GeroScience
https://doi.org/10.1007/s11357-026-02232-9

ORIGINAL ARTICLE

Prediction of cognitive test scores: a comparison of brain structure, health, demographic, and cognitive data across adulthood

Camilla Mendl-Heinisch · Nora Bittner · Tatiana Miller · Paulo Dellani · Fabian Bamberg · Klaus Berger · Patricia Bohmann · Josua A. Decker · Agnes Flöel · Karin Halina Greiser · Manuela Harries · Jan Kapar · Thomas Keil · Carolina J. Klett-Tammen · Lilian Krist · Thomas Kröncke · Michael Leitzmann · Thoralf Niendorf · Annette Peters · Tobias Pischon · Oliver Riedel · Steffen Ringhof · Christopher L. Schlett · Matthias B. Schulze · Mark O. Wielpütz · Kerstin Wirkner · Svenja Caspers · Christiane Jockwitz

Received: 9 February 2026 / Accepted: 16 March 2026

Abstract Cognitive performance prediction may help identify early cognitive decline. However, the heterogeneity of research findings impedes the identification of key predictors. This study used 21,877 participants (25–74 years) from the German National Cohort (NAKO Gesundheitsstudie, NAKO) to systematically predict cognitive test scores based on brain structure, demographic, health-related, and cognitive data. Importantly, validation analyses were performed across study sites and external samples (1000BRAINS). Higher predictability was observed in the total sample compared to age-specific subgroups (10% difference in explained variance). Demographic (e.g. age) and cognitive data (e.g. memory) outperformed brain structure (e.g. grey matter volume) and health-related data (e.g. hypertension). Cognitive tests were differentially predictable, most evident between episodic memory and motor speed ($R^2 \leq 0.32$ versus $R^2 \leq 0.18$). Differences in predictability between age groups finally highlight the importance of comparing prediction outcomes between

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11357-026-02232-9.

C. Mendl-Heinisch · N. Bittner · T. Miller · S. Caspers · C. Jockwitz
Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Jülich, Germany
e-mail: c.mendl-heinisch@fz-juelich.de

N. Bittner
e-mail: n.bittner@fz-juelich.de

T. Miller
e-mail: t.miller@fz-juelich.de

S. Caspers
e-mail: s.caspers@fz-juelich.de

C. Mendl-Heinisch · N. Bittner · T. Miller · P. Dellani · S. Caspers · C. Jockwitz (✉)
Institute for Anatomy I, Medical Faculty & University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
e-mail: c.jockwitz@fz-juelich.de

F. Bamberg · C. L. Schlett
Department of Diagnostic and Interventional Radiology, Medical Center – University of Freiburg, Faculty of Medicine – University of Freiburg, Freiburg, Germany
e-mail: fabian.bamberg@uniklinik-freiburg.de

C. L. Schlett
e-mail: christopher.schlett@uniklinik-freiburg.de

K. Berger
Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany
e-mail: bergerk@uni-muenster.de

P. Bohmann · M. Leitzmann
Department of Preventive Medicine and Epidemiology, University of Regensburg, Regensburg, Germany
e-mail: patricia.bohmann@ukr.de

Published online: 22 April 2026

M. Leitzmann
e-mail: michael.leitzmann@klinik.uni-regensburg.de

P. Bohmann
Department of Neurology, Medbo District Hospital and University Hospital of Regensburg, Regensburg, Germany

J. A. Decker · T. Kröncke
Department of Diagnostic and Interventional Radiology, University Hospital Augsburg, Augsburg, Germany
e-mail: Josua.Decker@uk-augsburg.de

T. Kröncke
e-mail: thomas.kroencke@uk-augsburg.de

A. Flöel
Department of Neurology, University Medicine Greifswald, Greifswald, Germany
e-mail: agnes.floeel@med.uni-greifswald.de

A. Flöel
German Center for Neurodegenerative Diseases (DZNE), Standort Rostock/Greifswald, Greifswald, Germany

K. H. Greiser
Division of Cancer Epidemiology, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany
e-mail: h.greiser@dkfz.de

M. Harries · C. J. Klett-Tammen
Department of Epidemiology, Helmholtz Centre for Infection Research, Brunswick, Germany
e-mail: manuela.harries@helmholtz-hzi.de

C. J. Klett-Tammen
e-mail: carolina.klett-tammen@helmholtz-hzi.de

J. Kapar · O. Riedel
Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
e-mail: kapar@leibniz-bips.de

O. Riedel
e-mail: riedel@leibniz-bips.de

J. Kapar
Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany

T. Keil · L. Krist
Institute of Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Berlin, Germany
e-mail: thomas.keil@charite.de

L. Krist
e-mail: lilian.krist@charite.de

T. Keil
Institute of Clinical Epidemiology and Biometry, University of Würzburg, Würzburg, Germany

T. Keil
State Institute of Health I, Bavarian Health and Food Safety Authority, Erlangen, Germany

T. Niendorf
Berlin Ultrahigh Field Facility (B.U.F.F.), Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany
e-mail: thoralf.niendorf@mdc-berlin.de

