep disorders. We explored 1000 combinations across seven parameters in Luna, an open-source spindle detector, and used four features of detected spindles (amplitude, density, duration, and peak frequency) to fit linear multiple regression models to predict cognitive scores.

Results: Spindle features (amplitude, density, duration, and mean frequency) were associated with the ability to predict raw fluid cognition scores (r = 0.503) and age-adjusted fluid cognition scores (r = 0.315) with the best spindle parameters. Fast spindle features generally showed better performance relative to slow spindle features. Spindle features weakly predicted total cognition and poorly predicted crystallized cognition regardless of parameter settings.

Conclusions: Our exploration of spindle detection parameters identified optimal parameters for studies of fluid cognition and revealed the role of parameter interactions for both slow and fast spindles. Our findings support sleep spindles as a sleep-based biomarker of fluid cognition.

Statement of Significance

With the recent surge of sleep research examining sleep spindles and their role in various neuropsychiatric conditions, automated spindle detection algorithms have become increasingly popular research tools. However, the selection of automated spindle detector parameters may influence the subset of detected spindles and their relevance to cognitive networks remains unclear. This study confirms that detected spindles and their functional relevance for studies of cognition vary depending on the parameter settings used. Our findings provide reference spindle detection parameters for other studies and highlight the importance of both optimizing parameter settings and accounting for interactions in studies of associations between spindles and cognition.

Key words: sleep spindle; EEG; cognition; sleep

Submitted: 16 April, 2021; **Revised:** 7 December, 2021

Introduction

Brain health is an emerging research focus that describes "the preservation of optimal brain integrity and mental and cognitive function at a given age in the absence of overt brain diseases that affect normal brain function" [1]. Because sleep disturbances [2] and neurocognitive diseases [3] are associated with heightened morbidity and mortality in older adults, sleep has become an increasingly popular therapeutic target for brain health. Changes in sleep macrostructure, including decreased sleep efficiency and duration and increased nighttime wakefulness and sleep fragmentation, have been linked to aging [4, 5] and an increased risk of cognitive impairment [6, 7].

More recently, age-related changes in sleep microstructure have also been described [8] and are notably more pronounced than changes in sleep macrostructure [9]. These findings have led to a surge of research using electroencephalogram (EEG) signals to extract sleep microfeatures, particularly sleep spindles, as electrophysiologic markers of neurodegenerative and psychiatric diseases [10–12]. Spindles are a hallmark of non-rapid eye movement (NREM) stage 2 sleep and are characterized by intermittent waxing and waning 11–16 Hz oscillations that last 0.5 to 3 s. Spindle mechanisms play a role in memory consolidation and relative sensory deafferentation of sleep, function as information carrier waves, and contribute to the cohesion of sleep itself. Although they vary greatly across individuals, spindles are a NREM EEG "fingerprint" [13] and show high heritability [14] and strong night-to-night stability [15].

Although spindles are sometimes visually recognizable, the spindle frequency range defined in the literature starts anywhere between 8 and 12 Hz [16, 17] and ends between 15 and 17 Hz [18, 19]. While it is commonly accepted that the spindle minimum duration is 0.5 s, the origins of this criterion are also unclear and not physiologically based, with studies reporting spindles as brief as 0.3 s [18, 20, 21]. Such open-endedness may not make a difference in the clinical practice of sleep medicine, where the primary use of spindle detection is scoring of NREM sleep. However, the use of spindle analysis as a precision tool to identify electrophysiologic markers of specific health outcomes requires more stringent and physiologically relevant criteria.

Spindles are further classified into "slow" and "fast" subtypes as a function of EEG topographical and frequency distributions. Slow spindles are preferentially detected in frontal regions while fast spindles are localized to central and parietal regions [22]. Similar to the definition of spindles themselves, there is no consensus on the delineation of these two subtypes, which has ranged from 12 Hz to 14 Hz across studies [9, 16, 17].

The current gold standard for spindle detection is visual inspection of the EEG. However, manual spindle detection by experts is time consuming, prone to errors, costly, limited to small data sets, and suffers from interrater variability due to subjective spindle definitions and varying expertise [20, 23]. Automatic detection methods, by contrast, allow for reproducible analysis and are being optimized to concur with visual inspection. This approach, however, suffers from the lack of consensus regarding spindle definitions and presumes that the physiological import of the spindle correlates with the current gold standard of visualization. Further, it overlooks potential associations between specific parameter settings of automated detection algorithms and the physiological relevance of detected spindles. Here, we hypothesized that parameter settings influence the association between sleep spindle features and cognition. To evaluate this hypothesis, we used cognitive measures as the gold standard and identified spindle detection parameters that best correlate with cognitive performance.

Methods

Participants

Adult patients (≥ 18 years of age) referred to the Massachusetts General Hospital Sleep Laboratory between November 2018 and October 2019 for overnight diagnostic polysomnography (PSG) were offered the option to return within 40 days of their PSG exam to complete a cognitive test battery. Participants provided written informed consent for participation in the study, which was approved by the Mass General Brigham Institutional Review Board.

Exclusion criteria included any baseline diagnosis of dementia or learning disability, prior exposure to the cognitive test battery, or inability to complete the battery due to English language non-fluency or motor, visual, or hearing impairments. Use of benzodiazepines was not an exclusion criterion. However, given their ability to increase spindle activity [24], we performed sensitivity analysis to quantify the effect of benzodiazepine usage on the overall results. Medication usage was collected through self-reports on the night of the PSG.

Cognitive test battery

All participants completed the NIH Toolbox Cognition Battery [25], one of the four core domains of the NIH Toolbox for Assessment of Neurological and Behavioral Function [26]. The NIH Toolbox Cognition Battery contains seven subtests, each of which measures a cognitive subdomain (Supplementary Table S1). Five subtests are categorized under fluid cognition: Dimensional Change Card Sort (DCCS), Flanker Inhibitory Control & Attention (ICA), List Sorting Working Memory (LSWM), Pattern Comparison Processing Speed (PCPS), and Picture Sequence Memory (PSM) tests. The other two subtests are categorized under crystallized cognition and assess language: Picture Vocabulary (PVT) and Oral Reading Recognition (ORR) tests. Absolute (uncorrected) and age-corrected standard scores are generated for each individual subtest and three composite cognitive measures: total cognition, fluid cognition, and crystallized cognition. Higher scores indicate better cognitive function for each cognitive domain. In adults, individual subtests have shown good test-retest reliability [26] and composite scores have shown good internal consistency, excellent test-retest reliability, and strong convergent and discriminant validities [27].

EEG recordings and signal processing

EEG and ECG recordings were collected from six channels during diagnostic PSG. EEG signals were recorded at 512 Hz and segmented into nonoverlapping 30-second epochs, which were manually scored by licensed sleep technicians following American Academy of Sleep Medicine standards [28] as part of clinical care. Each epoch was assigned one of the following labels: wake (W), rapid eye movement, Non-REM stage 1 (N1), Non-REM stage 2 (N2), or Non-REM stage 3 (N3). A sleep neurologist manually reviewed and checked sleep scores for quality control.

We extracted EEG signals from central channels referenced to the contralateral mastoid (C3-M2 and C4-M1) and a single ECG channel. EEG and ECG signals were resampled to 256 Hz and filtered with a zero-phase band-pass filter from 0.3 Hz to 35 Hz and 0.3 Hz to 40 Hz, respectively. A previous study showed that spindle density was higher across different studies when cardiac interference was removed from EEG signals [29]. Based on these findings, we applied ECG-correction to EEG signals. To remove ECG artifacts, we calculated a subject's instantaneous heart rate, smoothed it using a window size of 3 s, and detected R peaks using the Pan-Tompkins algorithm [29, 30]. The R peaks were then aligned with the EEG to create an average signature that was then subtracted from the EEG signal, as described previously [29]. We then retrieved the subject's average EEG signature by averaging over intervals of 0.5 s and subtracted any EEG signature that aligned with an R peak.

Following ECG correction, EEG signals were bandpass filtered between 0.1 Hz and 20 Hz. Signal artifacts were then detected and removed through a previously described filtering method [31]. Finally, to remove any extreme outliers, a threshold for outlier detection was set iteratively to remove epochs that were more than 3 standard deviations above or below the mean of four per-epoch summary metrics (the root mean square and three Hjorth parameters [32]). Only epochs from NREM stage 2 sleep were included for analysis.

Automated spindle detection

Sleep spindle features were generated using Luna (http://zzz.bwh.harvard.edu/luna/). The automated spindle detector component of Luna relies on a previously published Morlet wavelet transformation [33] that has been comprehensively evaluated [20]. Seven parameters were chosen for exploration: (1) cycles, which improve frequency resolution at higher values and improve temporal resolution at lower values; (2) target frequency (fc); (3) spindle quality metric (*q*); (4) multiplicative threshold for spindle detection; (5) minimum spindle duration (min); (6) maximum spindle duration (max); and (7) the maximum number of seconds between spindles across which they should be considered a single spindle and therefore merged, unless the resulting spindle is larger than the max parameter (merge). The range of exploration for each parameter is listed in Supplementary Table S2.

Statistical analyses

Pairing all possible parameter values resulted in 29 393 280 possible combinations. We sampled 1000 random combinations using Monte Carlo-based selection and ran Luna with each setting to extract spindle features. This method of parameter combinations selection was chosen because exploring all possible combinations of parameters is not possible, and systematic grid search experiments are known to be less efficient than random sampling [34]. A second exploratory analysis of 500 combinations was completed following the preliminary findings. Spindle features were averaged across both channels (C3-M2 and C4-M1) for each patient. Four spindle features of interest were selected a priori for cognitive performance prediction: spindle amplitude, density, duration, and mean spindle frequency. All cognitive measures were standardized (z-transform) prior to analysis.

Using these features, we fitted a linear regression model to predict cognitive performance on each of the NIH Toolbox subtests and on the overall composite uncorrected standard scores. Pearson's correlation was performed to compare measured cognitive scores with cognitive scores predicted by the optimized regression model. To avoid overfitting, we evaluated model performance using 10-fold cross validation. In detail, the dataset was randomly split into 10 folds where each fold contains the same number of PSGs (or approximately the same if not divisible by 10). A linear regression model was trained on 9 folds and tested on 1-fold, and this process rotated to ensure the whole dataset was used for testing. The reported correlation is the average of the 10 correlations across the 10 testing folds. Finally, a model was refitted on the whole dataset to get a single overall set of model coefficients. To generate 95% confidence intervals of the correlations and coefficients, we repeated the above process 1000 times using bootstrapping (sampling with replacement to create bootstrapped datasets of the same size as the original dataset). Confidence intervals are defined using the nonparametric 2.5th–97.5th percentiles of all bootstrapped estimates. A representative histogram of the bootstrapped values (and the point estimate from actual dataset) for the Pearson's correlations for fluid intelligence using the best parameter combination is shown in Supplementary Figure S1.

