

Prediction of cognitive performance by demographics, sleep, and brain morphometry: machine learning findings from ENIGMA-Sleep Working Group

Masoud Tahmasian

`m.tahmasian@fz-juelich.de`

Research Center Juelich <https://orcid.org/0000-0003-3999-3807>

Hanwen Bi

Institute of Neurosciences and Medicine (INM-7), Research Centre Jülich

Torbjörn Åkerstedt

Department of Clinical Neuroscience, Karolinska Institutet

Robin Bülow

University Medicine Greifswald <https://orcid.org/0000-0003-1884-5784>

Michele Deantoni

GIGA-CRC-Human Imaging, University of Liège

Alexander Drzezga

University Hospital Cologne <https://orcid.org/0000-0001-6018-716X>

Jeremy Elman

Department of Psychiatry, University of California San Diego

David Elmenhorst

Institute for Neuroscience and Medicine, Molecular Organization of the Brain (INM-2), Research Center Jülich

Eva-Maria Elmenhorst

Department of Sleep and Human Factors Research, German Aerospace Center

Ralf Ewert

Department of Internal Medicine B-Cardiology, Pneumology, Infectious Diseases, Intensive Care Medicine, University Medicine Greifswald

Christine Fennema-Notestine

Department of Psychiatry, University of California San Diego

Fabio Ferrarelli

University of Pittsburgh

Stefan Frenzel

Department of Psychiatry and Psychotherapy, University Medicine Greifswald

Charlotte Von Gall

Institute of Anatomy II, Medical Faculty, Heinrich Heine University

Sarah Genon

Research Centre Jülich <https://orcid.org/0000-0002-7087-7882>

Hans Grabe

University Medicine Greifswald

Phuong Thuy Nguyen Ho

Department of Radiology and Nuclear Medicine, Erasmus University Medical Centre

Sanne Hoepel

Department of Epidemiology, Erasmus MC University Medical Center Rotterdam

Felix Hoffstaedter

Research Centre Juelich <https://orcid.org/0000-0001-7163-3110>

Agustín Ibáñez

Trinity College Dublin <https://orcid.org/0000-0001-6758-5101>

Neda Jahanshad

Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of the University of Southern California

Ahmadreza Keihani

Department of Psychiatry, University of Pittsburgh, Pittsburgh

Vincent Küppers

Department of Nuclear Medicine, Faculty of Medicine and University Hospital Cologne

Wen Liu

Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich

Ahmad Mayeli

Department of Psychiatry, University of Pittsburgh

Nasrin Mortazavi

University of Liège

Julia Neitzel

Department of Radiology and Nuclear Medicine, Erasmus University Medical Centre

Gustav Nilsson

Department of Clinical Neuroscience, Karolinska Institutet

Matthew Panizzon

Department of Psychiatry, University of California San Diego

Julia Rupp

Department of Psychiatry, University of Pittsburgh

Amin Saberi

Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich

Christina Schmidt

GIGA-CRC-Human Imaging, University of Liège

Kai Spiegelhalder

University of Freiburg, Germany

Beate Stubbe

Department of Internal Medicine B-Cardiology, Pneumology, Infectious Diseases, Intensive Care Medicine, University Medicine Greifswald

Sandra Tamm

Department of Clinical Neuroscience, Karolinska Institutet

Sophia Thomopoulos

University of Southern California <https://orcid.org/0000-0002-0046-4070>

Paul Thompson

Imaging Genetics Center, Mark and Mary Stevens Neuroimaging & Informatics Institute, Keck School of Medicine of the University of Southern California, Marina del Rey, California, USA

<https://orcid.org/0000-0002-4720-8867>

Sofie Valk

Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich

Gilles Vandewalle

University of Liège <https://orcid.org/0000-0003-2483-2752>

Tina Thi Vo-Eckerle

Department of Psychiatry, University of California San Diego

Henry Völzke

University Medicine Greifswald

Laura Waite

Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich

Joseph Wexler

Department of Psychology, Stanford University

Katharina Wittfeld

University Medicine Greifswald <https://orcid.org/0000-0003-4383-5043>

Kaustubh Patil

Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich

Antoine Weihs

University Medicine Greifswald <https://orcid.org/0000-0002-7113-0027>

Simon Eickhoff

Research Centre Jülich

Federico Raimondo

Heinrich Heine University Düsseldorf <https://orcid.org/0000-0003-4087-8259>

Analysis**Keywords:**

Posted Date: June 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9914920/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Prediction of cognitive performance by demographics, sleep, and brain morphometry: machine learning findings from ENIGMA-Sleep Working Group

Hanwen Bi^{1,2}, Torbjörn Åkerstedt^{3,4}, Robin Bülow⁵, Michele Deantoni⁶, Alexander Drzezga^{7,8,9}, Jeremy A. Elman^{10,11}, David Elmenhorst^{9,7}, Eva-Maria Elmenhorst^{12,13}, Ralf Ewert¹⁴, Christine Fennema-Notestine^{10,11}, Fabio Ferrarelli¹⁵, Stefan Frenzel¹⁶, Charlotte von Gall¹⁷, Sarah Genon^{1,2}, Hans J. Grabe^{16,18}, Phuong Thuy Nguyen Ho¹⁹, Sanne J.W. Hoepel²⁰, Felix Hoffstaedter^{1,2}, Agustin Ibanez^{21,22,23,24,25}, Neda Jahanshad²⁶, Ahmadreza Keihani¹⁵, Vincent Küppers^{7,1}, Wen Liu^{1,2}, Ahmad Mayeli¹⁵, Nasrin Mortazavi⁶, Julia Neitzel^{19,20,27}, Gustav Nilsson^{3,28}, Matthew S. Panizzon^{10,11}, Julia S. Rupp¹⁵, Amin Saberi^{1,29,2}, Christina Schmidt⁶, Kai Spiegelhalter³⁰, Beate Stubbe¹⁴, Sandra Tamm³, Sophia I. Thomopoulos²⁶, Paul M. Thompson²⁶, Sofie L. Valk^{1,29,2}, Gilles Vandewalle⁶, Tina Thi Vo-Eckerle^{10,11}, Henry Völzke^{31,32}, Laura K. Waite¹, Joseph Wexler^{33,3}, Katharina Wittfeld¹⁶, Kaustubh R. Patil^{1,2}, Antoine Weihs^{16,18}, Simon B. Eickhoff^{1,2}, Federico Raimondo^{1,2,#}, Masoud Tahmasian^{1,2,7*#} & ENIGMA-Sleep Working Group

¹ Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich, Jülich, Germany

² Institute of Systems Neuroscience, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf, Germany

³ Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

⁴ Department of Psychology, Stockholm University, Stockholm, Sweden

⁵ Institute of Diagnostic Radiology and Neuroradiology, University Medicine Greifswald, Greifswald, Germany

⁶ GIGA-CRC-Human Imaging, University of Liège, Liège, Belgium

⁷ Department of Nuclear Medicine, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

⁸ German Center for Neurodegenerative Diseases (DZNE), Bonn-Cologne, Germany

⁹ Institute for Neuroscience and Medicine, Molecular Organization of the Brain (INM-2), Research Center Jülich, Jülich, Germany

¹⁰ Center for Behavior Genetics of Aging, Department of Psychiatry, University of California San Diego

¹¹ Department of Psychiatry, University of California San Diego

¹² Department of Sleep and Human Factors Research, German Aerospace Center, Cologne, Germany

¹³ Institute for Occupational, Social and Environmental Medicine, Medical Faculty, RWTH Aachen University, Aachen, Germany

¹⁴ Department of Internal Medicine B-Cardiology, Pneumology, Infectious Diseases, Intensive Care Medicine, University Medicine Greifswald

¹⁵ Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

¹⁶ Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany

¹⁷ Institute of Anatomy II, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany

¹⁸ German Center for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, Greifswald, Germany

¹⁹ Department of Radiology and Nuclear Medicine, Erasmus University Medical Centre, Rotterdam, the Netherlands

²⁰ Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam, Netherlands

²¹ Latin American Brain Health Institute (BrainLat), Universidad Adolfo Ibáñez, Santiago, Chile

²² Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland

²³ Department of Biophysics, School of Medicine, Istanbul Medipol University, 34815 Istanbul, Türkiye

²⁴ Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, 08005, Barcelona, Spain

²⁵ Cognitive Neuroscience Center (CNC), Universidad de San Andrés, Buenos Aires, Argentina

²⁶ Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA

²⁷ Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts

²⁸ International Globally Distributed Organization for Research and Education (IGDORE), Stockholm, Sweden

²⁹ Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

³⁰ Department of Psychiatry and Psychotherapy, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Germany

58 ³¹ Institute for Community Medicine, SHIP/Clinical-Epidemiological Research, University Medicine
59 Greifswald, Greifswald, Germany
60 ³² German Centre for Cardiovascular Research (DZHK), Partner Site Greifswald, Greifswald, Germany
61 ³³ Department of Psychology, Stanford University, Stanford, CA, USA

62
63 # shared senior authors.

64
65
66 * **Correspondence:** Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research
67 Center Jülich, 52428 Jülich, Germany, Email: m.tahmasian@fz-juelich.de
68

69 **Abstract**

70 Group-level studies have highlighted the roles of aging, poor sleep, and brain atrophy in cognitive
71 performance (CP) but have overlooked inter-individual variability. We predict CP from feature sets
72 (demographic, subjective/objective sleep parameters, and regional brain morphometry) using multisite
73 ENIGMA-Sleep data (n = 2,372). Linear and non-linear machine learning models were trained on the
74 largest cohort (n = 845), and the best-performing models were validated on independent cohorts.
75 Subsequently, based on the best-performing model on the largest cohort, we characterized feature
76 importance and interactions across all cohorts. We observed that a combination of demographic, sleep,
77 and brain parameters moderately predicted CP, with age emerging as the key predictor. Model
78 explanations further suggested that age was the primary driver of prediction models, while sleep played a
79 smaller role that varied across subgroups. These findings endorsed inter-individual variability and complex
80 interaction between aging, sleep, brain, and CP.

81

82

INTRODUCTION

Cognitive performance (CP) reflects individual ability across core neurocognitive domains, including attention, processing speed, executive function, and memory. In population studies, lower CP is associated with poorer educational and occupational attainment, reduced everyday functioning, maladaptive interpersonal relationships, and adverse mental health outcomes across the lifespan (Lövdén et al., 2020). Thus, poor CP, including attention and memory impairments, can lead to substantial healthcare and societal costs (Gläser et al., 2025). Notably, substantial inter-individual variability in CP exists, influenced by biological, lifestyle, and environmental factors. Identifying which factors most strongly contribute to CP at the individual level is therefore critical for informing early intervention strategies and guiding personalized approaches to cognitive health.

Aging across the lifespan is associated with cognitive decline in multiple domains and is an early symptom for the development of prevalent neurodegenerative diseases such as dementia (Brito et al., 2023). Fluid cognitive abilities, including attention, novel problem-solving, processing speed, literacy, numeracy, and episodic memory, tend to decline gradually from midlife onward, particularly in women (Brito et al., 2023; Hanushek et al., 2025). Accordingly, aging poses a major challenge to maintaining CP in rapidly aging populations. Besides aging, sleep is related to multiple aspects of CP (Tahmasian et al., 2026). Previous large-scale studies in a large aging population dataset (UK Biobank, UKB) observed that both long and short sleep durations were linked to impaired CP (Kyle et al., 2017; Li et al., 2025), and multi-organ ageing clocks (O'Toole et al., 2026). Non-linear relationships between sleep duration and CP and executive function have also been reported (Li et al., 2022; Tai et al., 2022). Longitudinal studies have shown that abnormal sleep duration and shift work are predictors of future cognitive impairment (Ell et al., 2023). Although these studies reported statistically significant group-level associations, the corresponding effect sizes were consistently very small (r range 0.0007 - 0.0057) (Keyes & Westreich, 2019; Lin et al., 2013; Spence et al., 2026; Sullivan & Feinn, 2012). Those UKB studies also primarily relied on a single self-reported question about sleep duration from the British population, whereas sleep behaviors show substantial regional and cultural variation (Coutrot et al., 2022; Willoughby et al., 2023). Moreover, individual sleep needs are flexible across individuals and dynamically related to CP (Fjell & Walhovd, 2024). Together, these factors may limit the generalizability of group-level findings to international cohorts.

The neurobiological substrates linking CP, aging, and sleep remain poorly understood, making the characterization of sleep-related variation in brain structure essential for elucidating the mechanisms underlying their interplay. Prior work incorporating genetic information has reported non-linear relationships among sleep duration, cognition, and brain structure (Li et al., 2022), with divergent patterns between short- and long- sleep duration groups (Li et al., 2025). An inverse U-shaped cross-sectional association between sleep duration and cortical thickness was reported, but longitudinal analysis did not show robust associations (Fjell, Sørensen, Wang, Amlie, Baaré, Bartrés-Faz, Bertram, et al., 2023). Other evidence suggests that long sleep duration is linked with higher grey matter volume of the basal ganglia, whereas insomnia symptoms were not associated with morphological measurements (Schiel et al., 2023; Weihs et al., 2023). At the individual level, structural brain data have been used to predict CP and sleep health separately, suggesting that morphometric measures may capture variability relevant to CP, although their predictive power has generally been modest (Genon et al., 2022; Raimondo et al., 2026; Wulan et al., 2024). Taken together, these findings suggest inconsistencies across studies regarding the interplay between aging, sleep, brain, and CP. Large-scale international datasets and advanced computational approaches are therefore needed to model this complexity and clarify the cognitive consequences of poor sleep, aging, and brain atrophy.

Machine learning (ML) provides an individual-level framework that can accommodate heterogeneous, non-linear associations, and particularly in unseen data (independent cohorts), thereby improving generalizability beyond traditional group-level statistical analyses in one cohort (Varoquaux & Poldrack,

2019). In this preregistered study (<https://osf.io/nhbkq/>), we leveraged multisite international data of the ENIGMA-Sleep Working group (Tahmasian et al., 2021) with both self-reported and objective (polysomnography-based) sleep measures to evaluate whether demographics (i.e., age, sex, body mass index (BMI), and depressive symptom scores), sleep, and regional (parcel-based) brain morphometry can predict CP in the general population. To assess the contribution of single and multiple domains to predicting CP, we systematically compared ML models based on demographics-only, sleep with demographics, and a combination of demographics, sleep, and brain morphometry, and complemented these comparisons with shuffled-feature controls. To evaluate the generalizability of ML models beyond a single cohort and under geographic variation in sleep behaviors (Coutrot et al., 2022; Willoughby et al., 2023), we trained multiple ML models in the largest cohort and assessed the best-performing models on independent cohorts, consistent with recommendations that external validation provides a more rigorous assessment of generalizability in neuroimaging prediction studies (Bi et al., 2026; Olfati et al., 2024; Rosenblatt et al., 2026). We then applied explainable ML using SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) to identify feature contributions and also Shapley Interaction Quantification (SHAP-IQ) (Muschalik et al., 2024) to characterize feature interaction patterns underlying model predictions. To examine whether these prediction patterns were homogeneous across participants, we additionally clustered participant-specific SHAP values and explored the transferability of the resulting clustering rules across cohorts.