A. Peters
Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Neuherberg, Germany
e-mail: peters_sekretariat@helmholtz-muenchen.de

A. Peters
Chair of Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, Munich, Germany

A. Peters
German Center for Mental Health (DZPG), Partner Site Munich, Munich, Germany

T. Pischon
Max Delbrück- Center for Molecular Medicine in the Helmholtz Association (MDC), Molecular Epidemiology Research Group, Berlin, Germany
e-mail: tobias.pischon@mdc-berlin.de

T. Pischon
Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Biobank Technology Platform, Berlin, Germany

T. Pischon
Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität Zu Berlin, Berlin, Germany

S. Ringhof
Department of Radiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
e-mail: steffen.ringhof@uniklinik-freiburg.de

M. B. Schulze
Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam Rehbruecke, Nuthetal, Germany
e-mail: mschulze@dife.de

M. B. Schulze
Institute of Nutritional Science, University of Potsdam, Nuthetal, Germany

M. O. Wielpütz
Department of Diagnostic Radiology and Neuroradiology, University Medicine Greifswald, Greifswald, Germany
e-mail: mark.wielpuetz@med.uni-greifswald.de

K. Wirkner
Leipzig Research Centre for Civilization Diseases (LIFE), Leipzig University, Leipzig, Germany
e-mail: kerstin.wirkner@uni-leipzig.de

adult lifespan and age-specific groups to elucidate general and age-sensitive predictors of cognitive test scores.

Keywords Machine learning analyses · Age decades · Brain structure · Demographic · Health-related · Cognitive functions · Prediction

Introduction

In light of the accelerating demographic change, promoting cognitive health into old age is of pivotal importance not only for society and economy, but also for the healthcare system and standard of life [1]. To this end, the interest in identifying markers for age-related cognitive decline is growing, since early identification of cognitive decline is one key for developing successful strategies for the prevention of cognitive dysfunction [2–4].

Machine learning (ML) tools have emerged as the prevailing solution in the search of markers for cognitive functioning, given their capacity to identify patterns in a high-dimensional space and to move analyses to the individual level [5–7]. Their initial focus has been on brain imaging markers, i.e. magnetic resonance imaging (MRI) derived, to predict cognitive abilities. Here, the best prediction results have been achieved in younger adults (age range 22–37) and for general intelligence using functional brain metrics with about 20% of variance explained [8–10]. However, when it comes to multimodal neuroimaging data, e.g. structural brain measures, functional (FC) and structural connectivity (SC), as well as to samples including older adults up to 85 years, lower predictability, with less than 10% explained variance, has been reported for both global and domain-specific cognitive test performance [2, 11–13]. Hence, MRI-derived brain data alone seems to be insufficient for reliable prediction of cognitive functioning across the lifespan.

Prior to ML, other factors, such as sociodemographic, health-related, and lifestyle variables, have been identified as contributors to cognitive decline and dementia. These factors may provide valuable insights in the search for reliable markers of cognitive dysfunction [14–16]. Cardiovascular risk factors such as diabetes or smoking and sociodemographic factors such as educational attainment, socioeconomic status (SES), sex, and age have been shown to exert substantial influence on cognitive functioning and represent important input features to predict cognitive performance across the adult lifespan [16–23]. For instance, sociodemographic and lifestyle data, e.g. physical activity, were shown to successfully predict processing speed (age range 34–97; up to 43% variance explained) [24]. At the same time, individual cognitive test scores have been shown to be highly inter-related; thus, measuring one specific cognitive ability can serve as a predictor for cognitive performance in other domains, i.e. performance in a verbal fluency test successfully predicted test scores in executive function tests and vice versa ($N > 200$; age range 20–55) [25, 26]. Hence, different data types and information, such as MRI-derived brain data, demographic, health-related, lifestyle items, and cognitive performance itself, contribute to successfully predicting cognitive abilities.

The Lancet Commission on Dementia elaborated further on this topic by underscoring the significance of integrating multiple risk factors to attain optimal predictability. The report recently identified 14 modifiable risk factors that may together account for 45% of dementia cases worldwide [16]. The authors emphasized that those risk factors typically emerge at distinct phases of the lifespan, e.g. education during childhood and hypertension in middle age. These findings underscore the importance of conducting age-specific prediction analyses and systematically comparing the predictability of cognitive

performance across distinct age groups. However, this poses a challenge with respect to statistical power, as large sample sizes are required for this type of analysis [27]. To date, most ML studies have focused on predicting cognitive performance either in younger or older populations, using diverse input features, algorithms, and cognitive tests [2–4, 8, 9, 13]. This heterogeneity complicates the comparison of results across studies. Nevertheless, initial indications point towards differences in the predictability of cognitive performance between younger and older adults [11, 28–31]. For instance, performance in executive functions and episodic memory may be better predicted from functional imaging data in older (age > 50) compared to younger (age < 40) adults [11, 28]. In order to unravel the complexities inherent in the age-dependent prediction of cognitive performance, and to identify time windows that are susceptible to potential influences from factors such as health-related conditions (e.g. hypertension) on cognitive performance, large-scale and lifespan-encompassing studies are required.