A modified percentile bootstrap method [35] was then performed to compare Pearson's correlation across models. This method accounts for any heteroscedasticity and non-normality in the data. To apply this method, we first identified the model with the largest Pearson's correlation value and then applied the bootstrapping method to calculate the confidence intervals of the differences between the best model and each of the remaining models. The top model was considered significantly better if the confidence interval comparing the two models did not include zero. Our first aim was to identify spindle detection parameters that best measured the raw association between spindles and cognition, regardless of confounding factors, using absolute (uncorrected) scores for data analysis. Our second aim was to identify the independent association between spindles and cognition after adjusting for age, using age-corrected scores, to explain the variation in cognition explained by spindle features after accounting for age.

To examine the role of parameter interactions in model performance, we visually inspected grouping behaviors for parameter combinations using a dimensionality reduction tool, t-distributed stochastic neighbor embedding (t-SNE [36, 37]). We also examined whether including spindle feature regression coefficients changed t-SNE grouping behaviors. Finally, we compared these t-SNE plots to a third that only included average spindle features for each combination. To identify the best hyperparameters for visualization, the perplexity parameter in t-SNE was varied and the best values were selected visually. All other parameters were set to default options. Parameter interactions were also examined through parallel coordinates [38]. For each parameter, a Cuzick test for trend [39] was computed to measure the trend of performance across parameter values. Following visualization, a final run of 495 selected parameter combinations were analyzed. Spindle features for top combinations at 12.5 Hz and 14.5–16 Hz were visually inspected by age and cognitive performance through scatter plots.

To evaluate our results in cognitively impaired patients, we reviewed polysomnograms acquired in the Sleep Laboratory at

Massachusetts General Hospital from 2009 to 2017 and identified patients with dementia and mild cognitive impairment (MCI). Dementia was defined, as in our prior work, using the following criteria: prescription of one or more dementia-related medication, record of dementia diagnosis in the patient's active, reoccurring medical problem list, Montreal Cognitive Assessment score ≤ 19 prior to the sleep study and <27 after the sleep study, or Mini-Mental State Examination score < 25. Criteria for MCI included record of MCI in the patient's active, reoccurring medical problem list or a Montreal Cognitive Assessment score between 20 and 25 prior to the sleep study and <27 after the sleep study. For both groups, exclusion criteria were age <50 years or history of developmental delay, brain tumor, neoplasm, stroke, brain injury/trauma, or seizure prior to the sleep study. Our final dataset included 215 dementia patients and 308 MCI patients. Spindle features (amplitude, density, duration, frequency) were extracted using the best and worst performing models. We investigated the association between spindle features and cognitive impairment status using a Cuzick test for trend.

External validation was performed using the Sleep Heart Health Study [40–43] (SHHS), which is a composite cohort overlapping with the Framingham Heart Study [44] (FHS). Participants were included if they completed a neuropsychological test battery [45] in the FHS within three years of their SHHS PSG exam date. Participants with incomplete Wechsler Memory Scale (WMS) data or who were diagnosed with cognitive impairment were excluded from analysis. A total of 476 participants were included in the validation dataset. Model performance was evaluated for the 1000 randomly selected parameter combinations from the preliminary stage, along with the best, default and worst combination parameters.

When examining associations between continuous variables, Pearson correlations were performed. Statistical significance of estimated associations was defined based on confidence intervals, presented following the format X [Y, Z]. Associations were considered statistically, significantly different from zero when their 95% interval did not include zero. To determine associations across discrete variables, a Cuzick's test for trend was used. For the Cuzick's test for trend, statistical significance was defined using a p-value < .05. All statistical analyses were performed using code written in-house using Python (https://www.python.org/). Source code used to produce the figures and complete regression analysis is available on our GitHub page: https://github.com/mghcdac/spindle_optimization.

Of note, Luna provides recommended default parameter values for spindle detection. Our goal in this work was to identify which part of the parameter space maximizes correlations with cognition rather than to specifically evaluate the Luna default parameters. Nevertheless, the reference parameters represent an important reference point, thus we provide comparisons with the default parameters throughout the manuscript.

Results

Patient characteristics

We enrolled 167 participants. Six were subsequently excluded: two participants did not complete cognitive measures and four had poor EEG signal quality. The final sample included 161 participants (89 women) with a mean age of 49 years. On the night of the PSG, 23 (14%) patients reported regular use of benzodiazepines. Participant demographics are described in Table 1. Score distributions on the NIH Toolbox cognitive battery and performance by age across all the subtests and composite tests are shown in Figures 1 and 2, respectively.

Overall performance trends

For fluid intelligence, 632/1000 models significantly predicted cognitive scores and had 95% confidence intervals that did not contain the null hypothesis value for either the correlation or regression coefficients. Of these models, 286 (45.25%) showed moderate correlation values (≥0.4). The best performing model (0.501 [0.175, 0.595]) had the following spindle detector parameters: cycles = 6, central frequency (fc) = 15.5 Hz, quality metric (q) = 0.3, threshold = 5.5, minimum spindle duration = 0.4 s, maximum spindle duration = 2.6 s, merge = 0.7 s. Model details and performance are shown in Table 2 and Figure 3, respectively. Specification curve analysis is shown in Supplementary Figure S2. Using a modified percentile bootstrap method to compare Pearson's correlations across all models, we found that the best performing model performed significantly better than 395 (40%) models. Of the 605 remaining models which did not significantly differ in accuracy, 360 (60%) were characterized by fast spindle activity (FFT average > 13Hz).

After examining each fluid subtest, we found that 34/62 (55%) models with similar performance were characterized by fast spindle activity for the Flanker ICA test, 116/280 (41%) for LSWM, 221/290 (76%) for DCCS, 262/458 (57%) for PCPS, and 337/528 (64%) for PSM. The best performing model for each subtest is shown in Table 2 and Figure 3, respectively.

With respect to crystallized intelligence, all models poorly predicted composite and subtest scores (Supplementary Figure S3). The parameter combination with the best performance showed negative correlation between predicted and true composite (–0.38 [–0.47, 0.06]), PVT (–0.39 [–0.48, 0.06]), and ORR (–0.35 [–0.45, 0.01]) scores (Figure 3).

For total intelligence, 299/1000 models significantly predicted cognitive scores and had 95% confidence intervals that did not contain the null hypothesis value for either the correlation or regression coefficients. Of these models, three (1%) showed moderate correlation values (≥0.4). The best performing model (0.407 [0.066, 0.495]) had the following spindle detector parameters: cycles = 7, central frequency (fc) = 14 Hz, quality metric (q) = 0.7, threshold = 6, minimum spindle duration = 0.4 s, maximum spindle duration = 3.5 s, merge = 1 s. Model details and performance are shown in Table 2 and Figure 3, respectively. Specification curve analysis is shown in Supplementary Figure S4. Using a modified percentile bootstrap method to compare Pearson's correlations across all models, we found that the best performing model performed significantly better than 301 (30%) models. Of the 699 models with similar performance, 203 (29%) were characterized by fast spindle activity (FFT average>13Hz).

Potential effect of benzodiazepine use on cognition for the best parameter combination was assessed using linear regression. The coefficient of benzodiazepine use (25.00 [–0.20, 43.54]) and its interaction with predicted scores (–25.34 [–45.30, 0.67]) were not statistically significant. Similarly, the effect of AHI on cognition for the best parameter combination was not significant in terms of its model coefficient (–9.33 [–38.00, 20.04]) or in its interaction with predicted scores (7.47 [–20.98, 34.40]). Although not statistically significant, the net effect of benzodiazepines,

Table 1. Patient characteristics (*N* = 161)

	Women (*N* = 89)	Men (*N* = 72)
Age, years, mean ± SD‡	46.8 ± 17.29	52.6 ± 18.52
Years of education, mean ± SD	16.0 ± 2.88	16.8 ± 3.11
Race		
White	63 (70.8%)	59 (81.9%)
Black	7 (7.9%)	4 (5.6%)
Hispanic/Latino	1 (1.1%)	0 (0%)
Asian	5 (5.6%)	4 (5.6%)
Multiracial	13 (14.6%)	5 (6.9%)
Hollingshead index, mean ± SD	47.5 ± 14.4	53.4 ± 13.9
Employed or self-employed, *n* (%)	49 (55%)	38 (53%)
Marital status (married), *n* (%)	36 (40%)	36 (50%)
Benzodiazepines	13 (14.6%)	10 (13.9%)
Current smoker, *n* (%)	2 (2%)	4 (6%)
AUDIT, median (IQR)‡	2.0 [0,3]	2 [0,4]
History of alcohol abuse, *n* (%)	6 (7%)	11 (15%)
History of substance abuse, *n* (%)	9 (10%)	12 (17%)
Body mass index, kg/m^2, median (IQR)	28.3 [23.6, 33.0]	28.4 [25.4, 32.5]
Diabetes, *n* (%)	4 (4.5%)	6 (8.3%)
Cardiovascular disease, % (*n*)	6 (6.7%)	8 (11.1%)
Charlson Comorbidity Index, median (IQR)	1.0 [0.0, 3.0]	1.0 [0.0,3.0]
PHQ-4, median (IQR)	2.0 [1.0, 4.0]	2 [0.0, 5.0]
Family history of dementia, n (%)*	41 (46%)	25 (35%)
PSQI, median (IQR)	9.29 [5, 12.8]	8.30 [5, 10]
AHI, median (IQR)§	2.2 [0.5, 5.7]	6.1 [1.8 12.8]
Normal (< 5), *n* (%)	64 (71.9%)	29 (40.3%)
Mild sleep apnea (5 ≤ AHI < 15), *n* (%)	21 (23.6%)	26 (36.1%)
Moderate sleep apnea (15 ≤ AHI < 30), *n* (%)	4 (4.5%)	12 (16.7%)
Severe sleep apnea (AHI ≥ 30), *n* (%)	0 (0%)	6 (8.3%)
PSG referral reasons		
Sleep apnea evaluation	55 (62%)	55 (76%)
Sleepiness	43 (48%)	35 (49%)
Insomnia	30 (34%)	21 (29%)
Non-restorative sleep	3 (3%)	1 (1%)
Restless legs syndrome	15 (17%)	18 (25%)
REM sleep behavior disorder	1 (1%)	1 (1%)

AUDIT, Alcohol Use Disorders Identification Test; AHI, apnea-hypopnea index (# apnea events per hour of sleep) at 4% desaturation for hypopnea. IQR, interquartile range; PHQ-4, Patient Health Questionnaire for Depression and Anxiety; PSQI, Pittsburgh Sleep Quality Index; SD, standard deviation. ‡<0.05; § <0.0001.
*Family history of one patient was unknown as patient was adopted.
‡<0.05.
§ <0.0001.

defined by the sum of the coefficient of benzodiazepine use (25) and its interaction with predicted scores (–25*predicted score), is negative and, thus, is correlated with worse predictions of cognition. This finding aligns with clinical findings that show altered sleep architecture profiles during use of benzodiazepines.