148

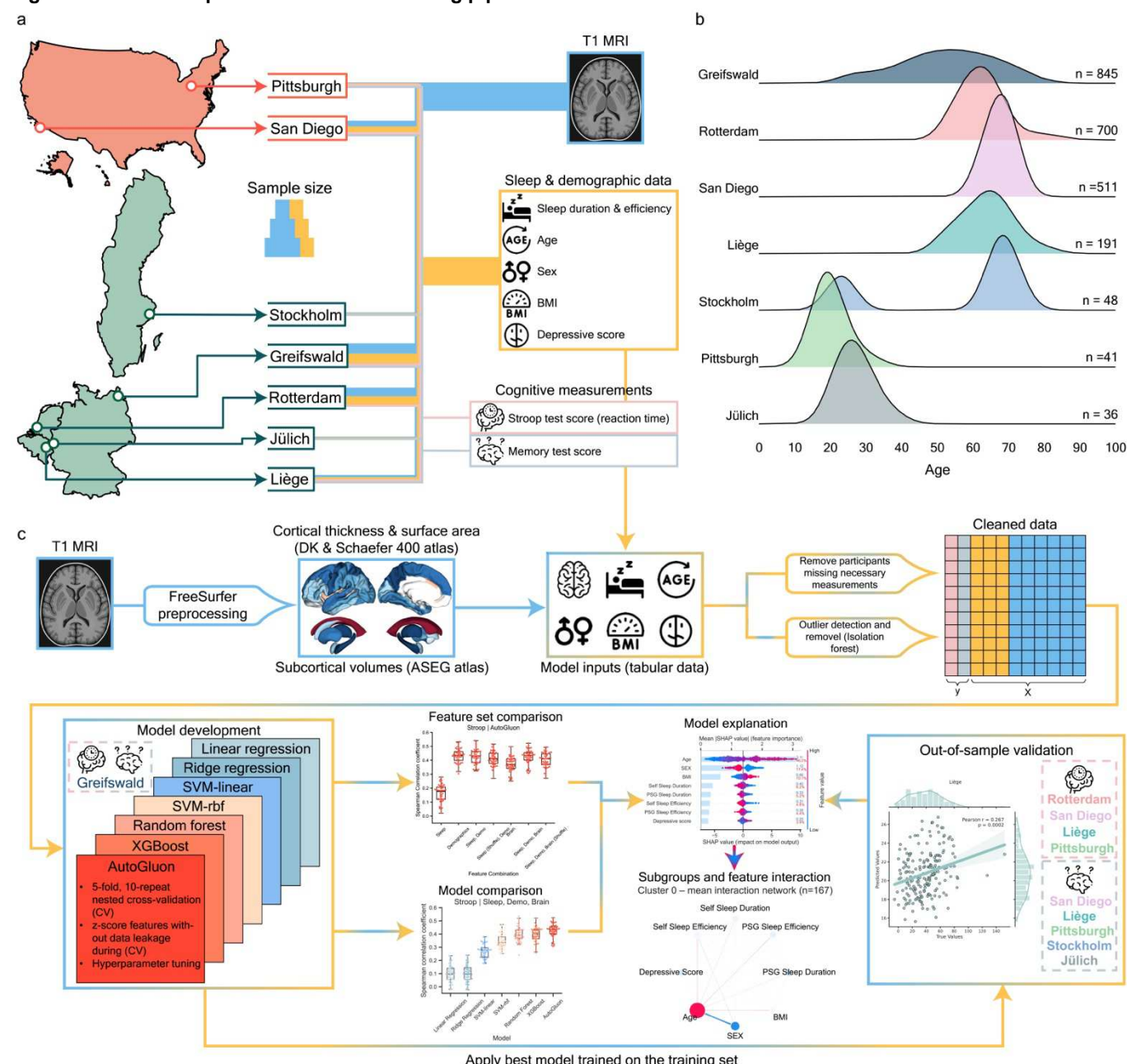
149 RESULTS

150 Overview of the framework and study cohort

151 We evaluated whether demographics, sleep, and regional brain morphometry features could predict CP
152 across seven international cohorts (n = 2,372 participants), including Greifswald (SHIP-Trend, n = 845)
153 (Völzke et al., 2022), Rotterdam (The Rotterdam Study, n = 700) (Ikram et al., 2024), San Diego (VETSA,
154 n = 511) (Kremen et al., 2013), Liège (COF and COGNAP, n = 191) (Deantoni et al., 2024), Stockholm
155 (Sleepy Brain, n = 48) (Nilsonne et al., 2020), Pittsburgh (PyNEL, n = 41) (LaGoy et al., 2022), and Jülich
156 (Somnosafe, n = 36) (Hennecke et al., 2021). The models used three feature domains, which were *i*)
157 demographics including age, sex, BMI, and depressive scores, *ii*) sleep duration and efficiency derived
158 from polysomnography (PSG) and self-reported questionnaires, and *iii*) parcel-based brain structural
159 measures from T1 MRI, including cortical thickness, surface area, and subcortical volume. The prediction
160 targets were cognitive measurements assessed by the Stroop (reaction time) and memory tests
161 (accuracy). To facilitate cross-cohort interpretation, we defined the Stroop test score uniformly as reaction
162 time and the memory test score as accuracy hereinafter. We trained the ML models in the Greifswald
163 cohort, which included 845 participants and provided the largest sample size, the broadest age range, and
164 both cognitive outcomes. Given differences in cognitive measurements and score scales across cohorts
165 (Supplementary Figure 1), Spearman's rank correlation coefficient was used as the primary performance
166 metric to assess whether participants with better observed cognitive performance also tended to receive
167 better predicted scores, without requiring exact agreement in score units. We then assessed the
168 generalizability of best-performing ML models in independent cohorts using out-of-cohort validation. We
169 first benchmarked seven linear and non-linear ML algorithms, then compared feature combinations to
170 assess whether sleep measurements and parcel-based brain morphometry improved prediction beyond
171 demographics. Based on the selected models, we used SHapley Additive exPlanations (SHAP) (Lundberg
172 & Lee, 2017) and Shapley Interaction Quantification (SHAP-IQ) (Muschalik et al., 2024) to characterize
173 feature importance and their interactions. To assess whether these explanation patterns were
174 homogeneous across participants or instead reflected distinct participant profiles, we further applied
175 spectral clustering to SHAP values in the Greifswald cohort and evaluated the transferability of the
176 resulting clustering rules in the remaining cohorts (Fig. 1c).

177
178

Figure 1. Data description and machine learning pipeline.



a, Data were collected from 2,372 participants across seven cohorts within the ENIGMA-Sleep working group. The datasets included demographic information, sleep measurements, cognitive performance scores, and structural brain imaging data derived from T1-weighted MRI scans. **b**, Cohorts differed in age distributions, ranging from young-adult to predominantly older-adult samples. Greifswald cohort spans early to late adulthood and was used for model development. Details for each cohort are provided in Supplementary Figure 1 and Supplementary Table 1. **c**, T1 MRI preprocessing was performed with FreeSurfer's recon-all command to derive parcel-based cortical and subcortical measures. Data were integrated into a tabular format and exclusion of participants with incomplete data and outlier removal. Models were trained in the Greifswald cohort with nested cross-validation. Validation of Stroop predictions was conducted using data from Rotterdam, San Diego, Liège, and Pittsburgh, while validation of memory predictions utilized data from San Diego, Liège, Stockholm, Pittsburgh, and Jülich. Model explanations were generated with SHAP and SHAP-IQ to quantify feature importance and interactions. Participant subgroups were identified by spectral clustering based on SHAP values.

Prediction of cognitive performance in the Greifswald cohort

We first compared models based on demographics, sleep, brain morphometry features, and combinations to predict the Stroop and memory test scores in the Greifswald cohort using 5-fold 10-repeat nested cross-validation. Across both outcomes, the demographics-only model showed the strongest performance

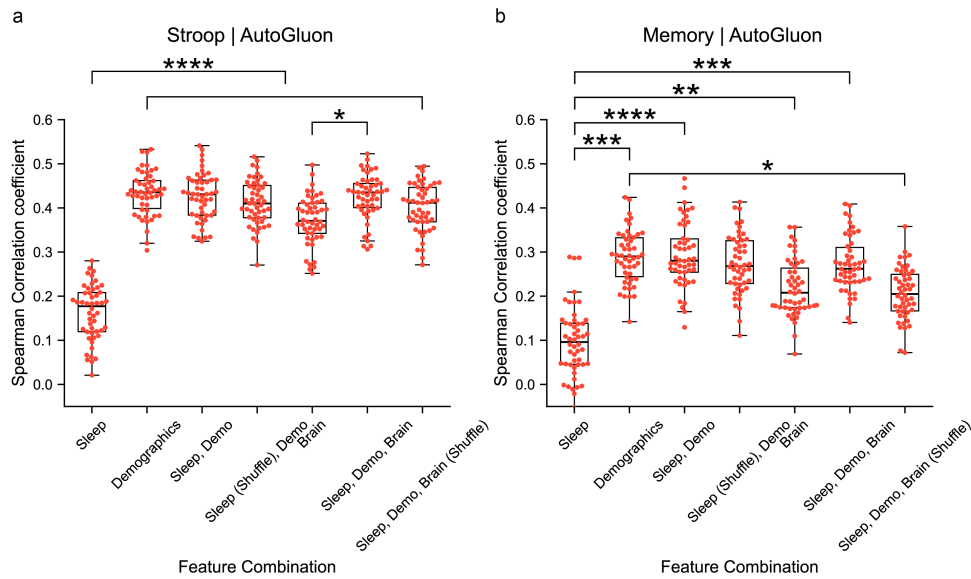
among single-domain feature sets, whereas sleep-only and brain-only models performed worse. Adding sleep to demographics or adding sleep and parcel-based brain morphometry to demographics did not significantly improve prediction beyond the demographics-only model. Non-linear models showed the most robust performance when the complete feature set was used. Therefore, AutoGluon was used as the baseline model for feature-set comparisons because it matched the best linear models with demographics-only and outperformed them with the complete feature set (Supplementary Figure 2).

Stroop test: With demographics-only features, AutoGluon achieved a robust prediction performance (test Spearman's $\rho = 0.43 \pm 0.05$, $R^2 = 0.15 \pm 0.05$, RMSE = 10.36 ± 1.48). Compared with the demographics-only model, the sleep-only model (test $\rho = 0.16 \pm 0.06$) performed worse ($P < 0.0001$), and the brain-only model yielded intermediate performance (test $\rho = 0.37 \pm 0.06$, $R^2 = 0.11 \pm 0.04$, RMSE = 10.60 ± 1.49). Adding sleep to demographics (test $\rho = 0.43 \pm 0.05$, $R^2 = 0.15 \pm 0.04$, RMSE = 10.34 ± 1.49) matched the demographics-only model and using all features together (test $\rho = 0.42 \pm 0.05$, $R^2 = 0.14 \pm 0.04$, RMSE = 10.42 ± 1.52), which was not significantly different from the demographics-only model. Shuffled-feature controls for sleep in sleep with demographics, and for brain in all feature sets, reduced performance slightly but not significantly (Fig. 2a, Supplementary Table 3).

Memory test: A similar pattern was observed for memory test score prediction. With demographics-only features, AutoGluon showed a robust prediction performance (test $\rho = 0.29 \pm 0.06$, $R^2 = 0.08 \pm 0.04$, RMSE = 15.08 ± 0.67). Compared with demographics-only, the sleep-only model (test $\rho = 0.10 \pm 0.08$) performed worse ($P < 0.001$), and the brain-only model yielded intermediate performance (test $\rho = 0.22 \pm 0.06$, $R^2 = 0.04 \pm 0.02$, RMSE = 15.42 ± 0.63). Adding sleep to demographics (test $\rho = 0.29 \pm 0.06$, $R^2 = 0.08 \pm 0.04$, RMSE = 15.06 ± 0.69) matched the demographics-only model and using all features together (test $\rho = 0.27 \pm 0.06$, $R^2 = 0.06 \pm 0.03$, RMSE = 15.23 ± 0.63), which was not significantly different from the demographic-only model. Shuffled-feature controls for sleep in sleep with demographics, and for brain in all feature sets, reduced performance slightly but not significantly (Fig. 2b, Supplementary Table 4).

Overall, in the Greifswald cohort, demographic variables captured most of the predictive signal for both Stroop and memory test scores, whereas sleep and parcel-based brain morphometry did not significantly improve prediction beyond the demographics-only model.

Figure 2. Prediction of cognitive performance by feature combinations in the Greifswald cohort using AutoGluon.



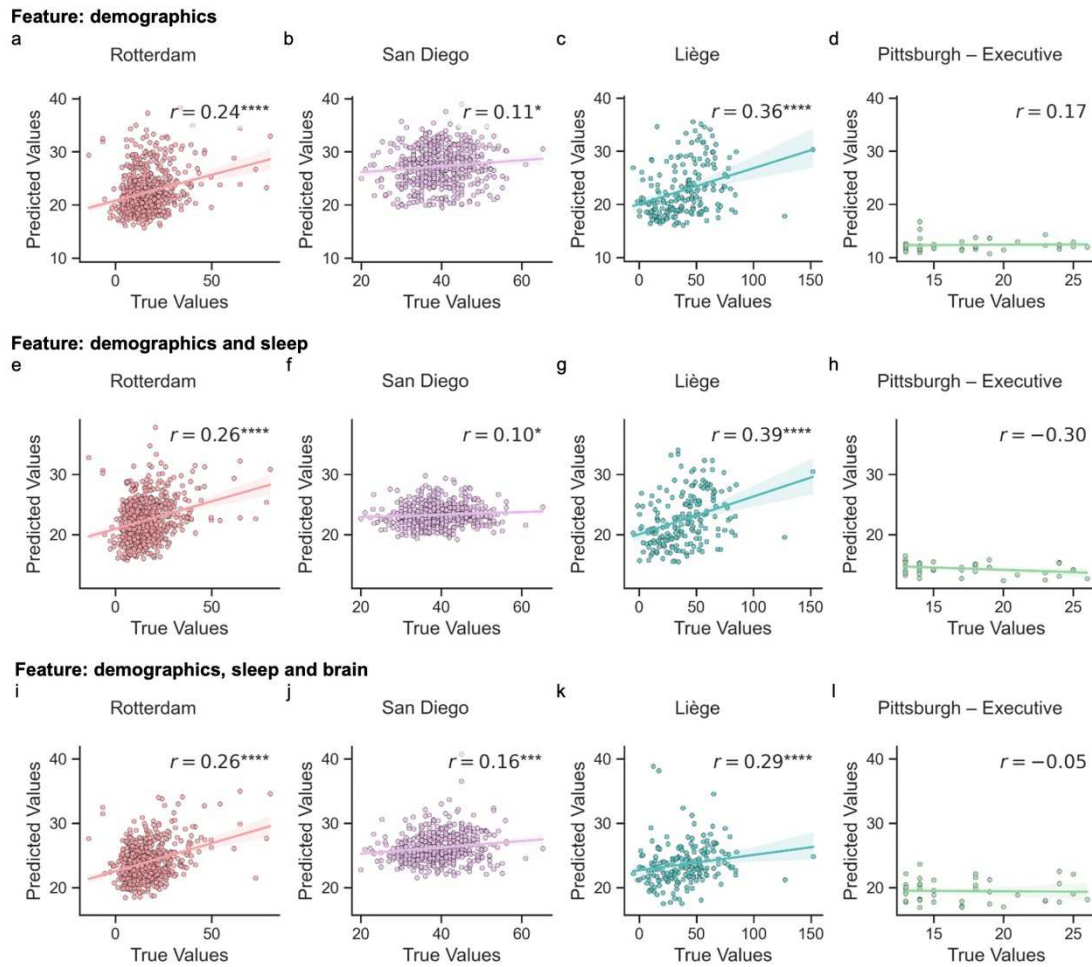
a, Stroop test score prediction across feature combinations by AutoGluon. Boxplots display test Spearman's rank correlation coefficient from 5-fold 10-repeat nested CV for feature combinations including sleep measurements only, demographics only, sleep with demographics, parcel-based brain morphometry only, and sleep, demographics with brain. Shuffled-feature controls for sleep measurements and parcel-based brain morphometry. **b**, Memory score prediction across the same feature combinations and evaluation scheme. Pairwise model comparison used Nadeau–Bengio corrected resampled paired t-tests on nested cross-validation scores with Bonferroni adjustment. Statistical significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

233 **Out-of-cohort validation of prediction models**

234 We next tested whether ML models trained in the Greifswald cohort generalized to independent cohorts.
 235 We selected AutoGluon as the best-performing model from the Greifswald cohort, which demonstrated
 236 robust, stable performance across feature sets. Detailed cohort-level validation results include all metrics for
 237 each feature set are provided from Supplementary Table 7 to Supplementary Table 22.