The current study comprehensively investigated the predictability of cognitive test scores from MRI-derived structural brain data, health-related variables, demographics, and cognitive performance across the lifespan (age range 25–74) in a large cross-sectional sample ($N = 21{,}877$) from the population-based German National Cohort (NAKO Gesundheitsstudie, NAKO) [32] using ML. Particularly, the study set out to systematically compare ML outcomes across (1) age decades (total sample vs age decades), (2) distinct modalities (structural brain data, health, demographics, cognition), (3) approaches (unimodal vs multimodal), and (4) different cognitive tests (executive functions, language, motor speed, episodic and working memory). In this context, we hypothesized that (1) prediction levels may differ between age decades; (2) health, demographic, and cognitive performance data improve the prediction performance; (3) multimodal models combining structural brain, health-related, demographics, and cognitive performance data outperform single modalities in the prediction; and (4) cognitive functions that are more strongly affected during the aging process, e.g. episodic memory, are best predicted. Crucially, to ensure the generalizability and replicability of our results, extensive validation analyses were performed across the NAKO imaging sites and in an out-of-distribution external dataset (1000BRAINS [33]).

Methods

Participants

Data for the current analyses were derived from the large population-based German National Cohort (NAKO-Gesundheitsstudie, NAKO), which aims at gaining a better understanding of the causes of major chronic diseases, such as cardiovascular diseases, cancer, or diabetes, identifying risk factors and developing effective disease prevention strategies [34]. It collects data across 18 different study centres in Germany, with five dedicated MR imaging sites (Augsburg, Berlin, Essen, Mannheim, and Neubrandenburg), each equipped with the same MR scanner (3 Tesla, MAGNETOM Skyra, Siemens Healthineers, Erlangen, Germany) and using the same scanning protocol [35].

A subset of 29,862 participants underwent the baseline MRI examination [35] and met the age criteria of the current study (age range 25–74). This age range allowed that sufficient numbers of participants were found in each of the age decades required for subsequent ML analyses. Missing imaging-derived, demographic (i.e. educational level [participants in training were excluded], socioeconomic status, and employment status), health-related, and/or cognitive data led to the exclusion of 7985 participants from the initial sample. A final sample of 21,877 participants (44% females, $M_{age} = 48.9$, $SD_{age} = 11.4$) was used for further analyses. This total sample was stratified into five age decades (for an overview of sample characteristics of each decade, see Table 1). Approval on the study protocol of the NAKO cohort was granted by all responsible local ethics committees (https://nako.de). All participants provided written informed consent prior to inclusion. The study procedure complies with the Declaration of Helsinki.

Structural brain data

In the current study, the focus was on MRI-derived structural brain data, following research that highlights

Table 1 Demographic information of sample regarding age and educational level

Sample	*N* (%)	Mean age in years (SD)	Education in years (SD)
Total	21,877 (44% females)	48.9 (11.4)	15.6 (2.3)
Decade 1 (25–34 years)	3010 (41% females)	29.6 (2.6)	15.7 (2.1)
Decade 2 (35–44 years)	4205 (42% females)	40.5 (2.8)	15.8 (2.3)
Decade 3 (45–54 years)	7592 (46% females)	49.4 (2.8)	15.5 (2.3)
Decade 4 (55–64 years)	4997 (46% females)	59.3 (2.9)	15.4 (2.3)
Decade 5 (65–74 years)	2073 (44% females)	67.2 (1.9)	15.4 (2.5)

For age and education, standard deviation (SD) appears in parentheses

how differences in cognitive function may already be captured at the whole-brain level in brain summary statistics, such as cortical grey matter volume (GMV), subcortical GMV, white matter volume (WMV), and white matter lesion (WML) load [36–38]. Across all NAKO MRI sites, structural imaging data were acquired on a 3T MR scanner (Magnetom Skyra, Siemens Healthineers, Erlangen, Germany) equipped with a 70-cm bore [39]. For brain structural analyses, a 3D T1-weighted magnetization prepared rapid acquisition gradient-echo (MPRAGE) technique was employed using: voxel size = 1 × 1 × 1 mm^3, TR = 2300 ms, TE = 2.98 ms, flip angle = 9°. Additionally, a T2-weighted 2D fluid-attenuated inversion recovery (FLAIR) technique was used: voxel resolution = 0.9 × 0.9 × 4 mm^3, TR = 9000 ms, TE = 100 ms, flip angle = 150°.