Isolating parameter settings from performance trends

To understand what drove specific parameter combinations to excel when predicting fluid cognition, we performed exploratory predictor analysis. Our first exploratory question was whether interactions between parameters were negligible. When visually inspected, certain parameter settings exhibited consistently poor performance (small correlation coefficients), such as combinations with fc < 11.5 Hz (Supplementary Figure S5). A Cuzick test for trend showed a significant trend of performance across parameter values for cycles, central frequency, and minimum spindle duration (cycles: $p = .002$; central frequency: $p < 2.2e\text{-}16$; minimum duration: $p = .02$; maximum duration: $p = .46$; quality metric: $p = .22$; threshold: $p = .19$; merge: $p = .87$).

Effect of parameter settings and interactions

t-SNE maps generated similar cluster arrangements with respect to both shape and relative location of Pearson's correlation values for three-dimensionally reduced features. Overall, t-SNE map topographies suggested discrete boundaries between models that poorly predicted cognition (small correlation coefficients) and moderately predicted cognition (larger correlation coefficients) when different parameter values were used (Figure 4A), and the delineation slightly improved with the addition of regression coefficients and when only average spindle features were used (Figure 4B and C).

To visualize parameter interactions, we used parallel coordinates plots. For slow spindle combinations, we observed that model performance peaked at fc = 12.5 Hz. When fc > 12 Hz, combinations with minimum spindle duration < 6 s and threshold

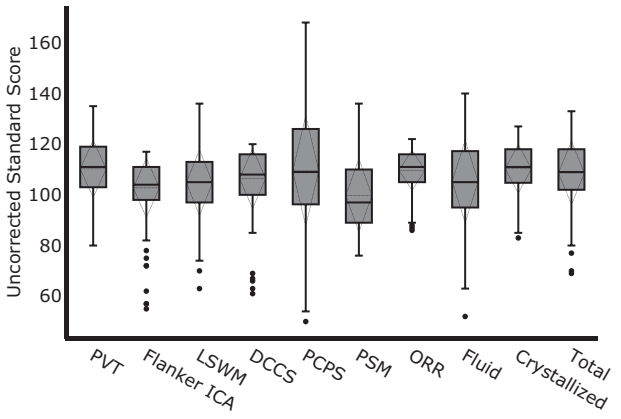

Figure 1. Patient performance on the NIH Toolbox Cognition Battery. Box plot of the absolute scores ($N = 161$) for each subtest and composite cognitive measures. The box signifies the upper and lower quartiles. The median is represented by a short black line within the box for each cognitive measure, while the mean is represented by a dashed line within each box. Standard deviation is represented by the dashed diamond and outliers are depicted by black dots. Greater variability in performance was seen for the fluid subtests (second to sixth boxes). Patients completed the cognitive assessment between 2 and 40 days following their polysomnography visit. DCCS, Dimensional Change Card Sort; ICA, Inhibitory Control & Attention; LSWM, List Sorting Working Memory; ORR, Oral Reading Recognition; PCPS, Pattern Comparison Processing Speed; PSM, Picture Sequence Memory; PVT, Picture Vocabulary Test.

values > 4 generally corresponded with better model performance. An inverse relationship was also seen between cycles and fc: model performance improved at lower cycles for higher fc values and at higher cycles for lower fc values. No clear parameter interactions or model performance patterns were found for maximum spindle duration, *q*, or merge settings.

Because parameter combinations were sampled randomly during preliminary analysis, the reliability of certain patterns was unclear. For example, although lower fc values appeared to favor higher cycles, we lacked parameter combinations that had cycles = 6, fc = 12.5 Hz, and threshold > 4. To address these uncertainties, we ran a select range of combinations (n = 500) after parallel coordinates visualization (Supplementary Table S3). Because *q* showed no trend and did not have any missing values of interest, it was set to zero for these runs. The top performing model from this run showed slightly improved performance 0.504 [0.198, 0.594]) and had the following parameters: cycles = 5, fc = 15, *q* = 0, threshold = 5, minimum spindle duration = 0.4 s, maximum spindle duration = 3.9 s, and merge = 0.7. Final inspection revealed that model performance consistently peaked for slow spindles when fc = 12.5, cycles = 12, threshold > 4.5, and minimum spindle duration < 0.5 s (Figure 5B). Fast spindle parameter combinations were less restrictive and resulted in larger correlations when fc > 14 Hz. At 14.5 Hz, performance improved when cycles = 7, threshold > 4.5, and minimum spindle duration < 0.5 s. Performance peaked between when fc was between 15 and 15.5 Hz, cycles = 5, and minimum spindle duration < 0.5 s. At fc = 15.5 Hz, no clear distinction was observed for thresholds > 4, while minimal improvement was seen at a threshold of 5 compared to 5.5 at fc = 15 Hz (Figure 5C).

Key predictive spindle features

The distribution of average spindle features across frequencies are shown in Supplementary Figure S6. When examining the top parameter combination for each central frequency, we a found a discrepancy between measured frequency values (FFT) and central frequency (fc) settings greater than 14 Hz (Table 3). Spindles detected by these combinations had an average FFT-based central frequency approximately 14 Hz.

Plotting spindle features detected by the top five parameter combinations from Table 3 against age revealed that spindle density declined with age across fast and slow spindle frequencies. However, in older patients, fast spindles resulted in a slightly higher frequency relative to younger patients, despite the decreased spindle density, and worse performance (smaller correlations) on fluid cognitive tests was linked with higher fast spindle FFT averages (Figure 6).

Age-adjusted performance trends

For our second aim, we used age-corrected scores to evaluate all parameter combinations. For fluid intelligence, 141/1500 models (1000 preliminary and 500 exploratory combinations) significantly predicted cognitive scores and had 95% confidence intervals that did not contain the null hypothesis value for either the correlation or regression coefficients. Of these models, 29 (20.56%) showed moderate correlation values (≥ 0.3). The best performing model (0.315 [0.109, 0.402]) had the following spindle detector parameters: cycles = 8, central frequency (fc) = 15 Hz, quality metric (*q*) = 0, threshold = 5.5, minimum spindle duration = 0.2 s, maximum spindle duration = 3.9 s, merge = 0.7s. Using a modified percentile bootstrap method to compare Pearson's correlations across all models, we found that the best performing model performed significantly better than the 859 (86%) nonsignificant models. Of the 141 remaining models which did not significantly differ in accuracy, 110 (78%) were characterized by fast spindle activity (FFT average > 13 Hz). Specification curve analysis is shown in Supplementary Figure S2.

When examining fluid subtests, we found only one significant model for the DCCS test, which was characterized by fast spindle activity. Additionally, 122/125 (98%) models with similar performance were characterized by fast spindle activity for PCPS and 89/97 (92%) for PSM tests. All models correlated poorly with Flanker ICA; including the model with the strongest correlation (–0.34 [–0.45, 0.05]). The best performing model (largest correlation coefficient) for LSWM showed a trend for significance (0.25 [–0.02, 0.36]).

Both crystallized (Supplementary Figure S3) and total (Supplementary Figure S4) intelligence composite and subtest scores were poorly predicted by all models. The parameter combination with the best performance showed negative correlation between predicted and true total composite (–0.34 [–0.53, 0.14]), crystallized composite (–0.40 [–0.45, 0.01]), PVT (0.31 [–0.45, 0.02]), and ORR (–0.38 [–0.49, 0.04]) scores.

Comparing performance trends for best and default settings

When absolute scores were used, the following default parameter combinations and resulted in better predictive performance (larger correlation coefficients): total (0.32 [0.02, 0.42]), fluid (0.43 [0.11, 0.52]), LSWM (0.31 [0.04, 0.42]), PCPS (0.37 [0.05, 0.47]), PSM (0.37 [0.05, 0.48]). In contrast, crystallized (–0.18 [–0.41, 0.10]), PVT (–0.21 [–0.41, 0.08]), Flanker ICA (0.26 [–0.04, 0.38]), DCCS (0.27 [–0.004, 0.378]), and ORR (–0.09 [–0.36, 0.11]) tests resulted in poor model performance (small correlation coefficients). Using a modified percentile bootstrap method to compare Pearson's correlations across all models, we only found a difference between