238 **Stroop test:** Out-of-cohort validation for Stroop test score prediction was assessed in the Rotterdam,
 239 San Diego, Liège, and Pittsburgh cohorts. In the Rotterdam cohort, the demographic-only model
 240 generalized significantly (Spearman $\rho = 0.24$, $P < 0.0001$), the model combining sleep with demographics
 241 performed similarly ($\rho = 0.26$, $P < 0.0001$), and the all-feature model also generalized ($\rho = 0.26$, $P <$
 242 0.0001). In the San Diego cohort, the demographics-only model generalized weakly but significantly ($\rho =$
 243 0.11 , $P < 0.05$) and the model combining sleep with demographics performed similarly ($\rho = 0.10$, $P < 0.05$),
 244 whereas the all-feature model showed stronger generalized performance ($\rho = 0.16$, $P < 0.001$). In the
 245 Liège cohort, the demographics-only model generalized significantly ($\rho = 0.36$, $P < 0.0001$), the model
 246 combining sleep with demographics generalized significantly ($\rho = 0.39$, $P < 0.0001$), and the all-feature
 247 model generalized significantly ($\rho = 0.29$, $P < 0.0001$). The Pittsburgh cohort contributed an executive-
 248 function proxy rather than a Stroop measure and did not show robust validation results (Fig. 3,
 249 Supplementary Table 3). Overall, Stroop test score prediction generalized across Rotterdam, San Diego,
 250 and Liège, with the relative performance of the feature sets varying across cohorts.

251
 252 **Figure 3. Out-of-cohort validation of Stroop test score prediction by AutoGluon.**

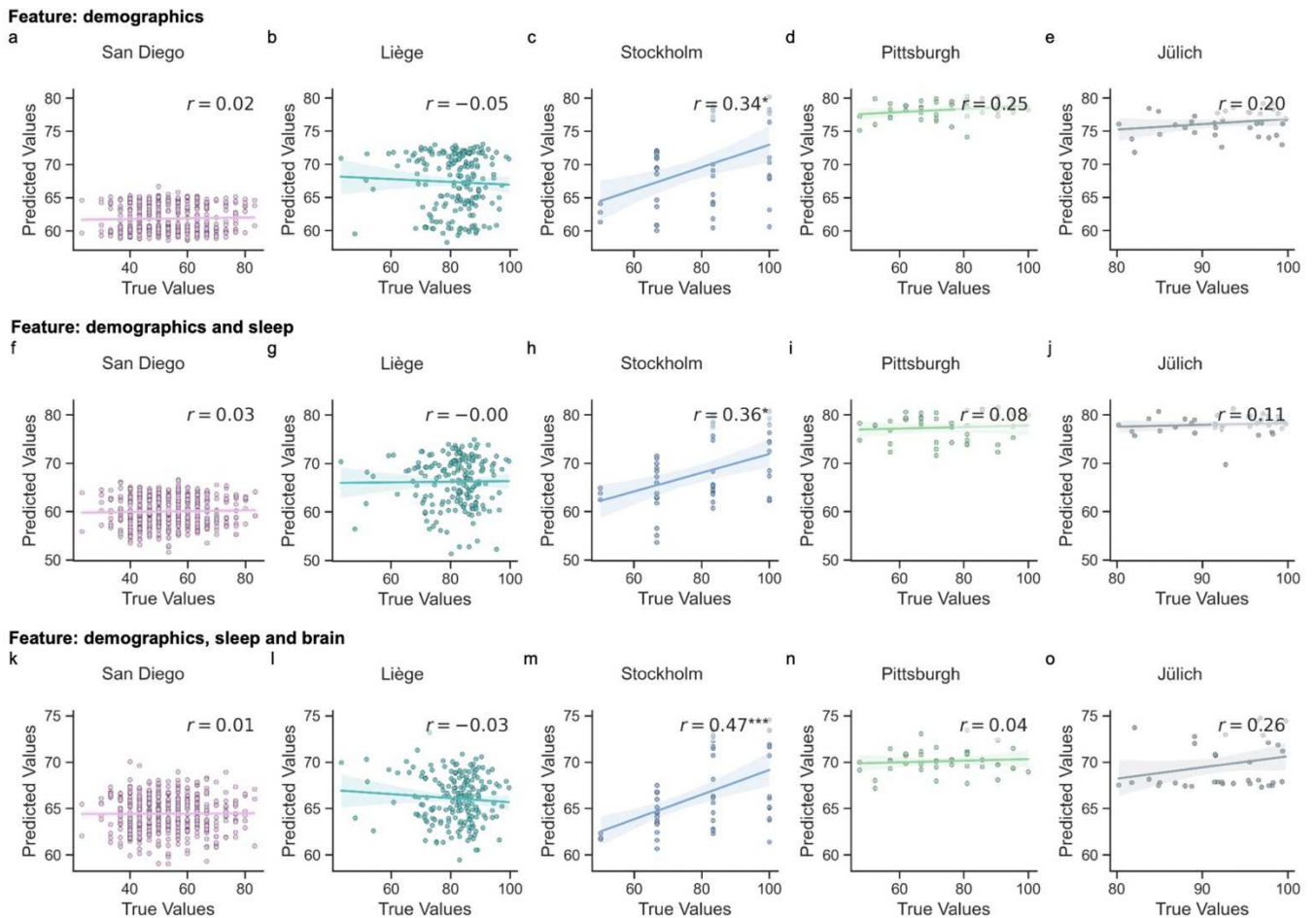


253
 254 AutoGluon models were trained in the Greifswald cohort and evaluated on four external cohorts. **a-d**, Model inputs are
 255 demographic features only. **e-h**, Model inputs are sleep measurements and demographics. **i-l**, Model inputs are sleep,
 256 demographics and parcel-based brain morphometry data. Within each group the panels from left to right correspond to Rotterdam,
 257 San Diego, Liège, and Pittsburgh. Each panel displays predicted versus true Stroop test scores. Significance of the Spearman's
 258 rank correlation coefficient is indicated by: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

259
260
261
262
263
264
265
266
267
268
269
270
271

Memory test: Based on data availability, out-of-cohort validation for memory test score prediction was assessed in the San Diego, Liège, Stockholm, Pittsburgh, and Jülich cohorts. The clearest validation was observed in Stockholm, where the demographics-only model generalized significantly ($\rho = 0.34$, $P < 0.05$), the model combining sleep with demographics generalized significantly ($\rho = 0.36$, $P < 0.05$), and the all-feature model showed the strongest generalized performance ($\rho = 0.47$, $P < 0.001$). In contrast, predictions were not validated in the San Diego or Liège cohorts due to differing scales (digit- and letter-based scores), and predictions in the Pittsburgh and Jülich cohorts were modest but not significant (Fig. 4, Supplementary Figure 3, Supplementary Table 4). Overall, out-of-cohort generalizability for memory test score prediction was more limited than for Stroop test score prediction and was most evident in the Stockholm cohort.

Figure 4. Out-of-cohort validation of memory test score prediction by AutoGluon.



272
273
274
275
276
277
278
279
280

AutoGluon models were trained in the Greifswald cohort and evaluated on five external cohorts. **a-e**, Model inputs are demographic features only. **f-j**, Model inputs are sleep measurements and demographics. **k-o**, Model inputs are sleep, demographics and parcel-based brain morphometry data. Within each feature combination, panels from left to right correspond to San Diego, Liège, Stockholm, Pittsburgh, and Jülich. Each panel displays predicted versus true memory test scores. The site-specific memory outcomes were WMS-III Digit Span in San Diego, n-back accuracy in Liège, spatial n-back accuracy in Stockholm, WMS-III Spatial Span in Pittsburgh, and spatial n-back accuracy from session 1 in Jülich. Significance of the Spearman's rank correlation coefficient is indicated by: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

281 Model explanation in the Greifswald cohort and out-of-cohort validation

282 To examine how individual features contributed to predictions, we used SHAP values to explain the
283 AutoGluon models trained on the Greifswald cohort. Here, we focused on models that use sleep and
284 demographics because they directly address how sleep measures contribute to prediction while remaining
285 low-dimensional enough for interpretable cross-cohort comparison. SHAP results for models including

286 parcel-based brain morphometry are shown in Supplementary Figures 4 and 5. We focused on models
287 that showed satisfactory external validity, defined a priori as a significant Spearman correlation between
288 observed and predicted scores. Under this criterion, model explanations for predicting the Stroop test
289 score are shown for Greifswald, San Diego, and Liège, and for predicting memory test score for Greifswald
290 and Stockholm.

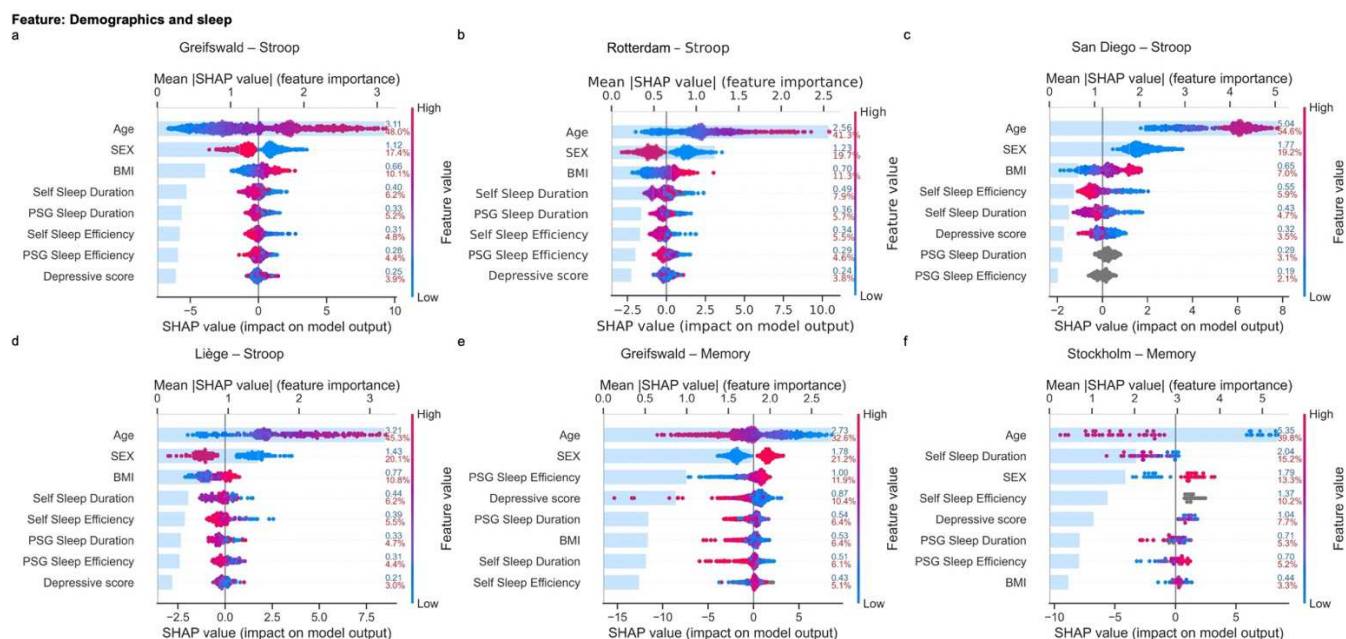
291 For Stroop test score prediction, age showed the largest mean absolute SHAP values in the sleep
292 and demographics models across all cohorts, followed by sex and BMI. Sleep duration and sleep efficiency
293 contributed only modestly, and depressive symptom scores showed consistently low importance (Fig. 5a-
294 d). In the models including all features, the Greifswald model was dominated by age and 3rd ventricle
295 volume, with additional contributions from prefrontal and cingulate cortical thickness parcels. Overall,
296 cortical thickness features carry more importance than cortical surface area. Similar patterns were
297 observed in the San Diego and Liège cohorts, where age, ventricular volume, and several frontal and
298 cingulate parcels remained among the most important contributors, despite minor site-specific shifts
299 (Supplementary Figure 4).

300 For memory test score prediction, age again showed the largest mean absolute SHAP values in the
301 sleep and demographic models in both Greifswald and Stockholm. Sex ranked second in Greifswald and
302 third in Stockholm. Sleep measurements contributed more to memory test score prediction than to Stroop
303 test score prediction. In Greifswald, polysomnography-based sleep measurements were more important
304 than self-reported sleep measures, whereas the reverse pattern was observed in Stockholm. BMI showed
305 consistently low importance in predicting memory test scores (Fig. 5e-f). When all features were included,
306 age remained the most important feature in Greifswald, followed by a broader pattern of brain
307 morphometry. Cortical thickness effects were most prominent in the ventromedial and orbitofrontal regions
308 and the occipital cortex, with additional contributions from cortical surface area and ventricular and other
309 subcortical volumes. Sleep variables contributed only marginally. The Stockholm cohort showed a similar
310 pattern, with age again ranking highest and overlapping frontal, occipital, and ventricular markers among
311 the most important features (Supplementary Figure 5).

312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332

333
334

Figure 5. Model explanation by SHAP for Stroop and Memory test score prediction.



335
336
337
338
339
340
341
342
343

a–d, Model explanations with SHAP for Stroop test score prediction using AutoGluon models with sleep and demographics in Greifswald (a), Rotterdam (b) San Diego (c), and Liège (d). **e–f**, Model explanations for memory test score prediction using the same feature combination in Greifswald (e) and Stockholm (f). Each panel shows a beeswarm plot of SHAP values for individual features, with points representing participants and point color indicating the underlying feature value. Each panel also includes a horizontal bar plot that displays the mean absolute SHAP value of each feature across all participants. Blue numbers next to each bar indicate the mean absolute SHAP value of that feature, and red numbers indicate the percentage contribution of that feature to the total mean absolute SHAP value across all features.

Feature interaction and subgroup analysis based on model explanation

345
346
347
348
349
350
351

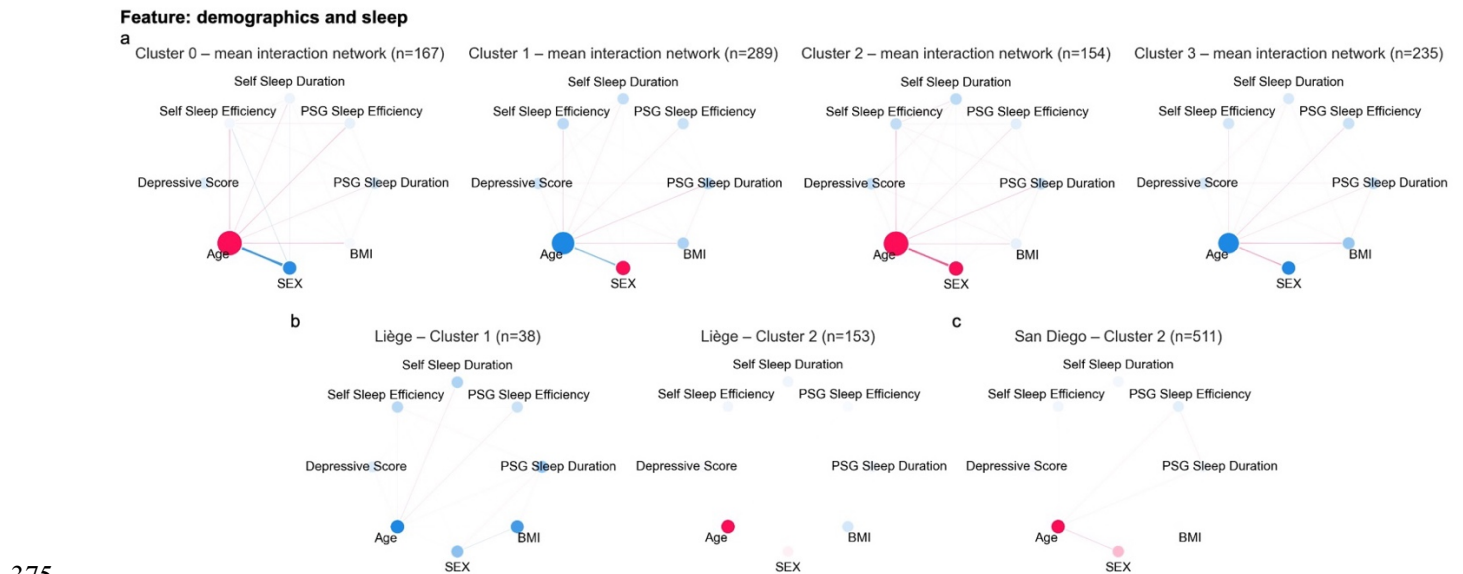
We next examined whether the contributions of sleep and demographic variables differed across participants by inspecting pairwise feature interactions and clustering participant-level SHAP profiles from the AutoGluon models trained with sleep and demographic features in the Greifswald cohort. SHAP dependence plots identified age as the dominant feature in both Stroop and memory test score prediction and suggested non-linear interactions between age and sleep duration, with shorter sleep in older participants and longer sleep in younger participants tending to be associated with worse predicted CP (Supplementary Figure 6).