Image preprocessing

T1-weighted 3D anatomical images were processed using the "recon-all" automated surface reconstruction pipeline of the FreeSurfer 7.1 Software package [40]. For a detailed description of the preprocessing steps, please refer to [40, 41] and the official documentation. After automated reconstruction, total cortical GMV, subcortical GMV, cerebellar GMV (left and right), total cerebral WMV, and estimated total intracranial volume (eTIV) were extracted using aseg.stats.

For the estimation of white matter hyperintensities, BIANCA, a well-established automated white matter hyperintensities segmentation algorithm, was used [42]. Before running the automated segmentation algorithm, data was preprocessed using the following steps: (1) production of brain extracted images using T2-FLAIR and T1-weighted data with CAT12 [43], (2) modality co-registration (T1 to FLAIR), and (3) normalization to template space (MNI152NLin6Asym) using ANTs [44]. After image preprocessing, manual segmentation of the WML was performed on FLAIR images of 120 participants from 1000BRAINS, a population-based aging cohort from Western Germany, and 64 participants from the NAKO (balanced in terms of age, sex, and cardiovascular factors) by a trained physicist [45]. BIANCA was then trained on the manually segmented FLAIR (reference base) data from 1000BRAINS and applied to NAKO data with the following parameters [42, 46]: spatial weighting (sw) = 8, no patch, no border for the non-lesion training points location (excluding 3 voxels close to the lesion's edge), fixed and unbalanced (FU) number of training points (lesion points = 5000, non-lesion points = 25,000 per participant), and lesion probability map threshold = 0.9. The accuracy of WML estimation was evaluated by computing the overlap between WML estimations from BIANCA and the manual masks for NAKO. The total WML volume (in mm^3) was estimated for each participant [42, 45].

Health-related data

In addition to the MRI assessments, participants also completed face-to-face interviews and touchscreen questionnaires. Furthermore, a study nurse-administered neuropsychological test battery and the participants underwent a series of medical examinations [32, 47]. In this context, a range of health-related data encompassing lifestyle behaviours as well as medical conditions were collected. For the current study,

ten different variables were selected as input data to the ML framework based on prior literature showing links to cognition: (1) body mass index, (2) smoking status (self-report; current/previous/non), (3) hypertension (based on blood pressure measurement; yes/no), (4) intake of antihypertensive medication (self-report; yes/no), (5) intake of cholesterol reducing medication (self-report; yes/no), (6) diabetes diagnosis (self-report; yes/no), (7) increased blood lipids (self-report; yes/no), (8) alcohol consumption (self-report; current/previous/never), (9) left hand grip strength, and (10) right hand grip strength (Fig. 1A; for a more detailed description, see Suppl. Table 1).

Demographic data

A series of demographic data were collected as part of the study protocol. For the purpose of the current analyses, variables of interest included (1) age (in years), (2) sex (male/female), (3) educational level (in years, ISCED-97 [48]), (4) socioeconomic status (International Socio-economic Index of Occupational Status [49], and (5) employment status (employed/unemployed/non-employed). These five variables were subsequently treated as the third input modality to ML. For detailed description, see Suppl. Table 1.

Cognitive performance

All participants took part in a neuropsychological examination, which covered different cognitive functions. In detail, the cognitive battery included (1) a semantic fluency test [number of correctly named animals within one minute], (2) a verbal episodic memory test [Immediate Recall: sum of words remembered from a word list (12 items) in two trials; Delayed Recall: sum of words remembered of a word list in delayed recall], (3) an executive function test [Stroop; time difference between naming ink colour of words with incongruent colour meaning and naming of colour of differently coloured boxes (Time T3–Time T2)], (4) a verbal working memory test [Digit Span; longest sequence of digits correctly repeated backwards], and (5) a motor speed test [Purdue Pegboard Test; number of correctly placed pin pairs in 30 s] [50] (for a more detailed description, see Suppl. Table 1).

Machine learning framework

A ML approach was chosen to investigate the predictability of cognitive test performance from brain structural, health-related, demographic, and cognitive data across different age decades. A schematic overview of the workflow can be found in Fig. 1. Input data to ML constituted (a) brain structural data: (1) GMV, (2) subcortical GMV, (3) right cerebellar GMV, (4) left cerebellar GMV, (5) WMV, (6) eTIV, and (7) WML (# of features = 7); (b) health-related data (# of features = 10); (c) demographic data (# of features = 5); and (d) cognitive data (# of features = 5; total # of input features = 26; Fig. 1A). Raw cognitive test scores were used as targets in the ML framework (total # of targets = 6; Fig. 1B). For each cognitive target, the remaining cognitive variables were used as input data with the only exception for immediate and delayed recall. Due to their high correlation, only one was included as a predictor [Immediate Recall] for the other cognitive variables. In case of predicting episodic memory immediate recall test scores, episodic memory delayed recall performance was excluded from the input data due to the risk of ML performance inflation and vice versa. ML performance estimations were obtained for all single modalities, for pairwise combinations, for three-way combinations and a four-way combination in the total sample and each decade separately (4 unimodal, 11 multimodal combinations; Fig. 1A). To obtain multimodal feature vectors, data was concatenated.