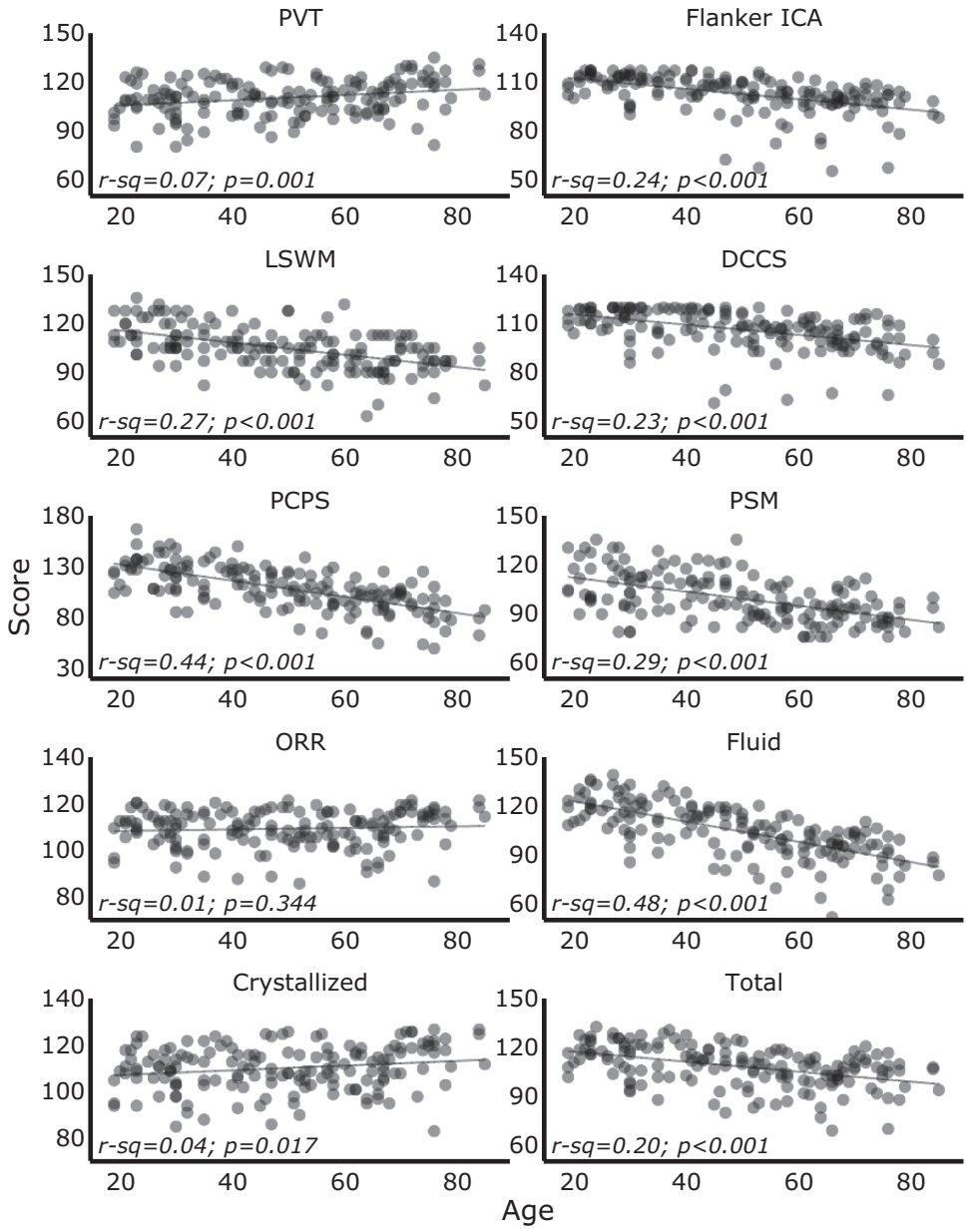

Figure 2. Performance on each subtest and composite measure on the NIH Toolbox Cognition Battery by age. Scatter plots of the absolute scores (N = 161) for each subtest and composite cognitive measures. Crystallized scores showed poor positive correlations with age, which appears to be driven by the PVT subtest. Fluid scores showed moderate to strong negative correlations with age. DCCS, Dimensional Change Card Sort; ICA, Inhibitory Control & Attention; LSWM, List Sorting Working Memory; ORR, Oral Reading Recognition; PCPS, Pattern Comparison Processing Speed; PSM, Picture Sequence Memory; PVT, Picture Vocabulary Test.

Table 2. Top performing model predicting total and fluid composite and subtest scores on the NIH Toolbox Cognition Battery

								FFT (Hz)		Duration (s)		Density (spm)		Amplitude (uV)		Pearson's *r*	
Test	cycles	fc (Hz)	q	th	min (s)	max (s)	merge (s)	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
Total Comp.	7	14	0.7	6	0.4	3.5	1	−2.40	[−4.26,−0.61]	2.63	[0.27, 4.56]	3.06	[1.14, 5.18]	1.05	[0.21, 2.49]	0.41	[0.07, 0.50]
Fluid Comp.	6	15.5	0.3	5.5	0.4	2.6	0.7	−6.48	[−9.30, −3.72]	0.49	[−2.41, 3.46]	4.30	[1.88, 6.93]	1.60	[0.13, 3.73]	0.50	[0.18, 0.60]
Fluid Comp.*	5	15	0	5	0.4	3.9	0.7	−5.08	[−7.42, −2.57]	−2.51	[−4.98, 0.29]	7.15	[4.67, 9.57]	2.31	[−0.05, 4.64]	0.50	[0.20,0.59]
PSM	7	16	0.6	3.5	0.9	3.1	0.8	−7.19	[−10.00, −4.76]	2.67	[0.54, 4.21]	2.13	[−0.52, 4.29]	1.79	[0.65, 3.31]	0.46	[0.11, 0.56]
DCCS	5	15.5	0	4.5	0.4	3.9	0.7	−4.13	[−6.00, −2.12]	−2.90	[−4.94, −0.32]	3.29	[1.52, 5.01]	2.28	[−0.85, 4.11]	0.45	[0.12, 0.57]
PCPS	5	15.5	0	4.5	0.4	3.9	0.7	−6.86	[−10.17, −3.18]	−5.54	[−9.00, −1.84]	7.42	[4.06, 10.49]	4.25	[1.07, 7.46]	0.47	[0.15, 0.57]
Flanker ICA	12	12.5	0	5	0.2	2.6	0.5	2.02	[0.93, 3.02]	0.19	[−1.03, 1.52]	3.21	[1.80, 4.59]	0.55	[−0.16, 1.14]	0.36	[0.15, 0.42]
LSWM	11	13	0.7	2.5	0.3	2.9	1	−0.46	[−2.25, 1.24]	4.55	[1.88,7.11]	1.36	[−1.21, 3.82]	0.25	[−1.58, 1.63]	0.37	[0.11, 0.46]

Linear multiple regression was used to predict cognition from 4 sleep spindle features: amplitude, density, duration, and FFT (mean spindle frequency). Pearson's correlation was then performed to compare measured cognitive scores with cognitive scores predicted by the optimized regression model. Sleep spindle features were generated using Luna.

comp, composite; fc, central frequency; DCCS, Dimensional Change Card Sort; ICA, Inhibitory Control & Attention; L, Lower; LSWM, List Sorting Working Memory; q, quality metric; PCPS, Pattern Comparison Processing Speed; PSM, Picture Sequence Memory; th, threshold; U, Upper.

*Top performing model from the final Luna run, which showed slight improvement in model performance.

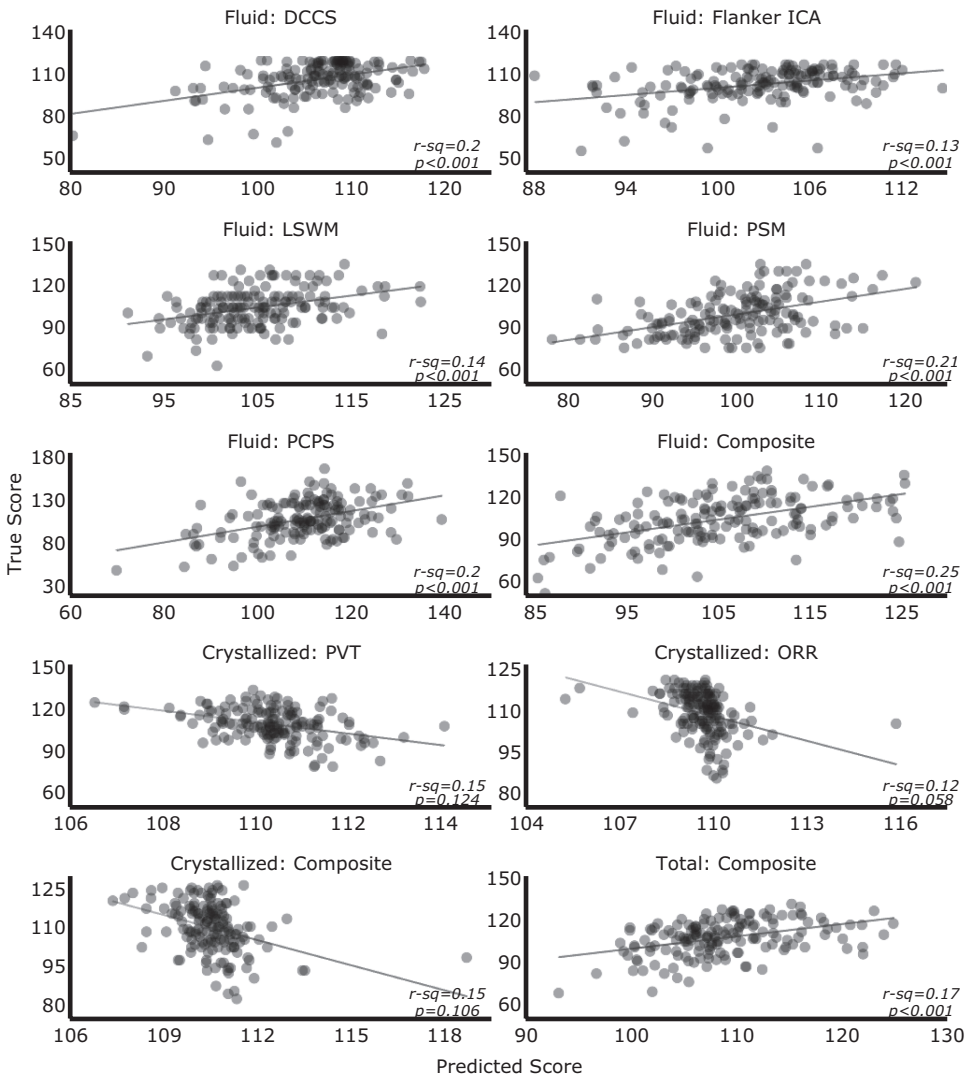

Figure 3. Sleep spindles moderately predict fluid cognition and poorly predict crystallized cognition. Scatter plots of the absolute scores (N = 161) and predicted scores for each subtest and composite cognitive measures on the NIH Toolbox Cognition Battery. True cognitive scores are compared with cognitive scores predicted by the optimized regression model for each cognitive test and measure. Sleep spindle features were generated using Luna. DCCS, Dimensional Change Card Sort; ICA, Inhibitory Control & Attention; LSWM, List Sorting Working Memory; ORR, Oral Reading Recognition; PCPS, Pattern Comparison Processing Speed; PSM, Picture Sequence Memory; PVT, Picture Vocabulary Test.