352
353
354
355
356
357
358
359
360
361
362
363
364
365

To summarize these heterogeneous explanation patterns, we clustered participant-level SHAP values into subgroups that reflected participants with similar combinations of feature importance for prediction, rather than similarity in the raw input variables. The number of clusters was determined by clustering validity metrics and UMAP visualization (Supplementary Figure 7), yielding four subgroups for Stroop test score prediction (Fig. 6a) and six subgroups for memory test score prediction (Fig. 7a, Supplementary Figure 8). Across these subgroups, age provided the principal axis of separation, whereas the contributions of sleep-related features varied between subgroups. SHAP-IQ further indicated that interactions among age, sex, sleep duration, sleep efficiency, and depressive score were present but generally weaker than the corresponding main effects (Fig. 6a, Fig.7a). We used a decision tree-based algorithm (Goix et al., 2020) to define clustering rules for each subgroup based on the original input values of the AutoGluon model. These rules allowed us to characterize each predicted subgroup profile by identifying the most distinctive variables and value ranges associated with subgroup assignment (Supplementary Tables 5 and 6). Detailed rule-based subgroup definitions, cluster sizes, and cluster-specific interaction patterns are reported in Supplementary Results 1.

When the clustering rules derived in the Greifswald cohort were applied to the external validation cohorts, transferability was cohort-specific, with two Stroop-related subgroup profiles recovered in Liège, one in San Diego (Fig. 6b-c), and three memory-related subgroup profiles in Stockholm (Fig. 7b). The corresponding interaction patterns were broadly similar to those observed in the matched Greifswald subgroups (Supplementary Figures 9 and 10). Supplementary Results 1 further describes which Greifswald-derived subgroup profiles were recovered externally and why incomplete recovery may reflect cohort differences in age distribution, sex composition, sleep measurements, and cognitive measurements.

Figure 6. Participants clustering and feature interaction analysis based on SHAP values for Stroop test score prediction.



SHAP-IQ interaction networks were computed from AutoGluon models trained with sleep measurements and demographic features for Stroop test score prediction. **a**, Cluster-specific interaction networks for the four SHAP-derived participant subgroups in the Greifswald cohort. **b**, Interaction networks for the two participant subgroups identified when the Greifswald-derived clustering rules were applied to the Liège cohort. **c**, Interaction network for the single participant subgroup identified when the same clustering rules were applied to the San Diego cohort. Node size indicates the mean SHAP value within each subgroup, and node color indicates the direction of the feature contribution to the predicted Stroop test score. Edge thickness and opacity indicate SHAP-IQ interaction strength. For Stroop test scores, higher predicted values indicate lower cognitive performance.

Figure 7. Participants clustering and feature interaction analysis based on SHAP values for memory test score prediction.

Feature: demographics and sleep



SHAP-IQ interaction networks were computed from AutoGluon models trained with sleep measurements and demographic features for memory test score prediction. **a**, Cluster-specific interaction networks for the six SHAP-derived participant subgroups in the Greifswald cohort. **b**, Interaction networks for the three participant subgroups identified when the Greifswald-derived clustering rules were applied to the Stockholm cohort. Node size indicates the magnitude of the mean SHAP value within each subgroup, and node color indicates the direction of the feature contribution to the predicted memory test score. Edge thickness and opacity indicate SHAP-IQ interaction strength. For memory test scores, higher predicted values indicate better cognitive performance.

Sensitivity analysis based on APOE-ε4

To assess the contribution of APOE-ε4 status to Stroop and memory test score prediction, we trained AutoGluon models that combined APOE with sleep-alone, sleep and demographic, and all features. Then, we compared them to models with shuffled APOE labels and to corresponding models without APOE-ε4 (Supplementary Figure 11). Across all feature sets and outcomes, including APOE-ε4 produced changes in test Spearman ρ that were small in magnitude relative to the corresponding models without APOE-ε4 and did not reach statistical significance in corrected t-tests. Consistent performance across models with observed and shuffled APOE-ε4 indicated that APOE-ε4 status contributed little to predict Stroop and memory test scores in our data.

DISCUSSION

The present study aimed to predict CP from sets of features (demographic, subjective/objective sleep parameters, and regional brain morphometry) across multisite data. Our findings demonstrated: (i) Non-

linear models (i.e., AutoGluon) yielded the most robust performance for both Stroop and memory test score predictions in the largest cohort (Greifswald). Model comparisons across feature combinations further indicated that, unlike demographic-only models, sleep-only features did not provide robust predictive value. Sleep and brain morphometry added limited predictive value to the demographic-only model. Shuffling the sleep and brain features additionally revealed that demographic variables were the primary contributors to prediction performance. In a Greifswald's subset with available genotype data, adding APOE- ϵ 4 status did not improve cross-validated prediction, indicating limited incremental predictive value in these models; (ii) Out-of-cohort validation of the best-performing predictive model trained on the Greifswald cohort for the Stroop test revealed generalizable patterns across the Rotterdam, San Diego, and Liege cohorts, despite differences in demographic characteristics, including age range, across cohorts. For the memory test, a generalizable pattern was observed only in the Stockholm cohort, highlighting greater variance in acquired memory tests across cohorts; (iii) SHAP-based model explanation showed that age is the main feature contributor for both Stroop and memory test score prediction across all cohorts; (iv) Subgroup analysis based on SHAP values and SHAP-IQ explanations showed heterogeneous patterns for Stroop test score predictions driven by age, sex, and their interaction. For memory test prediction, interactions among age, sex, polysomnography-based sleep efficiency and duration, and depressive score were observed.

Although sleep and brain morphometry are relevant to CP (Kyle et al., 2017; Li et al., 2022), demographic factors, particularly chronological age, emerged as the most robust predictor in our study. This pattern was observed in the primary training dataset and in out-of-sample validation cohorts, suggesting that age captures stable, generalizable variance in CP across independent samples. One possible explanation is that age serves as a proxy for the cumulative effects of biological processes across multiple organs (O'Toole et al., 2026; Tian et al., 2023), many of which are not fully captured by simple measures of sleep duration or regional brain structure. This result is consistent with recent work showing that demographic factors can rival or even outperform multimodal neuroimaging markers in predicting sleep health characteristics and cognitive outcomes, highlighting their strong role in explaining population-level variance (Raimondo et al., 2026; Ujma et al., 2023). Similarly, prior studies have reported that sleep EEG explains only modest variance in cognition and contributes relatively little unique explanatory power after demographic adjustment, typically accounting for approximately 2.5–10% of cognitive variation (Tyurina et al., 2025; Ujma et al., 2023). More broadly, predictive modelling studies in sleep and mental health have shown that neuroimaging can provide limited added predictive utility beyond behavioral measures, clinical variables, and demographic characteristics (Olfati et al., 2024; Raimondo et al., 2026). In population-based studies, the cognitive consequences of poor sleep tend to be modest and heterogeneous (Fjell et al., 2023) and depend on how sleep is operationalized, in terms of duration, efficiency, regularity, fragmentation, or longitudinal change (Moon et al., 2024; West et al., 2024). This heterogeneity likely reflects the inherently dynamic nature of sleep, whose cognitive effects may accumulate over time or depend on chronicity, intra-individual variability, and interactions with aging, affective state, cerebrovascular burden, and circadian regulation (Bi et al., 2026; Tahmasian et al., 2026). A similar interpretation may apply to brain morphometry. Our cross-sectional, parcel-based structural measures likely provide only a coarse anatomical snapshot of neural variation, whereas CP is likely shaped by more distributed and temporally dynamic processes. Although it has been shown that demographic characteristics are also superior to resting-state fMRI in predicting cognitive phenotypes (Omidvarnia et al., 2024). Of note, the amount of our brain features is larger than sleep and demographics. In addition, as a substantial proportion of structural brain variance is age-related, the predictive contribution of morphometric features may be attenuated by age when included in multivariable models (Bi et al., 2026; Komeyer et al., 2025; Kyle et al., 2017). These findings suggest that statistically significant group-level associations between sleep and CP in large-scale UKB data were mainly driven by large

sample sizes rather than robust effect magnitudes. This has raised concerns about the practical relevance of the reported small effect sizes in UKB (Keyes & Westreich, 2019; Lin et al., 2013; Spence et al., 2026; Sullivan & Feinn, 2012), and thereby, does not necessarily translate into robust, cross-cohort markers for individual-level prediction, indicating that sleep variables may retain explanatory relevance while offering only limited incremental predictive value.

Our clustering and subgroup analysis were intended to move beyond global SHAP averages and test whether the model relied on the same feature relationships across all participants or instead captured distinct interaction patterns. In this context, an interaction means that part of the model prediction is jointly attributed to a pair of features beyond their separate main effects, so the prediction cannot be fully explained by summing the individual contributions of those features alone. For Stroop prediction, the subgroup patterns indicated that age and sex defined the dominant interaction structure. For example, in older males, age and male sex jointly reinforced a less favorable prediction pattern, whereas in older females, the adverse age effect appeared partly offset by a more favorable sex and age-sex interaction contribution. Thus, even when age remained the dominant feature, its predictive role differed across sex groups rather than following a single, uniform gradient. For memory prediction, the interaction structure was more differentiated and suggested a stronger role for objective sleep within specific demographic contexts. In younger males with better PSG sleep efficiency and lower depressive burden, PSG sleep efficiency and its interaction with age aligned with more favorable predicted memory performance, whereas in older males, objective sleep duration and efficiency appeared to partly buffer an otherwise less favorable age-sex interaction. The repeated divergence between PSG-based and self-reported sleep measures across memory subgroups further suggests that objective and subjective sleep do not contribute interchangeably to prediction. Taken together, these findings suggest that age and sex define the dominant predictive background on which sleep exerts subgroup-specific modifying effects, rather than acting as a uniform predictor across the full sample (Rahimi-Eichi et al., 2025; Sattari et al., 2024; Stankeviciute et al., 2025). Importantly, these clusters should not be interpreted as biological subtypes. Instead, they capture recurring model decision patterns, indicating how demographics and sleep contribute to CP prediction in a heterogeneous and context-dependent manner.

Here, we used ML as it provides an explicit framework for individual-level prediction and for testing generalizability across independent cohorts with different age distributions, which cannot be inferred from within-sample associations alone (Komeyer et al., 2025; Olfati et al., 2024; Omidvarnia et al., 2024; Rosenblatt et al., 2026; Wu et al., 2023). This design allowed us to test whether models trained in the broad adult age range of the Greifswald cohort retained predictive performance in cohorts with different age ranges (Mendl-Heinisch et al., 2026; Olfati et al., 2024), while recognizing that transferability remained outcome and cohort-dependent. Recent studies in sleep and mental health underscore the utility of predictive modeling, enabling validation across independent datasets and quantification of the incremental value of candidate feature domains (Olfati et al., 2024). Importantly, recent large-scale neuroimaging studies further illustrate that ML can serve as a stringent test of whether high-dimensional brain features meaningfully improve out-of-sample prediction. In some cases, predictive performance remains near chance, thereby placing important constraints on mechanistic interpretation and tempering expectations for clinical translation (Olfati et al., 2024; Omidvarnia et al., 2024; Raimondo et al., 2026). Although many AI applications in sleep medicine focus on diagnostics, staging, or monitoring rather than cognitive performance itself, they highlight a methodological point that is directly relevant here, namely that real-world utility depends on rigorous validation, transparent reporting, and explicit handling of heterogeneity and bias (Bi et al., 2026; Jahrami et al., 2026). These considerations motivated our unified, leakage-controlled ML framework across multiple cohorts. By strictly separating model development from external evaluation and fitting all preprocessing steps within each training split, the analysis yields conservative estimates of generalization under realistic site heterogeneity. Finally, SHAP and SHAP-IQ added an

important explanatory layer by distinguishing dominant global effects from smaller context-dependent effects and by revealing subgroup-specific interaction patterns that were not evident from performance metrics alone.

Our study has several limitations, which mainly concern the existing predictor domains, the comparability of cognitive targets across cohorts, and the methodological constraints of cross-cohort ML validation (Bi et al., 2026; Raimondo et al., 2026). First, we only focused on sleep duration and sleep efficiency from PSG and the Pittsburgh Sleep Quality Index (PSQI), which represent a subset of sleep health dimensions. Future work should incorporate long-term wearable-derived measures to capture night-to-night variability at scale (Massar et al., 2021). Multimodal sleep data, including detailed biological signals and subjective reports, and emerging sleep-focused foundation models may further improve representation of learning across cohorts under rigorous train-test separation (Thapa et al., 2026). Second, the cognitive phenotypes showed limited comparability across cohorts and likely variable reliability, particularly for memory, which may have limited out-of-cohort prediction and reduced apparent generalizability across sites (Gell et al., 2024; Rosenblatt et al., 2026). This may be more salient in multi-cohort settings where CP measures differ in content and scaling, where harmonization supports valid external validation. A recent proposed framework removes site-related variance while preserving instrument effects and links tests via a shared latent ability scale to facilitate cross-cohort score conversion (Kennedy et al., 2024, 2025). However, it relies on item-level data and is less applicable when only total scores are available. In ML settings, harmonization may yield optimistic estimates if information from hold-out participants influences the transformations, which motivates leakage-free harmonization that fits only the training data (Nieto et al., 2026). As the included cohorts differed in age distributions (Fig. 1b), depressive scores (Supplementary Table 1), and implementation of the cognitive tasks, we did not harmonize features or targets across sites to avoid cross-site information leakage (Nieto et al., 2026). Longitudinal designs will be essential to determine whether sleep and multimodal brain measures (e.g., white matter integrity, functional connectivity, vascular burden, or brain age gap) predict future cognitive decline more strongly than baseline CP.

545

546 CONCLUSION

This study suggests that while sleep and brain structure make modest contributions in predicting CP, age remains the primary and generalizable predictor in our multisite, individual-level models. Overall, these results indicate a complex contribution of demographic characteristics, sleep, and brain structure to CP. Our findings underscore the need for longitudinal, multimodal, and densely sampled data, as well as personalized approaches, to better understand how aging, sleep, and brain structure jointly shape CP at the individual level.

553

554 METHODS

555 Study participants and measurements

Study participants: Participants were included from seven cohorts of the ENIGMA-Sleep Working Group, yielding a total of 2,372 participants (N = 938 females; median age = 58.8 years; range = 18.0 – 95.1 years). Details of the cohorts' demographic characteristics, depressive scores, APOE-ε4 carrier status, and sleep measurements are summarized in Supplementary Table 1. All contributing sites received approval from their respective local institutional review boards and ethics committees to collect and analyze data. Cohort-specific inclusion and exclusion criteria before data sharing are listed in Supplementary Table 2.

Sleep measurements: Sleep duration and sleep efficiency were included as sleep predictors. These measures were derived from PSG and self-reported sleep questionnaires where available. For questionnaire-based sleep measures, the PSQI was used across cohorts with self-reported sleep data,

566 and the Karolinska Sleep Questionnaire (KSQ) (Nordin et al., 2013) was used only in the Stockholm cohort.
567 The availability and modality of sleep measures differed across cohorts and are summarized in
568 Supplementary Table 1. We retained both objective and self-reported measures when both were available,
569 because they capture related but non-equivalent aspects of sleep.

570 **Demographic variables:** Age, sex, body mass index (BMI), and depressive scores were included,
571 where available, as the demographic and clinical feature set used for prediction. Age was included
572 because cognitive performance, sleep, and brain morphometry vary across adulthood. Sex was included
573 because it is relevant to inter-individual differences in sleep, brain structure, and cognition. BMI was
574 included because it is associated with sleep and overall health status, which may affect cognitive
575 performance. Depressive scores were included because depressive symptoms are associated with both
576 sleep and cognition. These variables were entered as predictors and were not regressed out from the
577 input features or cognitive outcomes. Depressive symptoms were assessed with the Patient Health
578 Questionnaire-9 (PHQ-9) in Greifswald, the Beck Depression Inventory-II (BDI-II) (Beck et al., 2011) in
579 Liège, the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1977) in Rotterdam and
580 San Diego, and the Hospital Anxiety and Depression Scale (HADS) (Zigmond & Snaith, 1983) in
581 Stockholm. PHQ-9 was converted to the Beck Depression Inventory-II (BDI-II) scale before analysis.
582 HADS scores in Stockholm were transformed to the BDI-II scale using a piecewise mapping based on
583 severity ranges, with HADS scores of 0–7 mapped to BDI-II 0–13, 8–10 to 14–19, 11–15 to 20–28, and
584 16–21 to 29–63, followed by rounding to the nearest integer. CES-D scores were retained on their original
585 scale. Depressive scores were not available in Pittsburgh or Jülich.