ML performance estimations were compared across four different algorithms, i.e. Ridge regression, linear support vector regression (linSVR), elastic net (EN) regression, and random forest (RF) regression, which are commonly employed in ML neuroimaging studies (Fig. 1C)[2]. In turn, ML model performance was evaluated using a repeated nested tenfold cross-validation (10 repeats) (Fig. 1D). Hyperparameter optimization was performed in the inner cross-validation to avoid data leakage (fivefold CV). Thereby, the following hyperparameter grids were searched: (i) regularization parameter C for linSVR (10^{-4} to 10^{1}, 10 steps, logarithmic scale), (ii) regularization

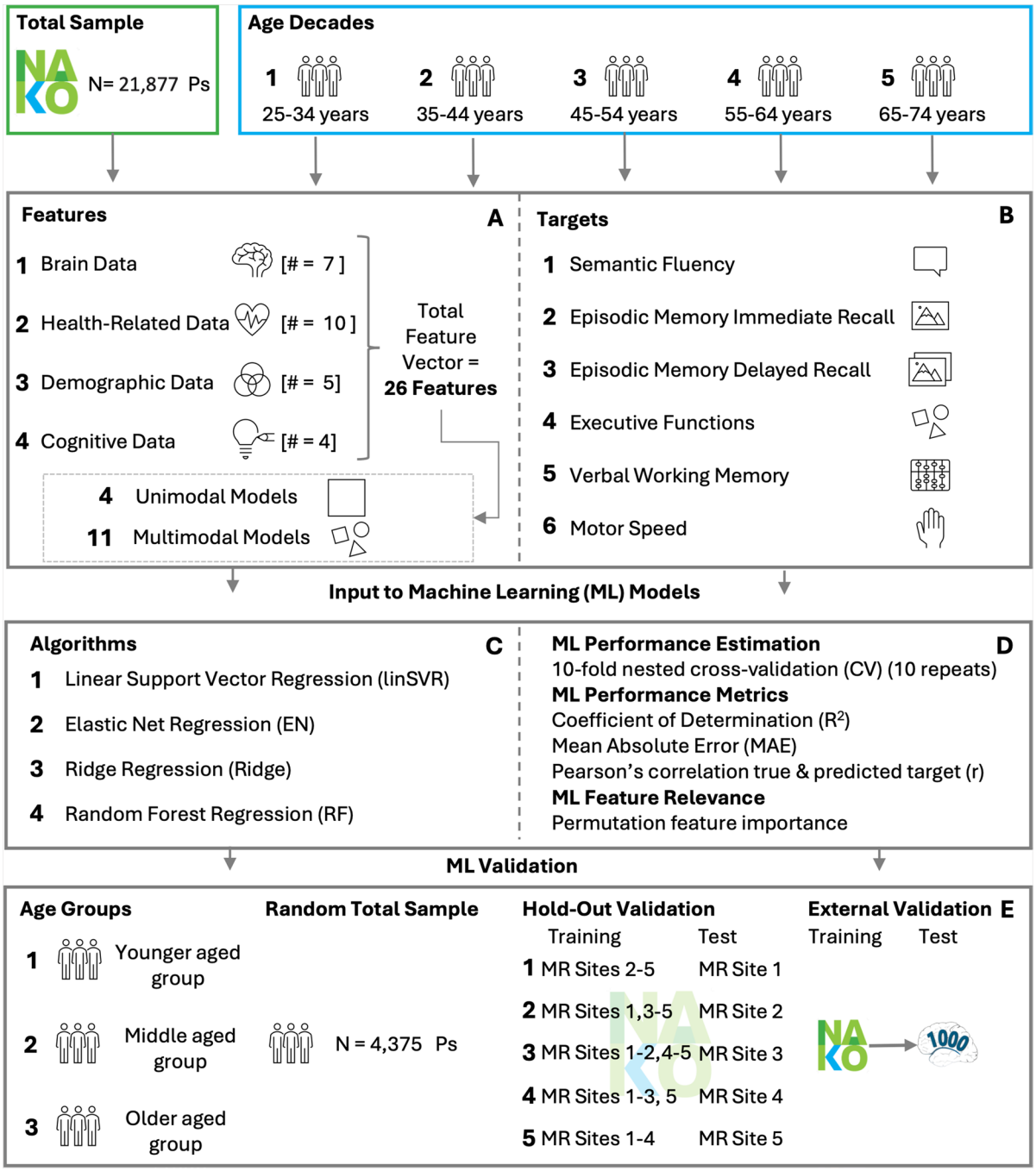


Fig. 1 Schematic illustration of the workflow. Sample description including the whole sample and age-specific groups (=Decades), **A**=included feature sets, **B**=targets to predict, **C**=algorithms used in the current study, **D**=settings for ML performance estimation as well as ML performance metrics, and **E**=ML validation analyses, including hold-out and external validation

parameter lambda λ for Ridge (10^{-3} to 10^{5}, 10 steps, logarithmic scale), (iii) regularization parameters lambda λ and alpha α for EN (λ 10^{-1} to 10^{2}, 10 steps, logarithmic scale; α 0.1 to 1, 10 steps), and (iv) number of trees and tree depth for RF (number of trees 10, 100, or 1000; tree depth 4, 6, 8, 10, 20, 40, none). For performance estimation, the mean absolute error (MAE), coefficient of determination (R^2), and the Pearson's correlation (r) between true and predicted targets were calculated (Fig. 1D). All ML analyses were performed using scikit-learn (version 0.22.1) in Python (https://scikit-learn.org/stable/index.html) [51].