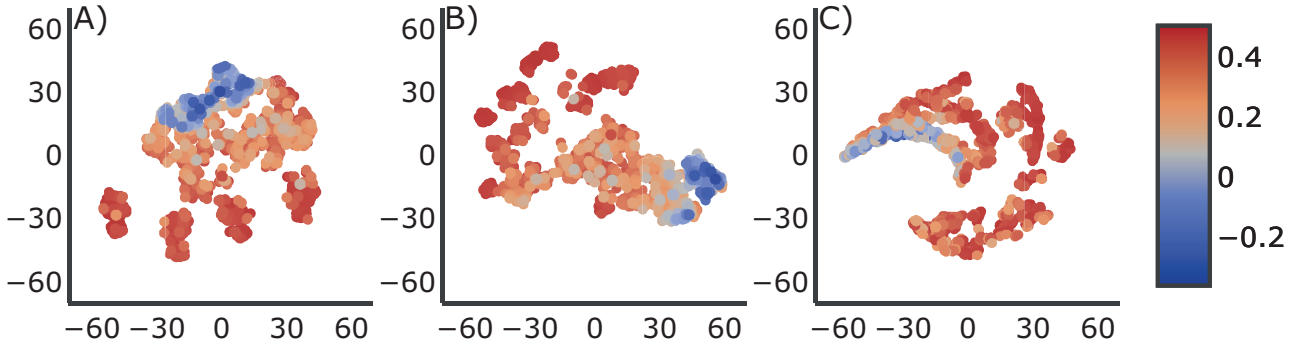

Figure 4. t-distributed stochastic neighbor embedding (t-SNE) visualization shows discrete boundaries between poor model performance and good performance defined by Pearson correlation values. t-SNE map topographies were generated using (A) parameter combinations, (B) parameters combinations with regression coefficients for the four spindle features, and (C) spindle features. The color bar represents Pearson *r* values. Hyperparameters were varied until a good visualization was obtained. Perplexity was set to 30.

the best performing and default parameter models for crystallized cognition and PVT scores: total (–0.09 [–0.6, 0.05]), fluid (–0.08 [–0.35, 0.21]), crystallized (–0.37 [–0.56, –0.06]), PVT (–0.43 [–0.60, –0.01]), Flanker ICA (–0.10 [–0.37, 0.09]), LSWM (–0.06 [–0.33, 0.20]), DCCS (–0.18 [–0.45, 0.13]), PCPS (–0.09 [–0.40, 0.19]), PSM (–0.09 [–0.39, 0.23]), and ORR (–0.28 [–0.56, 0.11]).

When age-adjusted scores were used, all default parameter combinations showed weak performance (small correlation

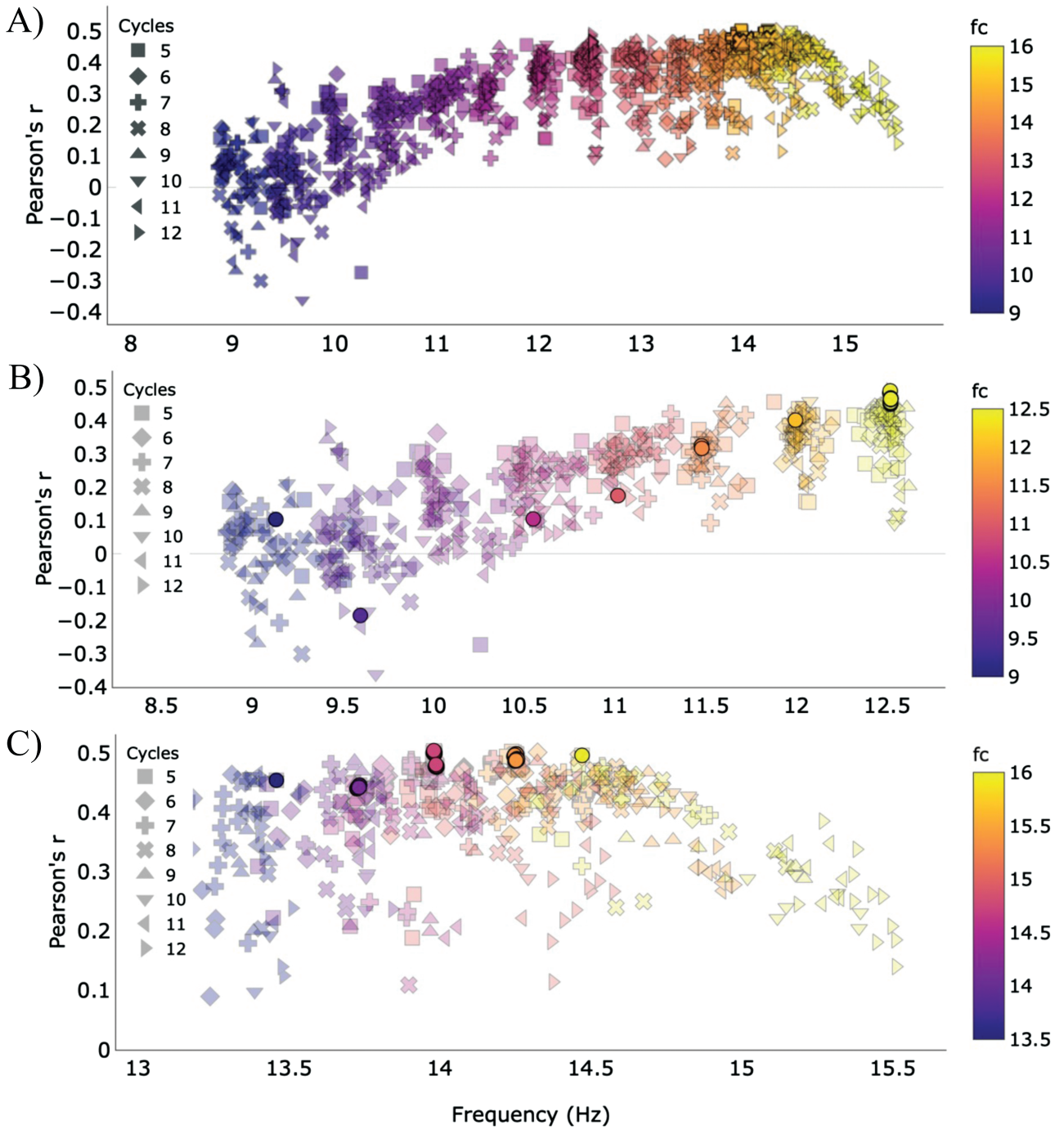

Figure 5. Model performance, defined by Pearson's *r*, for the different parameter combinations by central frequency (fc) and cycles. (A) Scatter plot of all parameter combinations with single fc value (N = 1495) shows that predicting cognition improves with higher central frequencies. (B) Scatter plot for slow spindle models (N = 610) shows that model performance peaks for slow spindles when the cycles parameter is set to 12 and central frequency (fc) is 12.5 Hz. Threshold > 4.5 and minimum spindle duration<0.5 s. (C). Scatter plot for fast spindle models (N = 818) shows good performance over 14 Hz that peaks when central frequency (fc) is 15–15.5 Hz and the cycles parameter is set to 5. Opaque circles represent threshold > 4.5 and minimum spindle duration < 0.5 s and cycles of 12 for (B) and 5 for (C).

coefficients): total (0.15 [–0.16, 0.27]), fluid (0.23 [–0.09, 0.34]), crystallized (–0.06 [–0.37, 0.16]), PVT (–0.095 [–0.38, 0.14]), Flanker ICA (0.05 [–0.23, 0.20]), LSWM (0.10 [–0.19, 0.23]), DCCS (0.03 [–0.28, 0.20]), PCPS (0.27 [–0.06, 0.38]), PSM (0.22 [–0.10, 0.34]), and ORR (–0.08 [–0.35, 0.13]). Using a modified percentile bootstrap method to compare Pearson's correlations across all models, we found no difference between the best performing and default parameter models: total (–0.12 [–0.43, 0.23]), fluid (–0.10 [–0.41, 0.24]), crystallized (–0.26 [–0.54, 0.19]), PVT (–0.28 [–0.59, 0.14]), Flanker ICA (–0.13 [–0.44, 0.21]), LSWM (–0.15 [–0.43, 0.14]), DCCS (–0.19 [–0.50, 0.16]), PCPS (–0.10 [–0.43, 0.21]), PSM (–0.10 [–0.42, 0.23]), and ORR (–0.27 [–0.55, 0.15]).

Performance trends in patients with cognitive impairment

To evaluate our results in cognitively impaired patients, we compared the association between cognitive impairment status and spindle features generated from the best and worst performing

Table 3. Luna parameters, spindle features (mean +/− SD), and Pearson correlation values ($p < .001$) for the top linear multiple regression models predictive of fluid cognition for each central frequency (fc) value

							FFT (Hz)		Duration (s)		Density (spm)		Amplitude (uV)			
cycles	fc (Hz)	q	th	Min (s)	Max (s)	Merge (s)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Pearson's r	95% CI
10	9	0.8	3.5	0.3	2.4	1	9.19	0.22	0.83	0.12	0.06	0.03	32.70	29.91	0.21	[−0.07, 0.32]
12	9.5	−0.4	2	0.3	2.3	0.3	9.42	0.13	0.66	0.05	6.03	0.93	29.93	42.64	0.38	[0.05, 0.48]
6	10	0	2	0.3	2.8	0.6	10.02	0.33	0.66	0.08	2.34	0.45	32.29	45.97	0.36	[0.05, 0.45]
5	10.5	0.6	5	0.6	3	0.9	10.82	0.48	0.89	0.12	0.12	0.10	42.62	55.87	0.34	[0.05, 0.45]
7	11	1	3	0.7	3.9	0.4	11.31	0.42	0.96	0.12	0.08	0.07	32.76	8.56	0.42	[0.24, 0.51]
5	11.5	1	2	0.5	3.1	0.3	11.88	0.55	0.74	0.08	0.21	0.18	30.15	8.72	0.46	[0.25, 0.56]
10	12	1	4.5	0.3	3.1	0.8	12.07	0.31	0.79	0.10	0.25	0.18	30.81	8.27	0.46	[0.27, 0.56]
12	12.5	0	5	0.2	2.6	0.5	12.52	0.21	0.87	0.09	1.77	0.66	34.04	46.15	0.49	[0.30, 0.55]
9	13	−0.4	6	0.2	3	0.3	12.95	0.22	0.84	0.11	1.33	0.61	36.60	44.88	0.47	[0.16, 0.56]
6	13.5	0.8	2.5	0.2	2.4	0.4	13.31	0.35	0.68	0.10	0.88	0.48	33.47	68.80	0.46	[0.11, 0.54]
9	14	−0.2	6	0.4	3.9	0.2	13.72	0.27	0.82	0.10	1.30	0.64	34.80	47.25	0.49	[0.16, 0.57]
7	14.5	0	5	0.2	3.9	0.7	13.91	0.34	0.76	0.10	1.25	0.71	33.36	53.52	0.49	[0.17, 0.58]
5	15	0	5	0.4	3.9	0.7	13.98	0.42	0.74	0.10	0.83	0.61	32.30	52.45	0.50	[0.20,0.59]
6	15.5	0.3	5	0.4	2.6	0.7	14.32	0.43	0.74	0.09	0.60	0.49	30.79	54.78	0.50	[0.18, 0.60]
5	16	0.4	5	0.2	2.7	0.5	14.47	0.44	0.71	0.09	0.53	0.46	27.42	57.20	0.50	[0.18, 0.58]

fc, central frequency; th, threshold; q, quality metric.