586 **Stroop test score:** The Stroop test score was used as an executive function-related outcome. It
587 was measured as reaction time, defined as the interference cost in seconds computed as the difference
588 between the interference and neutral conditions in the Greifswald, Rotterdam, and Liège cohorts. In the
589 San Diego cohort, the Stroop test score was indexed by the Stroop Interference Norm-Based T-Score
590 (Golden et al., 1978), which reflects performance on the color-word interference condition. The Pittsburgh
591 cohort did not provide a Stroop measure and contributed an executive function proxy using the
592 Neuropsychological Assessment Battery (NAB) Mazes Test. To ensure consistency across cohorts, in
593 which lower Stroop test scores indicate better cognitive performance, we inverted the San Diego and
594 Pittsburgh scores by subtracting each score from the sum of the cohort-specific maximum and minimum.
595 Although measurement protocols differed across cohorts, the scales of these Stroop test scores were
596 comparable (Supplementary Figure 1), and models were therefore trained in the Greifswald cohort and
597 evaluated on other independent cohorts to assess the generalizability of ML models.

598 **Memory test score:** The memory test score was used as an accuracy-based memory outcome.
599 It reflected accuracy for n-back working memory tasks in the Liège, Stockholm, and Jülich cohorts. In the
600 Liège cohort, working memory performance was measured using the n-back paradigm with three load
601 conditions: 1-, 2-, and 3-back. To unify this measure with accuracy-based memory scores from other
602 cohorts, we computed one memory test score per participant by taking the mean accuracy across the
603 three n-back loads. In the Jülich cohort, working memory was assessed with letter and spatial versions of
604 an n-back task in which participants indicated whether the current stimulus matched the one presented N
605 trials earlier. For each load level (1-, 2-, and 3-back), performance was summarized by the sensitivity (A')
606 based on hits, false alarms, and omissions (Hennecke et al., 2021). We derived a single working memory
607 score by averaging A' across the available n-back loads, including both letter and spatial versions.

608 In Stockholm, a spatial n-back paradigm was used, implemented in the WakeApp web app
609 (Holding et al., 2021). Participants were presented with a grid in which squares light up in sequence, and
610 the task was to remember the sequence. After each sequence, one square lit up with a number to ask
611 whether that square was in the position indicated by the number in the sequence, and the participant
612 responded yes or no. The test included 6 6-long sequences and 6 2-long sequences.

613 In the Greifswald cohort, memory was assessed using the Nuremberg Age Inventory (NAI) word
614 list (Oswald & Fleischmann, 1999), a German verbal learning and memory test that presents eight words
615 and requires immediate and delayed recall in the presence of distractor words. Immediate recall was
616 scored as the number of correctly recalled target words, and delayed recognition was scored as the
617 number of correctly recognized target words minus the number of falsely endorsed distractor words, with
618 a maximum of eight points for each condition. In the present study, we derived a memory accuracy score
619 by dividing the immediate and delayed sum scores by their maximum possible value and averaging the
620 two scaled scores.

621 In the San Diego cohort, memory performance was assessed using the Wechsler Memory Scale–
622 Third Edition (WMS-III) Digit Span Forward, Digit Span Backward, and Letter-Number Sequencing
623 (Wechsler, 1997). For Digit Span, we summed the forward and backward raw scores and divided the sum
624 by the maximum possible score of 30 to obtain a digit span accuracy proportion. For Letter Number
625 Sequencing, we divided the total raw score by the maximum possible score of 21 to obtain a sequencing
626 accuracy proportion.

627 The Pittsburgh cohort contributed memory measures from the WMS-III Spatial Span and Letter-
628 Number Sequencing. Spatial Span Forward and Backward raw scores were combined into a total score
629 between 0 and 32. Letter-Number Sequencing yielded a total raw score ranging from 0 to 21. To place
630 these measures on a common accuracy scale, we transformed each raw score into a percentage score
631 by dividing the Spatial Span total by its maximum of 32 and the Letter-Number Sequencing total by its
632 maximum of 21.

633 **APOE-ε4 carriership:** APOE-ε4 carriership: APOE-ε4 carrier status was summarized where
634 available and used in sensitivity analyses to explore whether genetic risk information improved the
635 prediction of Stroop and memory test scores beyond the main feature sets. In San Diego data, APOE
636 genotype was determined using PCR and HhaI restriction digest methods. In Greifswald, Rotterdam, Liège,
637 and Stockholm, the APOE genotype was determined from imputed rs429358 and rs7412 variants. APOE-
638 ε4 carriership by cohort is reported in Supplementary Table 1.

639

640 **Image acquisition and processing**

641 Each cohort collected T1-weighted structural MRI at participating sites. Images were processed in each
642 cohort by the standard ENIGMA workflow with FreeSurfer (<https://github.com/ENIGMA-gift/>). Cohort-
643 specific image acquisition parameters and FreeSurfer version are included in the Supplementary Table 2.
644 For each participant, parcel-based cortical thickness and surface area were calculated using the Schaefer
645 atlas (400 regions and 7 networks) (Schaefer et al., 2018). Parcel-based subcortical volumes were
646 calculated using the ASEG atlas, yielding 36 volumetric regions, including bilateral lateral ventricles,
647 inferior lateral ventricles, cerebellar white matter, cerebellar cortex, thalamus, caudate, putamen, pallidum,
648 hippocampus, amygdala, nucleus accumbens, ventral diencephalon, and vessels, as well as the third,
649 fourth, and fifth ventricles, cerebrospinal fluid, brainstem, and five corpus callosum segments (Fischl et al.,
650 2002).

651

652 **Feature integration and quality control**

653 For each cohort, demographic information, sleep measurements, and parcel-based brain morphometry
654 features were combined into a tabular dataset with one row per participant. Participants with missing
655 phenotypic or imaging features were excluded. To detect potential image-processing failures and extreme
656 multivariate outliers, we applied Isolation Forest (Liu et al., 2008) implemented by Scikit-learn (Pedregosa
657 et al., 2011) within each cohort. The Isolation Forest was fitted on all features and assigned each
658 participant an inlier (+1) or outlier (−1) label, and participants classified as outliers were excluded. This
659 procedure was used only to detect and exclude participants, and did not modify feature values, use

660 outcome information, or create any derived dataset for subsequent analysis. This step was treated as
661 cohort-level quality control before model training rather than as a cross-validation preprocessing
662 transformation. It was performed without using Stroop or memory test scores, and the resulting outlier
663 labels were not estimated from, or optimized against, prediction performance.

664 We did not apply cross-site data harmonization in this study because standard harmonization methods
665 such as ComBat (Johnson et al., 2007) and ComBat-GAM (Pomponio et al., 2020) can introduce data
666 leakage in ML pipelines when they are fitted across combined training and test data or across cohorts
667 before cross-validation (Nieto et al., 2026). In addition, previous work has suggested that appropriately
668 trained ML models can mitigate site and demographic-related bias in performance without explicit
669 harmonization (Wang et al., 2023).

670

671 **Machine learning models**

672 To comprehensively characterize the link between sleep, demographics, brain morphometry, and
673 cognitive performance, and to assess both linear and nonlinear relationships, we benchmarked a panel of
674 seven supervised learning algorithms. These ranged from simple linear models, which are linear
675 regression, ridge regression, and SVM with linear kernel, to nonlinear kernel and tree-based ensembles,
676 which are SVM with rbf kernel, random forest, and XGBoost, and an automated ML framework i.e.,
677 AutoGluon. AutoGluon, as an automated tabular learning framework, can infer feature types, handle
678 common preprocessing steps such as missing data and feature rescaling, and train a diverse set of base
679 learners that are combined through bagging, stacking, and a greedy weighted ensemble. This spectrum
680 of models also allowed us to evaluate whether associations between sleep, brain morphometry, and
681 cognition were better captured by models that can represent nonlinear relations (Li et al., 2022) and
682 whether these associations were robust across different algorithm choices.

683

684 **Model training and internal evaluation**

685 Model training took place in the Greifswald cohort, which had the largest sample size, covered the age
686 range of the other cohorts, and included both Stroop and memory test scores. Prior to model training,
687 parcel-based morphometric features were ratio-adjusted for brain size to mitigate global scaling effects
688 (Nicolaisen-Sobesky et al., 2022). Depressive scores were entered as model predictors within the
689 demographics and were not used for residualizing either predictors or cognitive outcomes. Specifically,
690 cortical thickness, surface area, and subcortical volume were normalized by overall cortical thickness, total
691 surface area, and total intracranial volume (TIV) at the individual participant level. To ensure
692 generalizability and obtain test scores from the Greifswald cohort, we employed a 5-fold, 10-repeat nested
693 cross-validation (CV), yielding 50 outer train-test splits for model evaluation. To maintain distributional
694 consistency across folds, data splitting was stratified based on a demographic label generated by binning
695 age into tertiles (based on the Greifswald cohort quantiles) and interacting these age bins with sex.

696 For linear regression, ridge regression, SVM with linear kernel, SVM with rbf kernel, random forest,
697 and XGBoost models, feature preprocessing and hyperparameter tuning were conducted within the
698 training data of each outer split. For each outer split, continuous variables in the training data were
699 standardized with z-scores. The fitted transformers were then applied to the corresponding test data for
700 that split. Within each outer training set, hyperparameters for linear regression, ridge regression, SVM with
701 linear and rbf kernels, and random forest were tuned by grid search using 5-fold inner CV with coefficient
702 of determination (R^2) as the optimization metric, implemented with Julearn (Hamdan et al., 2024) and
703 Scikit-learn (Pedregosa et al., 2011). For XGBoost, Optuna optimized hyperparameters using 5-fold inner
704 CV with R^2 as the optimization metric (Akiba et al., 2019). The best hyperparameters from each outer split
705 were used to refit the model on that split's full training data, and performance was then evaluated on that
706 split's test data.

For AutoGluon models, in each outer split, models were trained on the original feature values in the training data, since AutoGluon includes internal logic for preprocessing operations such as missing value imputation and one-hot encoding. We used AutoGluon's TabularPredictor in regression mode with R^2 as the optimization metric and the best-quality preset, which trains a diverse stack ensemble of bagged models and automatically configures hyperparameters, stacking and bagging based on dataset properties (Erickson et al., 2020). In each outer split, the single best-performing model obtained from the training procedure was refitted on the training data and evaluated on that split's test data.

We summarized performance across outer splits by averaging each test metric from each split to obtain model performance in the Greifswald cohort. Spearman correlation between observed and predicted scores was the primary performance metric because it evaluates whether participants with higher observed cognitive performance also tended to receive higher predicted scores, is less sensitive to extreme values than Pearson correlation, and is less dependent on absolute score scaling across different cognitive tests. We additionally reported rooted mean squared error (RMSE), mean absolute error (MAE), R^2 , and Pearson correlation. Linear regression, ridge regression, SVM with linear and rbf kernel, random forest, and XGBoost were then refitted with the best performing hyperparameters on the full Greifswald dataset for out-of-cohort validation. AutoGluon models were retrained on the whole Greifswald cohort using the same configuration as in nested CV for out-of-cohort validation.

Model comparison on the Greifswald cohort

To evaluate the contribution of different feature sets to prediction, we trained models for each ML algorithm on the Stroop test score and the memory test score using multiple input feature sets, which comprised sleep measurements only, demographics only, sleep with demographics, parcel-based brain morphometry only, and sleep plus demographics with brain. In addition, as negative control models, we trained models in which sleep measurements were randomly shuffled while demographics and brain morphometry features were left intact, and models in which brain morphometry features were randomly shuffled while sleep and demographics were left intact. These negative control models were used to assess whether sleep or brain features added information beyond demographics.

Statistical comparisons of model performance were carried out within the Greifswald cohort to compare feature sets within each algorithm and to compare algorithms within each feature set. Comparisons were based on model performance on the outer split and used the Nadeau–Bengio corrected resampled paired t-test on nested cross-validation scores with Bonferroni adjustment for all pairwise contrasts with threshold (Nadeau & Bengio, 1999). As all algorithms used the same outer fold splits, these pairwise comparisons were comparable. Statistical comparisons were implemented with Julearn (Hamdan et al., 2024).

Out-of-cohort validation in independent cohorts

For out-of-cohort validation, models were trained in the Greifswald cohort and then applied separately to the Rotterdam, San Diego, Liège, and Pittsburgh cohorts for Stroop test score prediction, and to the San Diego, Liège, Stockholm, Pittsburgh, and Jülich cohorts for memory test score prediction. Consistent with model training and evaluation in the Greifswald cohort, cortical thickness, cortical surface area, and subcortical volumes were normalized by overall cortical thickness, total cortical surface area, and total intracranial volume at the individual participant level in each validation cohort. Based on the algorithm comparison in the Greifswald cohort, we focused on out-of-cohort validation on XGBoost and AutoGluon, which showed robust and stable performance across feature sets. For XGBoost, the z-score preprocessing parameters estimated from the full Greifswald dataset were applied to the corresponding features in each validation cohort. For each feature set, the final model refitted in the Greifswald cohort was then applied to each validation cohort. AutoGluon models retrained in the Greifswald cohort for each

feature set were applied directly to the validation cohorts. Spearman correlation between observed and predicted scores was the primary performance metric, and we additionally reported RMSE, MAE, R^2 , and Pearson correlation. For both Spearman and Pearson correlation, we computed two-sided p-values to assess whether predictions were significant.

758

759 **Model explanation**

The contribution of different feature sets to prediction was assessed by model comparison, whereas the contribution of individual features within a model was quantified with SHAP values. SHAP values were computed with the SHAP toolbox (Lundberg & Lee, 2017). Model explanations focused on AutoGluon, which showed stable and comparatively strong predictions across feature sets. For each prediction target and feature set, we used the AutoGluon model retrained on the full Greifswald cohort and applied the SHAP PartitionExplainer, which better handles correlations among features, particularly for parcel-based brain features, using the full Greifswald dataset as background data to derive SHAP values for all participants in Greifswald. The same explainer was then applied to out-of-validation cohorts by passing the corresponding cohort data as new input while keeping the background data and AutoGluon model fixed.

AutoGluon models trained with all feature sets were explained in the Greifswald cohort. For reporting, we focused on two feature sets. First, sleep with demographics, which addresses how sleep measures and demographics contribute to prediction without imaging features. Second, all features, including sleep, demographics, and brain morphometry, show individual feature contributions when all features are available. Out-of-cohort SHAP analyses were restricted to cohorts in which the AutoGluon model achieved statistically significant out-of-cohort validation performance, defined as a significant Spearman correlation between observed and predicted scores (two-sided $P < 0.05$). We therefore report model explanations for the San Diego and Liège cohorts in Stroop prediction and for the Stockholm cohort in memory prediction to avoid overinterpreting feature attributions in cohorts with low prediction performance.

Global feature importance was summarized as the mean absolute SHAP value per feature and visualized with bar plots. Local feature attributions were represented by the distribution of SHAP values across participants and visualized with beeswarm plots. For Stroop test score prediction, higher SHAP values indicate that a feature contribution pushes the model towards larger predicted Stroop test scores, corresponding to worse executive performance. For memory test score prediction, higher SHAP values indicate that a feature contribution pushes the model towards larger predicted memory scores, corresponding to better memory performance.

786

787 **Feature interaction analysis based on model explanation**

SHAP dependence plots were used to qualitatively visualize how feature values are related to their contributions to model predictions and to suggest potential feature interactions. Because conventional SHAP values assign prediction deviations to individual features and may distribute interaction effects across single-feature attributions, they do not directly quantify pairwise interactions. To quantify pairwise interaction effects beyond the marginal feature attributions provided by SHAP, we used the Shapley Interaction Quantification (SHAP-IQ) toolbox (Muschalik et al., 2024). Similar to the SHAP analysis, we applied the SHAP-IQ TabularExplainer to the AutoGluon models retrained on the full Greifswald cohort, using the full Greifswald dataset as background data, and derived SHAP values and pairwise interaction values for all participants in Greifswald. The same explainer was then applied to out-of-validation cohorts by passing the corresponding cohort data as new input while keeping the background data and model fixed. Out-of-cohort SHAP-IQ analyses were restricted to cohorts in which the AutoGluon model achieved satisfactory external validation performance, which are the San Diego and Liège cohorts for Stroop

prediction and the Stockholm cohort for memory prediction. The construction and interpretation of interaction networks based on these quantities is described in the subgroup analysis section.