Feature importance

To assess the contribution of each input feature to the prediction, permutation feature importances were calculated on the testing dataset, for all targets and algorithms in each sample. By using this method, the relationship between input feature and target is broken in order to evaluate the relevance of a particular feature for the model. For complexity reduction, we focused on models in which all features were combined to extract important features for prediction. For further analyses, permutation feature importances were averaged across algorithms for each cognitive variable and sample. For interpretation purposes, the top 25% of the most relevant features were then selected and plotted for the total sample and each decade in a bubble plot using plotly (https://plotly.com/python/).

ML validation analyses

Further analyses were performed to validate our ML results. Initially, prediction performance was assessed in larger age groups to investigate if trends would persist after joining decades and making the age groups more similar in size. We therefore divided the total sample into three larger age groups [younger (Decades 1 and 2, $N=7215$, $M_{age}=35.9$, $SD_{age}=6.0$) vs. middle (Decade 3, $N=7592$, $M_{age}=49.4$, $SD_{age}=2.8$) vs. older age group (Decades 4 and 5, $N=7070$, $M_{age}=61.7$, $SD_{age}=4.5$)]. Additionally, to assess if the larger participant numbers in the total sample may partly explain our prediction results, we performed further ML analyses in a random subsample ($N=4375$, 44% females, $M_{age}=48.9$, $SD_{age}=11.3$) drawn from the total sample that preserved the distribution across the different age decades, but more closely matched the sample size in each decade (=total sample size/5 decades). Furthermore, a hold-out and external validation was performed to investigate the generalizability and reliability of ML results across sites and different cohorts (Fig. 1E). To assess generalizability across sites, a hold-out validation approach was chosen and the total sample was divided by MR scanning site [Site 1: $N=4670$, $M_{age}=49.2$, $SD_{age}=11.5$; Site 2: $N=4013$, $M_{age}=4$ 8.8, $SD_{age}=11.4$; Site 3: $N=4332$, $M_{age}=48.5$, $SD_{age}=11.5$; Site 4: $N=3759$, $M_{age}=47.9$, $SD_{age}=11.2$; Site 5: $N=5103$, $M_{age}=50.0$, $SD_{age}=11.1$]. Each MR scanning site served as held-out set once, while the remaining sites were used as internal training data (Fig. 1E). Best parameters from the internal training were applied to the held-out test set. For generalizability assessment across cohorts, an external validation setup was used. In this context, data from the population-based 1000BRAINS study ($N=838$, 46% females, age-range 25–74 years, $M_{age}=59.3$, $SD_{age}=11.6$) aimed at examining inter-individual variability in the aging brain served as an external test set [33]. In terms of ML, models were trained on the total NAKO sample for each cognitive target and tested on 1000BRAINS (Fig. 1E). For a more detailed information on the 1000BRAINS cohort, the imaging parameters, input and target data, please refer to the Suppl. Methods.

Model comparison

For better interpretability of results, comparisons to estimations of a reference model (dummy regressor; prediction of mean target of training set) were performed [2, 13]. Particularly, the percentage of folds, in which the actual models outperformed the reference model, was calculated. Furthermore, multimodal bonuses were calculated to evaluate whether multimodal models outperformed single modalities [2, 13]. In this context, B_{best} was calculated for each multimodal combination, which captures the difference in prediction performance of the multimodal combination and the best single modality.

Results

Prediction outcomes were compared between the total sample ($N=21,877$) and five age decades: Decade 1:

25–34 years; Decade 2: 35–44 years; Decade 3: 45–54 years; Decade 4: 55–64 years; and Decade 5: 65–74 years (Fig. 1). A ML model was defined using different algorithms (Fig. 1C) and settings (Fig. 1D) to predict cognitive test scores (Fig. 1B, targets) from different input features (Fig. 1A). Finally, results were extensively validated across sub-samples with different characteristics (e.g. sample size, geographical differences) as well as an external sample (Fig. 1E). For an overview of the workflow, see Fig. 1.