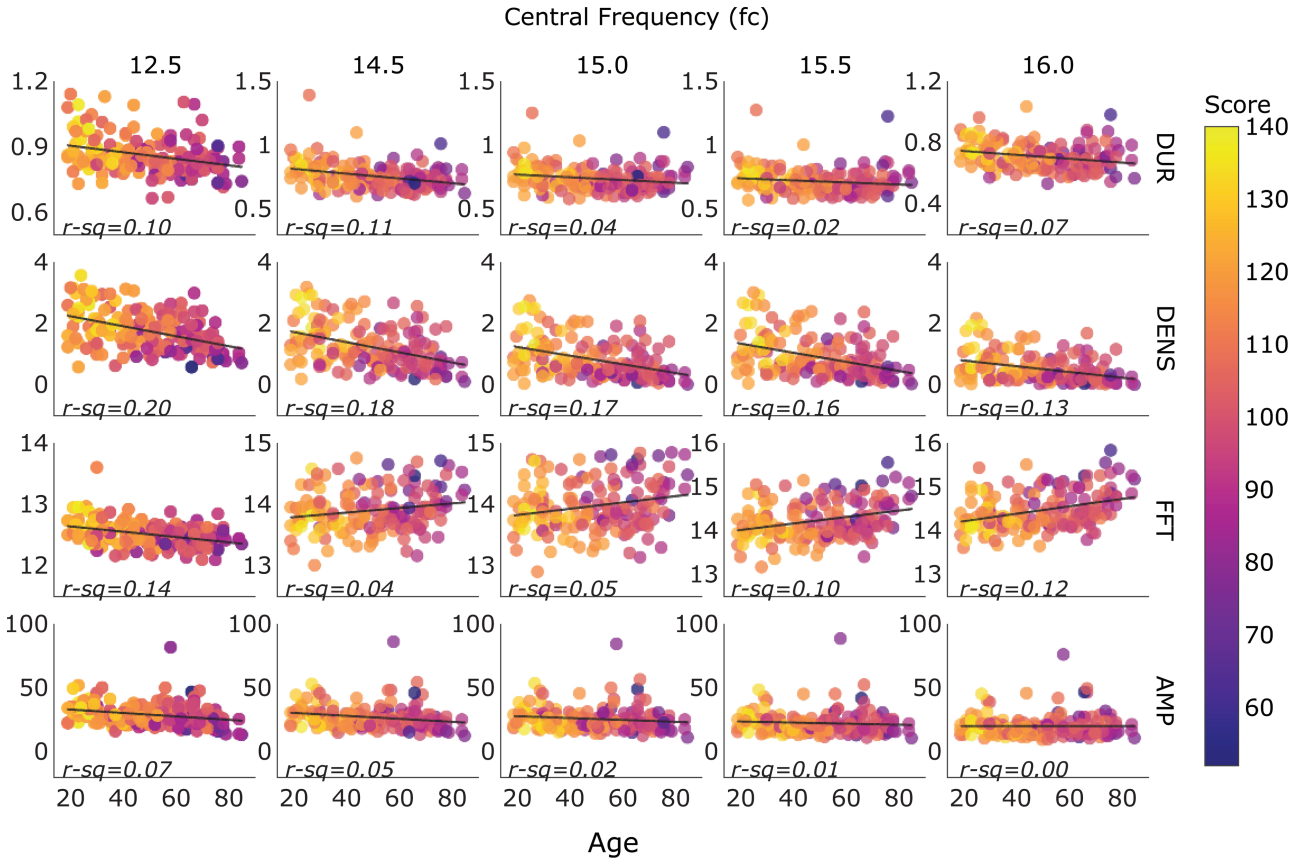

Figure 6. Spindle feature by age for the top 5 regression models by central frequency (fc). In older patients, fast spindles had a slightly higher frequency relative to younger patients, despite the decreased spindle density, and poor performance on fluid cognitive tests was linked with higher fast spindle FFT averages. The color bar represents absolute scores for fluid cognition from the NIH Toolbox Cognition Battery. Each data point represents one patient. AMP, Amplitude; FFT, Frequency; DENS, Density; DUR, Duration.

models across our original, non-dementia dataset and a dataset of 215 dementia and 308 MCI patients. Distributions of spindle features are shown in Figure 7. A Cuzick test for trend showed significant trends for all spindle features when the best performing parameter combination was selected (density: $p < 2.2e\text{-}16$; duration: $p = .03$; frequency: $p = 8.5e\text{-}11$; amplitude: $p = 3.8e\text{-}08$). For the worst performing combination, significant trends were found for spindle density and amplitude (density: $p = 6.7e\text{-}06$; duration: $p = 0.30$; frequency: $p = 0.54$; amplitude: $p = 5.9e\text{-}07$), although density values were close to 0 for all groups (Figure 7). Performance

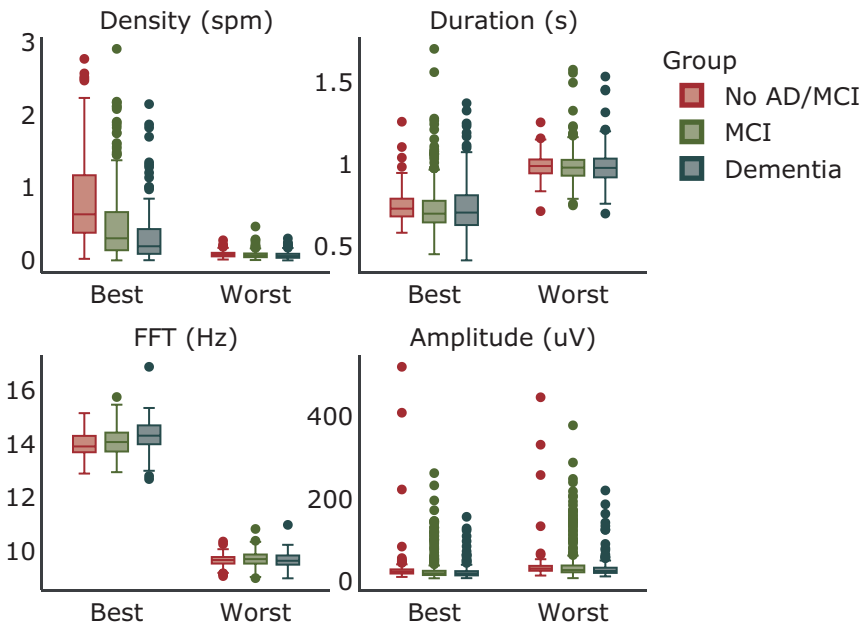

Figure 7. Distribution of spindle features by cognitive status for best and worst parameter combinations. Box plot of spindle density, duration, mean frequency (FFT), and amplitude are shown for each group and combination. The box signifies the upper and lower quartiles. The median is represented by a short line within the box for each group. Outliers are depicted by dots. AD, Alzheimer's Disease; MCI, Mild Cognitive Impairment.

for the 1000 randomly selected parameter combinations from the preliminary stage, along with the best, default and worst combination parameters are listed in Supplementary Table S4.

External validation

External validation of our results was performed using 476 participant data from SHHS/FHS. When the best performing parameter combination was selected, spindle features showed a significantly better ability to predict the WMS score compared to the worst performing combination (best: 0.27 [0.09, 0.38]; worst: –0.05 [–0.33, 0.13]) and similar performance with the default settings (0.23 [0.01, 0.35]). Performance for the 1000 randomly selected parameter combinations from the preliminary stage, along with the best, default and worst combination parameters is shown in in Supplementary Figure S7.

Discussion

Our key finding is that the strength of associations between cognition and sleep spindle features depends on the spindle detection parameters used and type of cognition being measured. Specifically, spindle features variably predict fluid cognition depending on spindle detection parameter values and poorly predict crystallized cognition regardless of the parameters chosen. The total cognition model resulted in a weaker association, which became nonsignificant when age-adjusted scores were used. Model performance for fluid composite and subtests was further influenced by spindle type (fast vs. slow). Fast spindle features generally showed better performance for predicting fluid cognition. Of the five fluid cognition subtests, only working memory showed preference for slow spindle features, with 59% of its significant models originating from slow spindle features. When age-adjusted scores were used, the performance gap between fast and slow spindle features widened, as 78% of fluid models were characterized by fast spindle features compared to 60% when absolute scores were used. Finally, interactions between parameters were noted for slow and fast spindle types. Overall, our findings provide evidence that parameter settings for automated spindle detection influence how well fluid cognition can be predicted and highlight the importance of considering parameter interactions when using automated spindle detectors.

Performance trends for best and default detector settings

When absolute scores were predicted, we found no significant difference between the best (0.503) and default (r = 0.427) parameter combinations at our sample size. Models predicting age-adjusted scores showed somewhat lower performance for all cognitive tests when default parameter settings were selected, and thus the variation in cognition explained by spindle features after accounting for age was smaller when default parameters are used, although the difference found was not statistically significant (modified bootstrap method) at our sample size. The purpose of predicting absolute versus age-adjusted scores are different, where the former assesses the association between sleep spindles and cognition, while the latter attempts to assess the independent contribution of spindle characteristics to cognition. Overall, we find that changing parameter settings can lead to variability in observed associations, with both higher and lower performance, although default parameters are generally representative of the overall associations.

Crystallized versus fluid cognition

When using optimized parameters to maximize correlation coefficients, the automated spindle detector extracted spindle features that moderately predicted (r = 0.503) fluid intelligence. Fluid intelligence relies on the capacity to use logic or abstract thinking to identify patterns and solve problems independently from accumulated knowledge. In

contrast, spindle features failed to accurately predict crystallized intelligence, which measures acquired knowledge and skills through experience and education. Although the best crystallized cognition model showed a moderate negative correlation between predicted and true scores, the association was not significant. To note, fluid and crystallized intelligence are strongly correlated and reflect higher-order general cognitive ability. Although these two constructs are interrelated, they are commonly evaluated and reported separately for research and treatment purposes [46] and are thought to be differentially influenced by different aspects of brain health [47].