Feature interactions analysis, derived from model explanations focused on sleep, with the demographic feature set. We restricted detailed interaction analyses to sleep with demographics, as this feature set directly targeted how sleep metrics interact with demographic factors in predicting cognitive performance and remained sufficiently low-dimensional for stable and interpretable interaction estimates. Including parcel-based brain morphometry features would have created a very high-dimensional interaction space with many weak, correlated effects.

SHAP-based subgroup analysis

To quantitatively characterize potential feature interactions that differed across participant groups, as suggested by the SHAP dependency plots, we clustered participant-level SHAP values to characterize potential feature interaction patterns that differed across subgroups. Subgroups were derived from SHAP values obtained from AutoGluon models trained on sleep and demographic features to predict Stroop and memory test scores in the Greifswald cohort. For each participant, we constructed an input vector consisting of SHAP values for all sleep and demographic features and used these SHAP vectors as the space for clustering. As different clustering algorithms impose different geometric and probabilistic assumptions on the data, we evaluated spectral clustering, k-means, and Gaussian mixture models. Algorithm choice and the number of clusters were guided by the gap statistic, silhouette score, Bayesian information criterion, and an elbow analysis of within-cluster sum of squares (Schubert, 2023), together with visual inspection of the resulting clustering structure. The final solution was selected as the combination of algorithm and cluster number that yielded clearly separated and interpretable subgroups. Uniform Manifold Approximation and Projection (UMAP) was used to embed the SHAP vectors into two dimensions for visualization (McInnes et al., 2020). Clustering algorithms and cluster validity metrics were implemented with scikit-learn, and UMAP was implemented with the Python UMAP package.

Each subgroup's model explanations were visualized using SHAP beeswarm plots and SHAP-IQ interaction networks. The beeswarm plots reveal which features dominate the predictions within a subgroup and how the directions and magnitudes of their effects vary across individuals, whereas the SHAP-IQ interaction networks expose subgroup-specific patterns of interacting features. For each subgroup's SHAP-IQ interaction network, we aggregated individual SHAP-IQ interaction values by averaging all participants in that cluster to obtain a subgroup-level interaction pattern.

To obtain human-readable cluster descriptions, we used SkopeRules (Goix et al., 2020) to learn decision rules that predicted SHAP-derived cluster labels from the original AutoGluon training features. We trained SkopeRules in the original feature space rather than on SHAP values because the raw variables are directly interpretable and consistently defined across cohorts. To transfer subgroup assignments to out-of-validation cohorts, the rules learned in the Greifswald cohort were applied directly as feature conditions. Participants were assigned to a subgroup when exactly one rule was satisfied and were left unassigned when no rule or more than one rule matched. For each subgroup in the out-of-validation cohorts, model explanations were also visualized using SHAP beeswarm plots and SHAP-IQ interaction networks.

Sensitivity analyses

Given its established relevance to cognitive performance and dementia, we conducted a sensitivity analysis to assess whether adding APOE-ε4 status improved prediction of Stroop and memory test scores from the other features. APOE-ε4 carrier status was available for a subset of participants (N = 785) in the Greifswald cohort. Within this cohort, we repeated the AutoGluon training and evaluation for Stroop and memory test score prediction using the same 5-fold 10-repeat nested CV design and evaluation metrics

847 as in the main AutoGluon models. Because APOE- ϵ 4 did not improve cross-validated prediction within the
848 Greifswald subset, APOE-based models were not advanced to out-of-cohort validation.

849 For both Stroop and memory test score prediction, we considered three feature configurations based
850 on the main AutoGluon models, that is, sleep-only, sleep plus demographic, and all features. For each
851 configuration, we trained two additional variants of the AutoGluon model: one model including APOE- ϵ 4
852 as an additional feature and another negative control model, including a randomly shuffled APOE- ϵ 4
853 feature. As in the main analyses, performance of the APOE- ϵ 4 models was summarized by the same set
854 of metrics, with Spearman correlation as the primary metric and MSE, MAE, R^2 , and Pearson correlation
855 also reported, averaged across outer test folds. To quantify the incremental predictive value of APOE- ϵ 4,
856 we compared model performance between models with and without APOE- ϵ 4 and between models with
857 observed and permuted APOE- ϵ 4 status using Nadeau–Bengio corrected resampled paired t-tests with
858 Bonferroni adjustment for all pairwise contrasts.

859

860 **Ethics & Inclusion statement**

861 This study was conducted within the ENIGMA-Sleep collaborative framework using de-identified data
862 contributed by participating cohorts. Each local site had received institutional ethical approval from the
863 relevant review board or ethics committee before data collection, and the present analysis of de-identified
864 cohort data was conducted in accordance with the applicable cohort-specific approvals, consent
865 procedures, and data-use agreements. Local cohort investigators were involved in data governance, data
866 curation, interpretation of cohort-specific information, manuscript revision, and authorship where they met
867 authorship criteria. Roles and responsibilities for data access, analysis, and manuscript preparation were
868 discussed among collaborators before and during the study. The study did not involve new participant
869 recruitment, transfer of biological materials, or procedures expected to cause stigmatization, incrimination,
870 discrimination, or additional personal risk to participants beyond the risks associated with the analysis of
871 de-identified health and neuroimaging data.

872

873 **Data availability**

874 The data used in this study were collected independently at participating sites and analyzed following the
875 standardized ENIGMA protocols. Due to participant privacy considerations, ethical restrictions, and site-
876 level data-sharing agreements, individual-level data cannot be made publicly available. Data requests will
877 be reviewed by the respective site and must comply with local ethics approvals and data-use agreements.
878 For the Greifswald SHIP-Trend cohort, data access applications should be submitted through the SHIP
879 transfer portal at <https://transfer.ship-med.uni-greifswald.de/FAIRequest/?lang=en> and will be reviewed
880 by the SHIP committee. For the San Diego cohort, instructions for data access requests are available on
881 the Vietnam Era Twin Study of Aging (VETSA) website at [https://psychiatry.ucsd.edu/research/programs-](https://psychiatry.ucsd.edu/research/programs-centers/vetsa/researchers.html)
882 [centers/vetsa/researchers.html](https://psychiatry.ucsd.edu/research/programs-centers/vetsa/researchers.html). De-identified VETSA Wave 1, 2, and 3 data are publicly available through
883 the National Archive of Computerized Data on Aging at
884 <https://www.icpsr.umich.edu/web/ICPSR/studies/38836>. For the Stockholm cohort, anonymized data are
885 available at <https://openneuro.org/datasets/ds000201>.

886

887 **Code availability**

888 This study was preregistered on the Open Science Framework (OSF, <https://osf.io/nhbkq/>). All analysis
889 tools are open-source in this study based on Python 3.10.18 using NumPy 2.1.3, pandas 2.3.3, SciPy
890 1.15.3, scikit-learn 1.7.2, XGBoost 3.0.5, AutoGluon 1.2.0, SHAP 0.49.1, SHAP-IQ 1.3.1, Skope-Rules
891 1.0.1, and UMAP-learn 0.5.7. Brain visualizations were generated based on Python 3.11.14 with NumPy
892 1.26.4, Nilearn 0.12.1, nibabel 5.3.2, neuromaps 0.0.5, brainspace 0.1.20, and surfplot 0.2.0. All scripts

for analyses in this study and Python environment specifications are provided in the GitHub repository (<https://github.com/harveybi/ENIGMA-Sleep-Cognitive-Performance>).

References (81 / typically recommend up to 50)

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, 2623–2631. <https://doi.org/10.1145/3292500.3330701>
- Beck, A. T., Steer, R. A., & Brown, G. (2011). *Beck Depression Inventory–II* [Dataset]. <https://doi.org/10.1037/t00742-000>
- Bi, H., Liu, W., & Tahmasian, M. (2026). Bridging artificial intelligence and sleep medicine: Current evidence, open challenges, and future directions. *Sleep Medicine Reviews*, 85, 102237. <https://doi.org/10.1016/j.smr.2026.102237>
- Brito, D. V. C., Esteves, F., Rajado, A. T., Silva, N., Araújo, I., Bragança, J., Castelo-Branco, P., & Nóbrega, C. (2023). Assessing cognitive decline in the aging brain: Lessons from rodent and human studies. *Npj Aging*, 9(1), 23. <https://doi.org/10.1038/s41514-023-00120-6>
- Coutrot, A., Lazar, A. S., Richards, M., Manley, E., Wiener, J. M., Dalton, R. C., Hornberger, M., & Spiers, H. J. (2022). Reported sleep duration reveals segmentation of the adult life-course into three phases. *Nature Communications*, 13(1), 7697. <https://doi.org/10.1038/s41467-022-34624-8>
- Deantoni, M., Reyt, M., Baillet, M., Dourte, M., De Haan, S., Lesoinne, A., Vandewalle, G., Maquet, P., Berthomier, C., Muto, V., Hammad, G., & Schmidt, C. (2024). Napping and circadian sleep–wake regulation during healthy aging. *Sleep*, 47(5), zsad287. <https://doi.org/10.1093/sleep/zsad287>
- Ell, J., Schiel, J. E., Feige, B., Riemann, D., Nyhuis, C. C., Fernandez-Mendoza, J., Vetter, C., Rutter, M. K., Kyle, S. D., & Spiegelhalter, K. (2023). Sleep health dimensions and shift work as longitudinal predictors of cognitive performance in the UK Biobank cohort. *SLEEP*, 46(6), zsad093. <https://doi.org/10.1093/sleep/zsad093>
- Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data* (arXiv:2003.06505). arXiv. <http://arxiv.org/abs/2003.06505>
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron*, 33(3), 341–355. [https://doi.org/10.1016/S0896-6273\(02\)00569-X](https://doi.org/10.1016/S0896-6273(02)00569-X)
- Fjell, A. M., Sørensen, Ø., Wang, Y., Amlie, I. K., Baaré, W. F. C., Bartres-Faz, D., Bertram, L., Boraxbekk, C.-J., Brandmaier, A. M., Demuth, I., Drevon, C. A., Ebmeier, K. P., Ghisletta, P., Kievit, R., Kühn, S., Madsen, K. S., Mowinkel, A. M., Nyberg, L., Sexton, C. E., ... Walhovd, K. B. (2023). No phenotypic or genotypic evidence for a link between sleep duration and brain atrophy. *Nature Human Behaviour*, 1–15. <https://doi.org/10.1038/s41562-023-01707-5>
- Fjell, A. M., Sørensen, Ø., Wang, Y., Amlie, I. K., Baaré, W. F. C., Bartres-Faz, D., Boraxbekk, C.-J., Brandmaier, A. M., Demuth, I., Drevon, C. A., Ebmeier, K. P., Ghisletta, P., Kievit, R., Kühn, S., Madsen, K. S., Nyberg, L., Solé-Padullés, C., Vidal-Piñeiro, D., Wagner, G., ... Walhovd, K. B. (2023). Is Short Sleep Bad for the Brain? Brain Structure and Cognitive Function in Short Sleepers. *The Journal of Neuroscience*, 43(28), 5241–5250. <https://doi.org/10.1523/JNEUROSCI.2330-22.2023>
- Fjell, A. M., & Walhovd, K. B. (2024). Individual sleep need is flexible and dynamically related to cognitive function. *Nature Human Behaviour*, 8(3), 422–430. <https://doi.org/10.1038/s41562-024-01827-6>
- Gell, M., Eickhoff, S. B., Omidvarnia, A., Küppers, V., Patil, K. R., Satterthwaite, T. D., Müller, V. I., & Langner, R. (2024). How measurement noise limits the accuracy of brain-behaviour predictions. *Nature Communications*, 15(1), 10678. <https://doi.org/10.1038/s41467-024-54022-6>
- Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nature Reviews Neuroscience*, 23(5), 307–318. <https://doi.org/10.1038/s41583-022-00584-7>
- Gläser, E., Kilimann, I., Platen, M., Hoffmann, W., Brosse, F., Buerger, K., Coenjaerts, M., Düzel, E., Ewers, M., Fliessbach, K., Frommann, I., Gemenetzi, M., Glanz, W., Hellmann-Regen, J., Incesoy, E. I., Janowitz, D., Jessen, F., Peters, O., Priller, J., ... Michalowsky, B. (2025). The economic burden of subjective cognitive decline, mild cognitive impairment and Alzheimer's dementia: Excess costs and associated clinical and risk factors. *Alzheimer's Research & Therapy*, 17, 142. <https://doi.org/10.1186/s13195-025-01785-9>

947 Goix, N., Birodkar, V., Gardin, F., Schertzer, J.-M., JEONG, H., kumar, manoj, Gramfort, A., Staley, T., Tour,
948 T. D. la, Deng, B., C, Pedregosa, F., Wu, L., Rokem, A., Jackson, K., & mrahim. (2020). *Scikit-learn-*
949 *contrib/skope-rules v1.0.1* [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.4316671>

950 Golden, C., Freshwater, S. M., & Golden, Z. (1978). *Stroop Color and Word Test*.
951 <https://doi.apa.org/doi/10.1037/t06065-000>

952 Hamdan, S., More, S., Sasse, L., Komeyer, V., Patil, K. R., Raimondo, F., & Initiative, for the A. D. N. (2024).
953 Julearn: An easy-to-use library for leakage-free evaluation and inspection of ML models. *Gigabyte*, 2024, 1–
954 16. <https://doi.org/10.46471/gigabyte.113>

955 Hanushek, E. A., Kinne, L., Witthöft, F., & Woessmann, L. (2025). Age and cognitive skills: Use it or lose it.
956 *Science Advances*, 11(10), eads1560. <https://doi.org/10.1126/sciadv.ads1560>

957 Hennecke, E., Lange, D., Steenbergen, F., Fronczek-Poncelet, J., Elmenhorst, D., Bauer, A., Aeschbach, D.,
958 & Elmenhorst, E.-M. (2021). Adverse interaction effects of chronic and acute sleep deficits on spatial working
959 memory but not on verbal working memory or declarative memory. *Journal of Sleep Research*, 30(4), e13225.
960 <https://doi.org/10.1111/jsr.13225>

961 Holding, B. C., Ingre, M., Petrovic, P., Sundelin, T., & Axelsson, J. (2021). Quantifying Cognitive Impairment
962 After Sleep Deprivation at Different Times of Day: A Proof of Concept Using Ultra-Short Smartphone-Based
963 Tests. *Frontiers in Behavioral Neuroscience*, 15. <https://doi.org/10.3389/fnbeh.2021.666146>

964 Ikram, M. A., Kieboom, B. C. T., Brouwer, W. P., Brusselle, G., Chaker, L., Ghanbari, M., Goedegebure, A.,
965 Ikram, M. K., Kavousi, M., de Knecht, R. J., Luik, A. I., van Meurs, J., Pardo, L. M., Rivadeneira, F., van Rooij,
966 F. J. A., Vernooij, M. W., Voortman, T., & Terzikhan, N. (2024). The Rotterdam Study. Design update and major
967 findings between 2020 and 2024. *European Journal of Epidemiology*, 39(2), 183–206.
968 <https://doi.org/10.1007/s10654-023-01094-1>

969 Jahrami, H., Husain, W., Trabelsi, K., Penzel, T., Hirshkowitz, M., Razjouyan, J., BaHammam, A. S., &
970 Sharafkhaneh, A. (2026). Artificial intelligence and sleep medicine II: A scoping review of applications,
971 advancements, and future directions. *Sleep Medicine Reviews*, 85, 102212.
972 <https://doi.org/10.1016/j.smr.2025.102212>

973 Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using
974 empirical Bayes methods. *Biostatistics*, 8(1), 118–127. <https://doi.org/10.1093/biostatistics/kxj037>