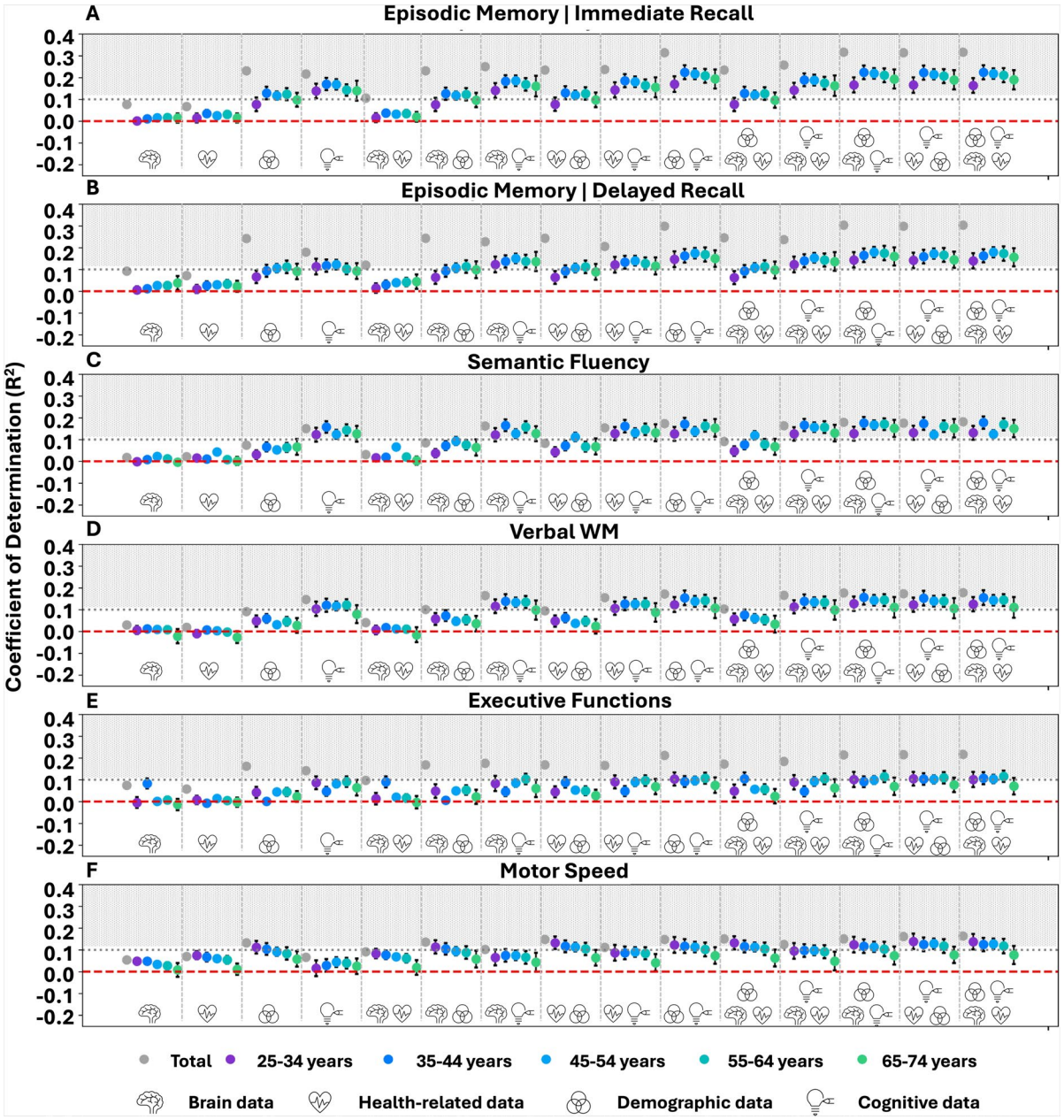
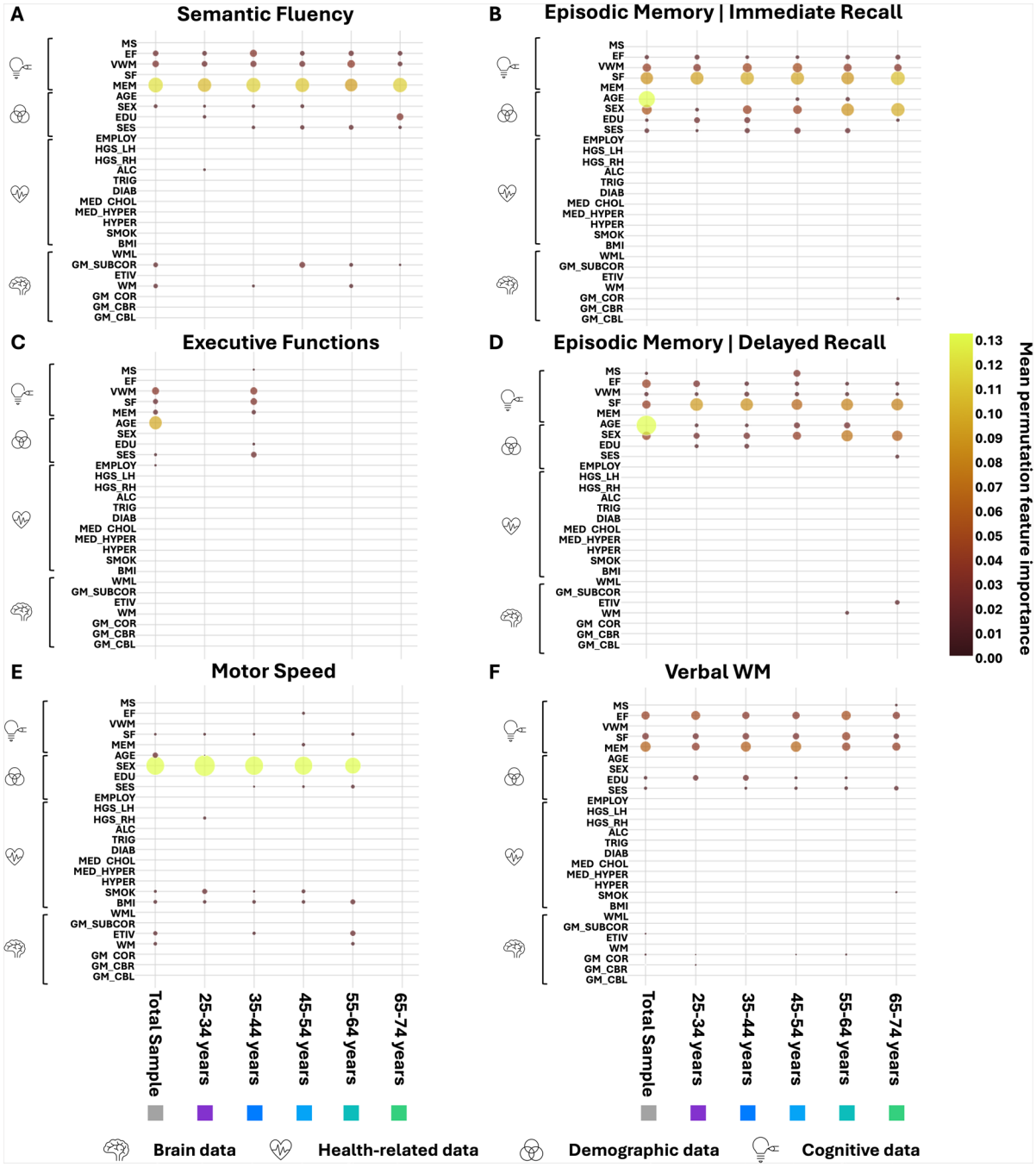