Our findings of fluid model performance agree with previous reports that show significant correlations between spindle features and fluid cognition [48, 49]. Using multimodal imaging (EEG and functional magnetic resonance imaging [fMRI]), one study observed associations between fluid cognition and specific brain region activations that were time-locked to spindle events [48]. Spindle amplitude was found to significantly correlate with three fluid cognition measures: deductive reasoning (r = 0.516), spatial planning (r = 0.444), and polygons (r = 0.445) subtests. Evidence for an association with spindle amplitude and general cognitive ability was also found in a meta-analysis [50]. Although one large study found no association between spindle features and general cognitive ability in a birth cohort of adolescents [51], this difference is attributable to small sample size, as meta-analysis found a correlation value of r ~ 0.2 using the same birth cohort [50]. For example, one study found an association between spectral power, maximal amplitude, duration, and frequency and cognitive impairment and specifically noted that spindle amplitude was associated with general cognitive ability, while spindle frequency was inversely related to executive function in patients with subjective cognitive complaints and MCI [52]. Because fluid cognition declines at earlier stages of Alzheimer's Disease (AD), and crystallized cognition is hypothesized to compensate for age-related declines in fluid cognition by increasing with age [53], a large crystallized > fluid cognition discrepancy may serve as sensitive marker for early AD pathological changes. These observations point to distinct neural circuitries underlying these two intelligence domains. Given the apparent greater fragility of fluid cognitive networks and its relation to early AD vulnerability, it supports the idea that spindles are a biomarker of fluid cognition and are less suited to measure crystallized knowledge.

Fluid cognition and spindle frequency

Of the fluid cognition models with equivalent performance, 360 (60%) were characterized by fast spindle activity (FFT average>13Hz). Similarly, models for each fluid subtest generally favored fast spindle activity. This observation was most clear for the DCCS (measures cognitive flexibility), where 76% of the best performing models originated from fast spindle features. Additionally, models for the Flanker ICA, PCPS, and PSM (measures visual episodic memory) tests consisted of 55%, 57%, and 68% fast spindle models, respectively. The LSWM test was the only subtest that showed preference for both slow models, with 59% of its top models corresponding to slow spindle activity. Notably, when age-adjusted scores were used for analysis, 100% of the best performing models originated from fast spindle features for the DCCS test. Similarly, 98% of PCPS models and 92% of PSM models originated from fast spindle features. In contrast the Flanker ICA and LSWM age-adjusted scores were poorly predicted by all models, although the latter showed a trend for significance when slow spindles were used.

Our finding that most fluid composite and subtest models showed better performance with fast spindle activity aligns with extensive literature supporting the selective association of fast spindles with cognitive functions characteristic of fluid intelligence [54-57]. For instance, one fMRI study assessed memory consolidation and found that fast sleep spindles selectively exhibited strong interaction with functional connectivity between the hippocampus and areas of the neocortex [55]. Other studies have linked episodic memory [56, 57], verbal learning [58], motor [59] and declarative [57] memory consolidation, and visuospatial reasoning ability [60] specifically to fast spindle activity.

Our findings are also in keeping with evidence that fast spindles are significantly reduced in AD [10] and MCI [10, 12] relative to healthy controls. The preferential role of fast spindles has also been shown in aging-related sleep microstructural changes [9] and with cognitive decline in AD and MCI [11]. In one study, the largest age-related sleep changes were decreases in delta power in N3, K-complex density, and fast spindle density. Notably, the effect size of age on fast spindle density was the largest across features, while no effects of age were seen on slow spindle density [9].

Although working memory was the only fluid subdomain that was predicted better by slow spindle features, this inconsistency might reflect distinct neural circuitries underlying the different spindle types. Although fast and slow spindles both overlap in their thalamic origin, fast spindle activity exclusively involves memory-related cortical regions, such as the hippocampus, while slow spindle activity involves the superior frontal gyrus [61], an area that has shown significant contribution to cognitive tasks like working memory [62]. Taken together, it appears that fast spindles influence all fluid intelligence-related tasks while slow spindles influence a limited number of these tasks.

Parameter interactions

Using a dimensionality reduction method, t-SNE [36, 37], we observed that parameter combinations alone were sufficient to visually separate models that resulted in small correlation coefficients from those that performed better and resulted in larger correlation coefficients. To assess the physiological basis for this clustering pattern we then confirmed with t-SNE that the raw spindle features were also capable of distinguishing weak vs. excellent performance.

Once parameter interactions were globally confirmed using t-SNE, we inspected these relationships using parallel coordinates [38]. Across both fast and slow spindle types, performance was optimized (correlation coefficients were maximized) when the threshold was greater than 4 and the minimum spindle duration was lower than 0.5 s. Slow spindles, however, showed consistently superior predictions for fluid intelligence when the central frequency was set to 12.5 Hz and cycles to 12, while fast spindles showed superior model performance when the central frequency was 15–15.5Hz and cycles was 5. Generally, the higher the central frequency, the worse it performed at higher cycles. Because the cycles parameter favors temporal resolution at lower values and frequency resolution at higher values, this observation may be attributed to shorter durations of faster spindles (Supplementary Figure S6). This distinction can lead

to confusion, when comparing fast spindle model performance. Fast spindles detected with a central frequency of 15–15.5Hz and cycles of 5 had an average measured frequency (FFT) of 14 Hz (pink and orange markers on Figure 5C). Ultimately, these settings detected spindles that had an average frequency of 14 Hz. When the central frequency was set to 14 Hz and the cycles was set to 12 (favoring frequency resolution) or 14.5 Hz and a lower cycles value, the detected spindles also had an average frequency of approximately 14 Hz yet showed inferior model performance (blue and purple markers on Figure 5C). Thus, even though these settings all detected spindles with a frequency of 14 Hz, model performance differed. This discrepancy may suggest a subset of spindles with a 14 Hz frequency with shorter durations that are important for fluid intelligence or could reflect a limitation of the spindle detector algorithm.

External validation

Using an internal validation dataset of patients with MCI and dementia, we found significant and strong trends between cognitive impairment status and each spindle feature when the best parameter combination was selected. Similarly, the best parameter combination showed superior performance in the external validation SHHS/FHS dataset, although overall performance was reduced compared to our MGH dataset. This reduction can be attributed to the different neuropsychological batteries used across the two datasets and the larger gap between cognitive testing and PSG date for SHHS/FHS (≤1095 days) compared to MGH (≤40 days).

Limitations

Our analysis was limited to central EEG derivations, which may be important given development of wearable EEG acquisition systems which focus on frontal or even occipital derivations. Because slow spindles are mostly attributed to frontal regions, model performance may have favored fast spindle features. However, one previous study found that spindle density at central derivations were generally highly predictive of global spindle density for both fast and slow types, suggesting that central derivations should efficiently capture within-channel variance in slow spindles at other locations [29]. Our analysis was also restricted to a single spindle detector; thus, findings do not directly generalize directly to other spindle detection methods. A recent paper compared the performance of seven spindle detection algorithms, although evaluation was restricted to default parameter settings [20].

Other limitations of this study include sample bias, as the patients were recruited from a single center, low racial/ethnic (76% White) and socioeconomic diversity, and variation in the number of days between PSG visit and cognitive assessment. Finally, our optimal parameters are specific for the NIH Toolbox cognition battery and may not translate to other measures of cognition.

Conclusion

In summary, when using automated spindle detectors with the intention of examining brain health, parameter settings and interactions should be taken into consideration as they moderate the detected spindle features and their physiological implications. By setting cognitive measures as the gold standard when tuning spindle detection parameters, spindle features showed significant association with the ability to predict absolute fluid cognition scores (r = 0.503). Our analysis further identifies two parameters of space that show weak vs moderate predictive performance. Fast spindle features also showed better performance relative to slow spindle features, although future studies are needed to evaluate the functional difference between these spindle types.

Supplementary Material

Supplementary material is available at *SLEEP* online.

Disclosure Statement

Financial Disclosure: This research was conducted while MBW was a Breakthroughs in Gerontology Grant recipient, supported by the Glenn Foundation for Medical Research and the American Federation for Aging Research. MBW also received funding from grants from the National Institutes of Health (NIH-NINDS 1K23NS090900, 1R01NS102190-01, 1R01NS102574, 1R01NS107291, 1RF1AG064312) and the American Academy of Sleep Medicine (AASM) through an AASM Foundation Strategic Research Award. HS received funding from the Glenn Foundation for Medical Research and the American Federation for Aging Research. RJT received support from an American Association of Sleep Medicine Foundation Strategic Research Category-I award.

Nonfinancial Disclosure: MBW is co-founder of Beacon Biosignals. RJT discloses: (1) Licensed patent for the ECG-spectrogram to MyCardio, LLC; (2) Licensed patent to DeVilbiss-Drive for auto-CPAP algorithm; (3) unlicensed patent for a device to treat central and complex sleep apnea using low concentration CO_2 adjunct to positive airway pressure; (4) consultant to Jazz Pharmaceuticals, GLG Councils, and Guidepoint Global. All other authors have no conflicts of interest to declare.

Authors' Contributions

NA performed analysis and drafted and reviewed the manuscript. HS was part of the data analysis, critical review, and revision of the manuscript. EMY was part of data management, critical review, and revision of the manuscript. RAT, MW, MDSC were part of data acquisition and critical review and revision of the manuscript. WG, LWD, EK, AO, JS, JR, SSC were part of critical review and revision of the manuscript. RJT contributed to design of the work, supervision (mentorship), critical review, and revision of the manuscript. MBW contributed to design of the work, supervision (oversight and leadership), critical review, and revision of the manuscript.

References

1. Wang Y, *et al*. What is brain health and why is it important? *BMJ*. 2020;**371**:m3683. doi:10.1136/bmj.m3683.
2. Lubetkin EI, *et al*. Burden of disease due to sleep duration and sleep problems in the elderly. *Sleep Health* 2018;**4**:182–187. doi:10.1016/j.sleh.2017.11.007.