975 Kennedy, E., Vadlamani, S., Lindsey, H. M., Lei, P.-W., Jo-Pugh, M., Thompson, P. M., Tate, D. F., Hillary, F.
976 G., Dennis, E. L., Wilde, E. A., for the ENIGMA Clinical Endpoints Working Group, Adamson, M., Alda, M.,
977 Alonso-Lana, S., Ambrogi, S., Anderson, T. J., Arango, C., Asarnow, R. F., Avram, M., ... Zunta-Soares, G. B.
978 (2024). Bridging big data in the ENIGMA consortium to combine non-equivalent cognitive measures. *Scientific*
979 *Reports*, 14(1), 24289. <https://doi.org/10.1038/s41598-024-72968-x>

980 Kennedy, E., Vadlamani, S., Lindsey, H. M., Peterson, K. S., Dams O'Connor, K., Agarwal, R., Amiri, H. H.,
981 Andersen, R. K., Babikian, T., Baron, D. A., Bigler, E. D., Caeyenberghs, K., Delano-Wood, L., Disner, S. G.,
982 Dobryakova, E., Eapen, B. C., Edelstein, R. M., Esopenko, C., Genova, H. M., ... Dennis, E. L. (2025). Linking
983 Symptom Inventories Using Semantic Textual Similarity. *Journal of Neurotrauma*, 42(11–12), 1008–1020.
984 <https://doi.org/10.1089/neu.2024.0301>

985 Keyes, K. M., & Westreich, D. (2019). UK Biobank, big data, and the consequences of non-representativeness.
986 *The Lancet*, 393(10178), 1297. [https://doi.org/10.1016/S0140-6736\(18\)33067-8](https://doi.org/10.1016/S0140-6736(18)33067-8)

987 Komeyer, V., Nieto, N., Eickhoff, S. B., Raimondo, F., & Patil, K. R. (2025). Overview of Challenges in Brain-
988 Based Predictive Modeling: Toward Meaningful Predictive Insights. *Biological Psychiatry*.
989 <https://doi.org/10.1016/j.biopsych.2025.09.003>

990 Kremen, W. S., Franz, C. E., & Lyons, M. J. (2013). VETSA: The Vietnam Era Twin Study of Aging. *Twin*
991 *Research and Human Genetics*, 16(1), 399–402. <https://doi.org/10.1017/thg.2012.86>

992 Kyle, S. D., Sexton, C. E., Feige, B., Luik, A. I., Lane, J., Saxena, R., Anderson, S. G., Bechtold, D. A., Dixon,
993 W., Little, M. A., Ray, D., Riemann, D., Espie, C. A., Rutter, M. K., & Spiegelhalter, K. (2017). Sleep and
994 cognitive performance: Cross-sectional associations in the UK Biobank. *Sleep Medicine*, 38, 85–91.
995 <https://doi.org/10.1016/j.sleep.2017.07.001>

996 LaGoy, A. D., Mayeli, A., Smagula, S. F., & Ferrarelli, F. (2022). Relationships between rest-activity rhythms,
997 sleep, and clinical symptoms in individuals at clinical high risk for psychosis and healthy comparison subjects.
998 *Journal of Psychiatric Research*, 155, 465–470. <https://doi.org/10.1016/j.jpsychires.2022.09.009>

999 Li, Y., Gong, W., Sahakian, B. J., Huang, S., Zhang, W., Zhao, Y., Ma, L., Naismith, S., Yu, J., Jia, T., Cheng,
1000 W., & Feng, J. (2025). Divergent biological pathways linking short and long sleep durations to mental and
1001 physical health. *Nature Mental Health*, 1–15. <https://doi.org/10.1038/s44220-025-00395-6>

1002 Li, Y., Sahakian, B. J., Kang, J., Langley, C., Zhang, W., Xie, C., Xiang, S., Yu, J., Cheng, W., & Feng, J. (2022).
1003 The brain structure and genetic mechanisms underlying the nonlinear association between sleep duration,
1004 cognition and mental health. *Nature Aging*, 2(5), 425–437. <https://doi.org/10.1038/s43587-022-00210-2>

1005 Lin, M., Lucas, H. C., & Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. *Information*
1006 *Systems Research*, 24(4), 906–917. <https://doi.org/10.1287/isre.2013.0480>

1007 Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on*
1008 *Data Mining*, 413–422. <https://doi.org/10.1109/ICDM.2008.17>

1009 Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberger, U., & Tucker-Drob, E. M. (2020). Education and
1010 Cognitive Functioning Across the Life Span. *Psychological Science in the Public Interest: A Journal of the*
1011 *American Psychological Society*, 21(1), 6–41. <https://doi.org/10.1177/1529100620920576>

1012 Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural*
1013 *Information Processing Systems*, 30.
1014 https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

1015 Massar, S. A. A., Chua, X. Y., Soon, C. S., Ng, A. S. C., Ong, J. L., Chee, N. I. Y. N., Lee, T. S., Ghosh, A., &
1016 Chee, M. W. L. (2021). Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and
1017 self-report data. *Npj Digital Medicine*, 4(1), 90. <https://doi.org/10.1038/s41746-021-00466-9>

1018 McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for*
1019 *Dimension Reduction* (arXiv:1802.03426). arXiv. <https://doi.org/10.48550/arXiv.1802.03426>

1020 Mendl-Heinisch, C., Bittner, N., Miller, T., Dellani, P., Bamberg, F., Berger, K., Bohmann, P., Decker, J. A.,
1021 Flöel, A., Greiser, K. H., Harries, M., Kapar, J., Keil, T., Klett-Tammen, C. J., Krist, L., Kröncke, T., Leitzmann,
1022 M., Niendorf, T., Peters, A., ... Jockwitz, C. (2026). Prediction of cognitive test scores: A comparison of brain
1023 structure, health, demographic, and cognitive data across adulthood. *GeroScience*.
1024 <https://doi.org/10.1007/s11357-026-02232-9>

1025 Moon, C., Schneider, A., Cho, Y.-E., Zhang, M., Dang, H., & Vu, K. (2024). Sleep duration, sleep efficiency,
1026 and amyloid β among cognitively healthy later-life adults: A systematic review and meta-analysis. *BMC*
1027 *Geriatrics*, 24(1), 408. <https://doi.org/10.1186/s12877-024-05010-4>

1028 Muschalik, M., Baniecki, H., Fumagalli, F., Kolpaczki, P., Hammer, B., & Hüllermeier, E. (2024). shapiq:
1029 Shapley Interactions for Machine Learning. *Advances in Neural Information Processing Systems*, 37, 130324–
1030 130357. <https://doi.org/10.52202/079017-4141>

1031 Nadeau, C., & Bengio, Y. (1999). Inference for the Generalization Error. *Advances in Neural Information*
1032 *Processing Systems*, 12.
1033 [https://proceedings.neurips.cc/paper_files/paper/1999/hash/7d12b66d3df6af8d429c1a357d8b9e1a-](https://proceedings.neurips.cc/paper_files/paper/1999/hash/7d12b66d3df6af8d429c1a357d8b9e1a-Abstract.html)
1034 [Abstract.html](https://proceedings.neurips.cc/paper_files/paper/1999/hash/7d12b66d3df6af8d429c1a357d8b9e1a-Abstract.html)

1035 Nicolaisen-Sobesky, E., Mihalik, A., Kharabian-Masouleh, S., Ferreira, F. S., Hoffstaedter, F., Schwender, H.,
1036 Maleki Balajoo, S., Valk, S. L., Eickhoff, S. B., Yeo, B. T. T., Mourao-Miranda, J., & Genon, S. (2022). A cross-
1037 cohort replicable and heritable latent dimension linking behaviour to multi-featured brain structure.
1038 *Communications Biology*, 5(1), 1–11. <https://doi.org/10.1038/s42003-022-04244-5>

1039 Nieto, N., Eickhoff, S. B., Jung, C., Reuter, M., Diers, K., Kelm, M., Lichtenberg, A., Raimondo, F., & Patil, K.
1040 R. (2026). Impact of leakage on data harmonization in machine learning pipelines in class imbalance across
1041 sites. *Neurocomputing*, 133146. <https://doi.org/10.1016/j.neucom.2026.133146>

1042 Nilsson, G., Tamm, S., D'Onofrio, P., Thuné, H. Å., Schwarz, J., Lavebratt, C., Liu, J. J., Månsson, K. N.,
1043 Sundelin, T., Axelsson, J., Fransson, P., Kecklund, G., Fischer, H., Lekander, M., & Åkerstedt, T. (2020). *The*
1044 *Stockholm Sleepy Brain Study: Effects of Sleep Deprivation on Cognitive and Emotional Processing in Young*
1045 *and Old [Dataset]*. Openneuro. <https://doi.org/10.18112/OPENNEURO.DS000201.V1.0.3>

1046 Nordin, M., Åkerstedt, T., & Nordin, S. (2013). Psychometric evaluation and normative data for the Karolinska
1047 Sleep Questionnaire: Evaluation of the KSQ. *Sleep and Biological Rhythms*, 11(4), 216–226.
1048 <https://doi.org/10.1111/sbr.12024>

1049 Olfati, M., Samea, F., Faghihroohi, S., Balajoo, S. M., Küppers, V., Genon, S., Patil, K., Eickhoff, S. B., &
1050 Tahmasian, M. (2024). Prediction of depressive symptoms severity based on sleep quality, anxiety, and gray
1051 matter volume: A generalizable machine learning approach across three datasets. *eBioMedicine*, 108, 105313.
1052 <https://doi.org/10.1016/j.ebiom.2024.105313>

1053 Omidvarnia, A., Sasse, L., Larabi, D. I., Raimondo, F., Hoffstaedter, F., Kasper, J., Dukart, J., Petersen, M.,
1054 Cheng, B., Thomalla, G., Eickhoff, S. B., & Patil, K. R. (2024). Individual characteristics outperform resting-
1055 state fMRI for the prediction of behavioral phenotypes. *Communications Biology*, 7(1), 771.
1056 <https://doi.org/10.1038/s42003-024-06438-5>

1057 Oswald, W. D., & Fleischmann, U. M. (1999). *Nürnberger-Alters-Inventar:(NAI); NAI-testmanual und-textband*.
1058 Verlag für Psychologie, Hogrefe.

1059 O'Toole, C. K., Song, Z., Anagnostakis, F., Yang, Z., Tian, Y. E., Duggan, M. R., Zou, C., Leng, Y., Cai, Y., Bai,
1060 W., Fu, C. H. Y., Rafii, M. S., Aisen, P., Wang, G., De Jager, P. L., Zeng, J., Oh, H. S.-H., Zhou, X., Walker, K.
1061 A., ... Wen, J. (2026). Sleep chart of biological ageing clocks in middle and late life. *Nature*, 1–11.
1062 <https://doi.org/10.1038/s41586-026-10524-5>

1063 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,
1064 Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É.
1065 (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(null), 2825–2830.
1066 <https://doi.org/10.5555/1953048.2078195>

1067 Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M.,
1068 Satterthwaite, T. D., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C.,
1069 Fripp, J., Koutsouleris, N., Wolf, D. H., ... Davatzikos, C. (2020). Harmonization of large MRI datasets for the
1070 analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208, 116450.
1071 <https://doi.org/10.1016/j.neuroimage.2019.116450>

1072 Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General
1073 Population. *Applied Psychological Measurement*, 1(3), 385–401.
1074 <https://doi.org/10.1177/014662167700100306>

1075 Rahimi-Eichi, H., Baker, J. T., Fjell, A. M., & Buckner, R. L. (2025). Age- and sex-related differences in sleep
1076 patterns and their relations to self-reported sleep and mood. *SLEEP Advances*, 6(4), zpaf079.
1077 <https://doi.org/10.1093/sleepadvances/zpaf079>

1078 Raimondo, F., Bi, H., Komeyer, V., Kasper, J., Primus, S., Hoffstaedter, F., Mandal, S., Waite, L., Winkelmann,
1079 J., Oexle, K., Eickhoff, S. B., Tahmasian, M., & Patil, K. R. (2026). Can we predict sleep health based on brain
1080 features? A large-scale machine learning study using UK Biobank. *Brain Communications*, fcag016.
1081 <https://doi.org/10.1093/braincomms/fcag016>

1082 Rosenblatt, M., Foster, M. L., Adkinson, B. D., Tejavibulya, L., Khaitova, M., Ye, J., Sun, H., Rodriguez, R. X.,
1083 Camp, C. C., Chinta, A., McCusker, M. C., Han, L., Fields, C. T., Mehta, S., & Scheinost, D. (2026). External
1084 validation improves generalizability, replicability and reproducibility in predictive models for neuroimaging.
1085 *Nature Methods*, 1–11. <https://doi.org/10.1038/s41592-026-03115-9>

1086 Sattari, N., Mander, B. A., Dave, A., Lui, K. K., Sprecher, K. E., Chappel-Farley, M. G., Riedner, B. A., Bendlin,
1087 B. B., & Benca, R. M. (2024). Sex Modulates the Influence of APOE Genotype on Sleep and Sleep-Dependent
1088 Memory Consolidation in Older Adults. *Alzheimer's & Dementia*, 20(S1), e090492.
1089 <https://doi.org/10.1002/alz.090492>

1090 Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T.
1091 T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI.
1092 *Cerebral Cortex*, 28(9), 3095–3114. <https://doi.org/10.1093/cercor/bhx179>

1093 Schiel, J. E., Tamm, S., Holub, F., Petri, R., Dashti, H. S., Domschke, K., Feige, B., Goodman, M. O., Jones,
1094 S. E., Lane, J. M., Ratti, P.-L., Ray, D. W., Redline, S., Riemann, D., Rutter, M. K., Saxena, R., Sexton, C. E.,
1095 Tahmasian, M., Wang, H., ... Spiegelhalter, K. (2023). Associations between sleep health and grey matter
1096 volume in the UK Biobank cohort ($n = 33\,356$). *Brain Communications*, 5(4), fcad200.
1097 <https://doi.org/10.1093/braincomms/fcad200>

1098 Schubert, E. (2023). Stop using the elbow criterion for k-means and how to choose the number of clusters
1099 instead. *SIGKDD Explor. Newsl.*, 25(1), 36–42. <https://doi.org/10.1145/3606274.3606278>

1100 Spence, J. P., Mostafavi, H., Ota, M., Milind, N., Gjorgjieva, T., Smith, C. J., Simons, Y. B., Sella, G., & Pritchard,
1101 J. K. (2026). Specificity, length and luck drive gene rankings in association studies. *Nature*, 649(8098), 918–
1102 925. <https://doi.org/10.1038/s41586-025-09703-7>

1103 Stankeviciute, L., Tort-Colet, N., Fernández-Arcos, A., Sánchez-Benavides, G., Minguillón, C., Fauria, K., Holst,
1104 S. C., Garcés, P., Mueggler, T., Zetterberg, H., Blennow, K., Iranzo, Á., Suárez-Calvet, M., Gispert, J. D.,
1105 Molinuevo, J. L., Grau-Rivera, O., & Study, for the A. (2025). Associations between objective sleep metrics
1106 and brain structure in cognitively unimpaired adults: Interactions with sex and Alzheimer's biomarkers.
1107 *Alzheimer's & Dementia*, 21(6), e70353. <https://doi.org/10.1002/alz.70353>

1108 Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—Or Why the P Value Is Not Enough. *Journal of Graduate*
1109 *Medical Education*, 4(3), 279–282. <https://doi.org/10.4300/JGME-D-12-00156.1>

1110 Tahmasian, M., Aleman, A., Andreassen, O. A., Arab, Z., Baillet, M., Benedetti, F., Bresser, T., Bright, J., Chee,
1111 M. W. L., Chylinski, D., Cheng, W., Deantoni, M., Dresler, M., Eickhoff, S. B., Eickhoff, C. R., Elvsåshagen, T.,
1112 Feng, J., Foster-Dingley, J. C., Ganjgahi, H., ... Zarei, M. (2021). ENIGMA-Sleep: Challenges, opportunities,
1113 and the road map. *Journal of Sleep Research*, 30(6). <https://doi.org/10.1111/jsr.13347>

1114 Tahmasian, M., Elberse, J. D., Ahmadi, R., Liu, W., Rosenzweig, I., Genon, S., Eickhoff, S. B., & Mander, B.
1115 A. (2026). Sleep disturbances and Alzheimer's disease: A multiscale approach from exposome to neurobiology
1116 and precision medicine. *GeroScience*. <https://doi.org/10.1007/s11357-026-02152-8>