3. Todd S, *et al*. Survival in dementia and predictors of mortality: a review. *Int J Geriatr Psychiatry*. 2013;**28**:1109–1124. doi:10.1002/gps.3946.
4. Ohayon MM, *et al*. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep*. 2004;**27**:1255–1273. doi:10.1093/sleep/27.7.1255.
5. Moraes W, *et al*. Effects of aging on sleep structure throughout adulthood: a population-based study. *Sleep Med*. 2014;**15**:401–409. doi:10.1016/j.sleep.2013.11.791.
6. Wennberg AMV, *et al*. Sleep disturbance, cognitive decline, and dementia: a review. *Semin Neurol*. 2017;**37**:395–406. doi:10.1055/s-0037-1604351.
7. Blackwell T, *et al*. Associations of objectively and subjectively measured sleep quality with subsequent cognitive decline in older community-dwelling men: the MrOS sleep study. *Sleep*. 2014;**37**:655–663. doi:10.5665/sleep.3562.
8. Crowley K, *et al*. The effects of normal aging on sleep spindle and K-complex production. *Clin Neurophysiol*. 2002;**113**:1615–1622. doi:10.1016/s1388-2457(02)00237-7.
9. Schwarz JFA, *et al*. Age affects sleep microstructure more than sleep macrostructure. *J Sleep Res*. 2017;**26**:277–287. doi:10.1111/jsr.12478.
10. Rauchs G, *et al*. Is there a link between sleep changes and memory in Alzheimer's disease? *Neuroreport*. 2008;**19**:1159–1162. doi:10.1097/wnr.0b013e32830867c4.
11. Gorgoni M, *et al*. Parietal fast sleep spindle density decrease in alzheimer's disease and amnesic mild cognitive impairment. *Neural Plast*. 2016;**2016**:8376108. doi:10.1155/2016/8376108.
12. Westerberg CE, *et al*. Concurrent impairments in sleep and memory in amnestic mild cognitive impairment. *J Int Neuropsychol Soc*. 2012;**18**:490–500. doi:10.1017/s135561771200001x.
13. De Gennaro L, *et al*. An electroencephalographic fingerprint of human sleep. *Neuroimage*. 2005;**26**:114–122. doi:10.1016/j.neuroimage.2005.01.020.
14. Goldschmied JR, *et al*. Spindles are highly heritable as identified by different spindle detectors. *Sleep* 2020. doi:10.1093/sleep/zsaa230.
15. Silverstein LD, *et al*. The stability of the sigma sleep spindle. *Electroencephalogr Clin Neurophysiol*. 1976;**40**:666–670. doi:10.1016/0013-4694(76)90142-5.
16. Holz J, *et al*. EEG Sigma and slow-wave activity during NREM sleep correlate with overnight declarative and procedural memory consolidation. *J Sleep Res*. 2012;**21**:612–619. doi:10.1111/j.1365-2869.2012.01017.x.
17. Ayoub A, *et al*. Differential effects on fast and slow spindle activity, and the sleep slow oscillation in humans with carbamazepine and flunarizine to antagonize voltage-dependent Na+ and Ca2+ channel activity. *Sleep*. 2013;**36**:905–911. doi:10.5665/sleep.2722.
18. Laventure S, *et al*. Beyond spindles: interactions between sleep spindles and boundary frequencies during cued reactivation of motor memory representations. *Sleep*. 2018:41. doi:10.1093/sleep/zsy142.
19. Molle M, *et al*. Fast and slow spindles during the sleep slow oscillation: disparate coalescence and engagement in memory processing. *Sleep*. 2011;**34**:1411–1421. doi:10.5665/sleep.1290.
20. Warby SC, *et al*. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nat Methods*. 2014;**11**:385–392. doi:10.1038/nmeth.2855.
21. Dimitrov T, *et al*. Sleep spindles comprise a subset of a broader class of electroencephalogram events. *Sleep*. 2021. doi:10.1093/sleep/zsab099.
22. Alfonsi V, *et al*. Spatiotemporal dynamics of sleep spindle sources across NREM sleep cycles. *Front Neurosci*. 2019;**13**:727. doi: 10.3389/fnins.2019.00727.
23. Lacourse K, *et al*. Massive online data annotation, crowdsourcing to generate high quality sleep spindle annotations from EEG data. *Sci Data*. 2020;**7**:190. doi:10.1038/s41597-020-0533-4.
24. Feinberg I, *et al*. Effects of hypnotics on the sleep EEG of healthy young adults: new data and psychopharmacologic implications. *J Psychiatr Res*. 2000;**34**:423–438. doi:10.1016/s0022-3956(00)00038-8.
25. Weintraub S, *et al*. Cognition assessment using the NIH Toolbox. *Neurology*. 2013;**80**:S54–S64. doi:10.1212/wnl.0b013e3182872ded.
26. Weintraub S, *et al*. The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: validation in an adult sample. *J Int Neuropsychol Soc*. 2014;**20**:567–578. doi:10.1017/s1355617714000320.
27. Heaton RK, *et al*. Reliability and validity of composite scores from the NIH toolbox cognition battery in adults. *J Int Neuropsychol Soc*. 2014;**20**:588–598. doi:10.1017/s1355617714000241.
28. Berry RB, *et al*. *The AASM Manual for the Scoring of Sleep and Associated Events*. Darien, IL: American Academy of Sleep Medicine; 2015.
29. Purcell SM, *et al*. Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource. *Nat Commun*. 2017;**8**:15930. doi:10.1038/ncomms15930.
30. Pan J, *et al*. A real-time QRS detection algorithm. *IEEE Trans Biomed Eng*. 1985;**32**:230–236. doi:10.1109/TBME.1985.325532.
31. Buckelmuller J, *et al*. Trait-like individual differences in the human sleep electroencephalogram. *Neuroscience*. 2006;**138**:351–356. doi:10.1016/j.neuroscience.2005.11.005.
32. Hjorth B. EEG analysis based on time domain properties. *Electroencephalogr Clin Neurophysiol*. 1970;**29**:306–310. doi:10.1016/0013-4694(70)90143-4.
33. Wamsley EJ, *et al*. Reduced sleep spindles and spindle coherence in schizophrenia: mechanisms of impaired memory consolidation? *Biol Psychiatry*. 2012;**71**:154–161. doi:10.1016/j.biopsych.2011.08.008.
34. Bergstra J, *et al*. Random search for hyper-parameter optimization. *J Mach Learn Res*. 2012;**13**:281–305.
35. Wilcox RR. Comparing Pearson correlations: dealing with heteroscedasticity and nonnormality. *Commun Stat - Simul Comput*. 2009;**38**:2220–2234. doi:10.1080/03610910903289151.
36. van der Maaten L, *et al*. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;**9**:2579–2605.
37. van der Maaten L. Accelerating t-SNE using tree-based algorithms. *J Mach Learn Res*. 2014;**15**:3221–3245.
38. Johansson J, *et al*. Evaluation of parallel coordinates: overview, categorization and guidelines for future research. *IEEE Trans Vis Comput Graph*. 2016;**22**:579–588. doi:10.1109/TVCG.2015.2466992.
39. Cuzick J. A Wilcoxon-type test for trend. *Stat Med*. 1985;**4**:87–90. doi:10.1002/sim.4780040112.
40. Redline S, *et al*. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. *Sleep*. 1998;**21**:759–767.
41. Quan SF, *et al*. The sleep heart health study: design, rationale, and methods. *Sleep*. 1997;**20**:1077–1085.
42. Zhang GQ, *et al*. The national sleep research resource: towards a sleep data commons. *J Am Med Inform Assoc*. 2018;**25**:1351–1358. doi:10.1093/jamia/ocy064.

43. Dean DA, *et al*. Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep*. 2016;**39**:1151-1164. doi: 10.5665/sleep.5774
44. Kannel WB, *et al*. An investigation of coronary heart disease in families. The Framingham offspring study. *Am J Epidemiol*. 1979;**110**:281–290. doi:10.1093/oxfordjournals.aje.a112813.
45. Farmer ME, *et al*. Neuropsychological test performance in Framingham: a descriptive study. *Psychol Rep*. 1987;**60**:1023–1040. doi:10.1177/0033294187060003-201.1.
46. Jaeggi SM, *et al*. Improving fluid intelligence with training on working memory. *Proc Natl Acad Sci U S A*. 2008;**105**:6829–33. doi: 10.1073/pnas.0801268105.
47. Akshoomoff N, *et al*. VIII. NIH toolbox cognition battery (CB): composite scores of crystallized, fluid, and overall cognition. *Monogr Soc Res Child Dev*. 2013;**78**:119–132. doi:10.1111/mono.12038.
48. Fang Z, *et al*. Brain activation time-locked to sleep spindles associated with human cognitive abilities. *Front Neurosci*. 2019;**13**:46. doi:10.3389/fnins.2019.00046.
49. van den Berg NH, *et al*. Sleep stages and neural oscillations: a window into sleep's role in memory consolidation and cognitive abilities. *Handb Behav Neurosci*. 2019;**30**:455–470.
50. Ujma PP. Sleep spindles and general cognitive ability – A meta-analysis. *Sleep Spindl Cortical Up States*. 2018:1–17.
51. Pesonen A-K, *et al*. The associations between spindle characteristics and cognitive ability in a large adolescent birth cohort. *Intelligence*. 2019;**72**:13-9.
52. Taillard J, *et al*. Non-REM sleep characteristics predict early cognitive impairment in an aging population. *Front Neurol*. 2019;**10**:197. doi:10.3389/fneur.2019.00197.
53. Li Y, *et al*. Complementary cognitive capabilities, economic decision making, and aging. *Psychol Aging*. 2013;**28**:595–613. doi:10.1037/a0034172.
54. Ferini-Strambi L, *et al*. Sleep microstructure and memory function. *Front Neurol*. 2013;**4**:159. doi:10.3389/fneur.2013.00159.
55. Andrade KC, *et al*. Sleep spindles and hippocampal functional connectivity in human NREM sleep. *J Neurosci*. 2011;**31**:10331–10339. doi:10.1523/JNEUROSCI.5660-10.2011.
56. Saletin JM, *et al*. The role of sleep in directed forgetting and remembering of human memories. *Cereb Cortex*. 2011;**21**:2534–2541. doi:10.1093/cercor/bhr034.
57. van der Helm E, *et al*. Sleep-dependent facilitation of episodic memory details. *PLoS One*. 2011;**6**:e27421. doi:10.1371/journal.pone.0027421.
58. Lafortune M, *et al*. Sleep spindles and rapid eye movement sleep as predictors of next morning cognitive performance in healthy middle-aged and older participants. *J Sleep Res*. 2014;**23**:159–167. doi:10.1111/jsr.12108.
59. Barakat M, *et al*. Fast and slow spindle involvement in the consolidation of a new motor sequence. *Behav Brain Res*. 2011;**217**:117–121. doi:10.1016/j.bbr.2010.10.019.
60. Fogel SM, *et al*. The function of the sleep spindle: a physiological index of intelligence and a mechanism for sleep-dependent memory consolidation. *Neurosci Biobehav Rev*. 2011;**35**:1154–1165. doi:10.1016/j.neubiorev.2010.12.003.
61. Schabus M, *et al*. Hemodynamic cerebral correlates of sleep spindles during human non-rapid eye movement sleep. *Proc Natl Acad Sci USA*. 2007;**104**:13164–13169. doi:10.1073/pnas.0703084104.
62. du Boisgueheneuc F, *et al*. Functions of the left superior frontal gyrus in humans: a lesion study. *Brain*. 2006;**129**:3315–3328. doi:10.1093/brain/awl244.