1117 Tai, X. Y., Chen, C., Manohar, S., & Husain, M. (2022). Impact of sleep duration on executive function and
1118 brain structure. *Communications Biology*, 5(1), 1–10. <https://doi.org/10.1038/s42003-022-03123-3>

1119 Thapa, R., Kjaer, M. R., He, B., Covert, I., Moore IV, H., Hanif, U., Ganjoo, G., Westover, M. B., Jennum, P.,
1120 Brink-Kjaer, A., Mignot, E., & Zou, J. (2026). A multimodal sleep foundation model for disease prediction. *Nature*
1121 *Medicine*, 1–11. <https://doi.org/10.1038/s41591-025-04133-4>

1122 Tian, Y. E., Cropley, V., Maier, A. B., Lautenschlager, N. T., Breakspear, M., & Zalesky, A. (2023).
1123 Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nature*
1124 *Medicine*, 29(5), 1221–1231. <https://doi.org/10.1038/s41591-023-02296-6>

1125 Tyurina, D. D., Stasenkov, S. V., Lushnikov, K. V., & Vedunova, M. V. (2025). Using Machine Learning Methods
1126 to Predict Cognitive Age from Psychophysiological Tests. *Healthcare*, 13(24), 3193.
1127 <https://doi.org/10.3390/healthcare13243193>

1128 Ujma, P. P., Bódizs, R., Dresler, M., Simor, P., Purcell, S., Stone, K. L., Yaffe, K., & Redline, S. (2023).
1129 Multivariate prediction of cognitive performance from the sleep electroencephalogram. *NeuroImage*, 279,
1130 120319. <https://doi.org/10.1016/j.neuroimage.2023.120319>

1131 Varoquaux, G., & Poldrack, R. A. (2019). Predictive models avoid excessive reductionism in cognitive
1132 neuroimaging. *Current Opinion in Neurobiology, Machine Learning, Big Data, and Neuroscience*, 55, 1–6.
1133 <https://doi.org/10.1016/j.conb.2018.11.002>

1134 Völzke, H., Schössow, J., Schmidt, C. O., Jürgens, C., Richter, A., Werner, A., Werner, N., Radke, D., Teumer,
1135 A., Ittermann, T., Schauer, B., Henck, V., Friedrich, N., Hannemann, A., Winter, T., Nauck, M., Dörr, M., Bahls,
1136 M., Felix, S. B., ... Kocher, T. (2022). Cohort Profile Update: The Study of Health in Pomerania (SHIP).
1137 *International Journal of Epidemiology*, 51(6), e372–e383. <https://doi.org/10.1093/ije/dyac034>

1138 Wang, R., Chaudhari, P., & Davatzikos, C. (2023). Bias in machine learning models can be significantly
1139 mitigated by careful training: Evidence from neuroimaging studies. *Proceedings of the National Academy of*
1140 *Sciences*, 120(6), e2211613120. <https://doi.org/10.1073/pnas.2211613120>

1141 Wechsler, D. (1997). *Wechsler Memory Scale—Third Edition (WMS-III)*. The Psychological Corporation.

1142 Weihs, A., Frenzel, S., Bi, H., Schiel, J. E., Afshani, M., Bülow, R., Ewert, R., Fietze, I., Hoffstaedter, F.,
1143 Jahanshad, N., Khazaie, H., Riemann, D., Rostampour, M., Stubbe, B., Thomopoulos, S. I., Thompson, P. M.,
1144 Valk, S. L., Völzke, H., Zarei, M., ... Group, for the E. N. G. through M.-A. (ENIGMA)-S. W. (2023). Lack of
1145 structural brain alterations associated with insomnia: Findings from the ENIGMA-Sleep Working Group. *Journal*
1146 *of Sleep Research*, 32(5), e13884. <https://doi.org/10.1111/jsr.13884>

1147 West, R., Wong, R. T. C., Park, J.-E., Lee, S. W., Mudiyanse, D. E., Liu, Z., Ma, D., & Stockman, L. (2024).
1148 Sleep duration, chronotype, health and lifestyle factors affect cognition: A UK Biobank cross-sectional study.
1149 *BMJ Public Health*. <https://doi.org/10.1136/bmjph-2024-001000>

1150 Willoughby, A. R., Alikhani, I., Karsikas, M., Chua, X. Y., & Chee, M. W. L. (2023). Country differences in
1151 nocturnal sleep variability: Observations from a large-scale, long-term sleep wearable study. *Sleep Medicine*,
1152 110, 155–165. <https://doi.org/10.1016/j.sleep.2023.08.010>

1153 Wu, J., Li, J., Eickhoff, S. B., Scheinost, D., & Genon, S. (2023). The challenges and prospects of brain-based
1154 prediction of behaviour. *Nature Human Behaviour*, 7(8), 1255–1264. <https://doi.org/10.1038/s41562-023-01670-1>

1155

1156 Wulan, N., An, L., Zhang, C., Kong, R., Chen, P., Bzdok, D., Eickhoff, S. B., Holmes, A. J., & Yeo, B. T. T.
1157 (2024). Translating phenotypic prediction models from big to small anatomical MRI data using meta-matching.
1158 *Imaging Neuroscience*, 2, imag-2–00251. https://doi.org/10.1162/imag_a_00251

1159 Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica*
1160 *Scandinavica*, 67(6), 361–370. <https://doi.org/10.1111/j.1600-0447.1983.tb09716.x>

Acknowledgements

A.I is supported by grants from the Multi-partner consortium to expand dementia research in Latin America [ReDLat2, supported by Fogarty International Center (FIC), National Institutes of Health, National Institutes of Aging (R01 AG057234, R01 AG075775, R01 AG21051, R01 AG083799, CARDS-NIH, R01 AG057234), Alzheimer's Association (SG-20-725707), Rainwater Charitable Foundation – The Bluefield project to cure FTD, and Global Brain Health Institute)], ANID/FONDECYT Regular (1250091 and 1210176 and 1220995); ANID/PIA/ANILLOS ACT210096; JPI JPND-Care, DISCeRN 2025 - Health and Social Care Research with a Focus on the Moderate and Late Stages of Neurodegenerative Diseases; FONDEF ID20I10152, and ANID/FONDAP 15150012; Wellcome Trust award for BRAIN-CLIMA: Investigating the Combined Impact of Heat and Air Pollution on Blood-Brain Barrier Integrity and Brain Aging in Latin America, (335293/Z/25/Z), Wellcome Leap CARE Program (Grant Number: CARE-2025-0883490149) for the project “Advancing Female-Specific Predictive Models and Risk Assessment Tools for Alzheimer’s Disease in the US and Latin America, and the CliCBrain (Horizon ID: 101236426; DOI 10.3030/101236426, Marie Skłodowska-Curie Actions - MSCA). The Study of Health in Pomerania (SHIP) is part of the Community Medicine Research Net (CMR) (<http://www.medizin.uni-greifswald.de/icm>) of the University of Greifswald funded by grants from the German Federal Ministry of Education and Research (BMBF, grants 01ZZ96030 and 01ZZ0701). The MRIs in SHIP were supported by a joint grant from Siemens Healthineers, Erlangen, Germany, and the Federal State of Mecklenburg-Western Pomerania. PSG assessment was in part supported by the German RLS organization (Deutsche Restless Legs Vereinigung). The contents of this publication are solely the responsibility of the authors and do not represent the official views of these institutions. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The VETSA is supported by the National Institute on Aging at the National Institutes of Health (grant numbers R01s AG076838, AG050595, AG022381, and AG064955). The content of this manuscript is the responsibility of the authors and does not represent official views of NIA/NIH, or the Veterans’ Administration. Numerous organizations provided invaluable assistance in the conduct of the VET Registry, including: U.S. Department of Veterans Affairs, Department of Defense; National Personnel Records Center, National Archives and Records Administration; Internal Revenue Service; National Opinion Research Center; National Research Council, National Academy of Sciences; the Institute for Survey Research, Temple University. The ENIGMA Working Group acknowledges the NIH Big Data to Knowledge (BD2K) award for foundational support and consortium development (U54 EB020403 to Paul M. Thompson). For a complete list of ENIGMA-related grant support please see here: <http://enigma.ini.usc.edu/about-2/funding/>.

1199 **Author information**

1200 **Authors and Affiliations**

1201 **Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich,**

1202 **Jülich, Germany**

1203 Hanwen Bi, Sarah Genon, Felix Hoffstaedter, Vincent Küppers, Wen Liu, Amin Saberi, Sofie L. Valk, Laura

1204 K. Waite, Kaustubh R. Patil, Simon B. Eickhoff, Federico Raimondo & Masoud Tahmasian

1205

1206 **Institute of Systems Neuroscience, Medical Faculty and University Hospital Düsseldorf, Heinrich**

1207 **Heine University, Düsseldorf, Germany**

1208 Hanwen Bi, Sarah Genon, Felix Hoffstaedter, Wen Liu, Amin Saberi, Sofie L. Valk, Kaustubh R. Patil,

1209 Simon B. Eickhoff, Federico Raimondo & Masoud Tahmasian

1210

1211 **Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden**

1212 Torbjörn Åkerstedt, Gustav Nilsson, Sandra Tamm & Joseph Wexler

1213

1214 **Department of Psychology, Stockholm University, Stockholm, Sweden**

1215 Torbjörn Åkerstedt

1216

1217 **Institute of Diagnostic Radiology and Neuroradiology, University Medicine Greifswald, Greifswald,**

1218 **Germany**

1219 Robin Bülow

1220

1221 **GIGA-CRC-Human Imaging, University of Liège, Liège, Belgium**

1222 Michele Deantoni, Nasrin Mortazavi, Christina Schmidt & Gilles Vandewalle

1223

1224 **Department of Nuclear Medicine, Faculty of Medicine and University Hospital Cologne, University**

1225 **of Cologne, Cologne, Germany**

1226 Alexander Drzezga, David Elmenhorst, Vincent Küppers & Masoud Tahmasian

1227

1228 **German Center for Neurodegenerative Diseases (DZNE), Bonn-Cologne, Germany**

1229 Alexander Drzezga

1230

1231 **Institute for Neuroscience and Medicine, Molecular Organization of the Brain (INM-2), Research**

1232 **Center Jülich, Jülich, Germany**

1233 Alexander Drzezga & David Elmenhorst

1234

1235 **Center for Behavior Genetics of Aging, Department of Psychiatry, University of California San**

1236 **Diego**

1237 Jeremy A. Elman, Christine Fennema-Notestine, Matthew S. Panizzon & Tina Thi Vo-Eckerle

1238

1239 **Department of Psychiatry, University of California San Diego**

1240 Jeremy A. Elman, Christine Fennema-Notestine, Matthew S. Panizzon & Tina Thi Vo-Eckerle

1241

1242 **Department of Sleep and Human Factors Research, German Aerospace Center, Cologne, Germany**

1243 Eva-Maria Elmenhorst

1244

1245 **Institute for Occupational, Social and Environmental Medicine, Medical Faculty, RWTH Aachen**
 1246 **University, Aachen, Germany**
 1247 Eva-Maria Elmenhorst
 1248
 1249 **Department of Internal Medicine B-Cardiology Pneumology, Infectious Diseases Intensive Care**
 1250 **Medicine, University Medicine Greifswald**
 1251 Ralf Ewert & Beate Stubbe
 1252
 1253 **Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA**
 1254 Fabio Ferrarelli, Ahmadrza Keihani, Ahmad Mayeli & Julia S. Rupp
 1255
 1256 **Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald,**
 1257 **Germany**
 1258 Stefan Frenzel, Hans J. Grabe, Katharina Wittfeld & Antoine Weihs
 1259
 1260 **Institute of Anatomy II, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany**
 1261 Charlotte von Gall
 1262
 1263 **German Center for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, Greifswald,**
 1264 **Germany**
 1265 Hans J. Grabe & Antoine Weihs
 1266
 1267 **Department of Radiology and Nuclear Medicine, Erasmus University Medical Centre, Rotterdam,**
 1268 **the Netherlands**
 1269 Phuong Thuy Nguyen Ho & Julia Neitzel
 1270
 1271 **Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam,**
 1272 **Netherlands**
 1273 Sanne J.W. Hoepel & Julia Neitzel
 1274
 1275 **Latin American Brain Health Institute (BrainLat), Universidad Adolfo Ibáñez, Santiago, Chile**
 1276 Agustin Ibanez
 1277
 1278 **Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland**
 1279 Agustin Ibanez
 1280
 1281 **Department of Biophysics, School of Medicine, Istanbul Medipol University, 34815 Istanbul,**
 1282 **Türkiye**
 1283 Agustin Ibanez
 1284
 1285 **Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, 08005, Barcelona,**
 1286 **Spain**
 1287 Agustin Ibanez
 1288
 1289 **Cognitive Neuroscience Center (CNC), Universidad de San Andrés, Buenos Aires, Argentina**
 1290 Agustin Ibanez
 1291

1292 **Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck**
1293 **School of Medicine of the University of Southern California, Los Angeles, CA, USA**
1294 Neda Jahanshad, Sophia I. Thomopoulos & Paul M. Thompson
1295
1296 **Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts**
1297 Julia Neitzel
1298
1299 **International Globally Distributed Organization for Research and Education (IGDORE), Stockholm,**
1300 **Sweden**
1301 Gustav Nilssonne
1302
1303 **Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany**
1304 Amin Saberi & Sofie L. Valk
1305
1306 **Department of Psychiatry and Psychotherapy, Medical Center – University of Freiburg, Faculty of**
1307 **Medicine, University of Freiburg, Germany**
1308 Kai Spiegelhalder
1309
1310 **Institute for Community Medicine, SHIP/Clinical-Epidemiological Research, University Medicine**
1311 **Greifswald, Greifswald, Germany**
1312 Henry Völzke
1313
1314 **German Centre for Cardiovascular Research (DZHK), Partner Site Greifswald, Greifswald,**
1315 **Germany**
1316 Henry Völzke
1317
1318 **Department of Psychology, Stanford University, Stanford, CA, USA**
1319 Joseph Wexler
1320
1321 **Consortia**
1322 **ENIGMA Sleep Working Group**
1323 Hanwen Bi, Michele Deantoni, David Elmenhorst, Fabio Ferrarelli, Hans J. Grabe, Sanne J.W. Hoepel,
1324 Felix Hoffstaedter, Neda Jahanshad, Ahmad Mayeli, Nasrin Mortazavi, Julia Neitzel, Gustav Nilssonne,
1325 Amin Saberi, Christina Schmidt, Kai Spiegelhalder, Sandra Tamm, Sophia I. Thomopoulos, Paul M.
1326 Thompson, Sofie L. Valk, Gilles Vandewalle, Henry Völzke, Joseph Wexler, Kaustubh R. Patil, Antoine
1327 Weihs, Simon B. Eickhoff, Federico Raimondo, Masoud Tahmasian
1328
1329 **Contributions**
1330 H.B. and M.T. designed the research, analyzed and interpreted the results, and wrote the paper.
1331 H.B. conducted the research, analyzed the data, made the figures, and prepared, reviewed and
1332 published the code. F.R. contributed to the study design and interpretation of the results. R.B.,
1333 R.E., S.F., H.J.G., B.S., H.V., K.W., A.W., P.T.N.H., S.J.W.H., J.N., J.A.E., C.F.-N., M.S.P.,
1334 T.T.V.-E., M.D., N.M., C.S., G.V., T.Å., G.N., S.T., J.W., F.F., A.K., A.M., J.S.R., D.E. and E.-M.E.
1335 contributed data. P.T.N.H. and S.J.W.H. conducted out-of-cohort validation analyses in the
1336 Rotterdam data. F.H., V.K., W.L., S.L.V., K.R.P., C.V.G., A.I., K.S., S.I.T., P.M.T., N.J., S.B.E.,
1337 A.W., A.S., S.G., T.T.V.-E., C.F.-N., K.W., S.F., A.D., and L.K.W. provided methodological,

1338 analytic, conceptual, consortium, or neuroimaging support and critically reviewed the paper. F.R.
1339 and M.T. supervised the project. All authors edited the paper and approved its final contents.

1340
1341 **Corresponding author**

1342 Correspondence to Masoud Tahmasian.

1343
1344 **Conflict of interest**

1345 H.J.G. has received travel grants and speaker honoraria from Neuraxpharm, Servier, Indorsia, and
1346 Janssen-Cilag. All other authors declare no conflicts of interest.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials.pdf](#)