Fig. 2 Prediction performance for different cognitive test scores [**A**–**F**] based on structural brain data, health, demographics and cognition in the total sample and across age decades. Mean Coefficient of Determination (R^2) across algorithms. Error bars represent mean standard deviation (SD) across folds. Grey shaded areas indicate ML results $R^2 > 0.1$. Verbal WM = verbal working memory



Prediction results for cognitive tests across age decades

Across all cognitive tests, input features, and algorithms used, best prediction performance was found in the total sample. Here, prediction levels, as measured by the coefficient of determination (R^2), reached up to 0.32 and explained up to 10% more variance in the total sample than in the different age decades (total: R^2 range 0.02–0.32, in 99.8–100% of folds $R^2 >$ dummy regressor, Decades 1–5: R^2 range −0.03–0.22, in 46–100% of folds

◂**Fig. 3** Mean permutation feature importances shown for each cognitive variable [**A**–**F**] and each sample. Feature importances only displayed for models exceeding a prediction performance $R^2 > 0.1$ in the four-way combination. Larger circles mirror greater feature importance. GM_CBL = grey matter volume of the left cerebellum; GM_CBR = grey matter volume of the right cerebellum; WM = white matter volume; ETIV = estimated total intracranial volume; GM_COR = total cortical grey matter volume; GM_SUBCOR = subcortical grey matter volume; WML, =white matter lesion load; BMI = body mass index; SMOK = smoking status; HYPER = hypertension diagnosis; MED_HYPER = antihypertensive medication; MED_CHOL = cholesterol reducing medication; DIAB = diabetes diagnosis; TRIG = heightened blood lipids & triglycerides; ALC = alcohol; HGS_RH = right hand grip strength; HGS_LH = left hand grip strength; EMPLOY = employment status; SES = socioeconomic status; EDU = educational level; MS = motor speed; SF = semantic fluency; MEM = episodic memory intermediate recall; EF = executive functions; VWM/Verbal WM = verbal working memory

$R^2 >$ dummy regressor; Fig. 2, Suppl. Tables 2–3). This effect was found particularly for executive functions, immediate and delayed episodic memory (total: R^2 range 0.06–0.32, in 99.8–100% of folds $R^2 >$ dummy regressor; Decades 1–5: R^2 range: −0.01–0.22, in 46–100% of folds $R^2 >$ dummy regressor).

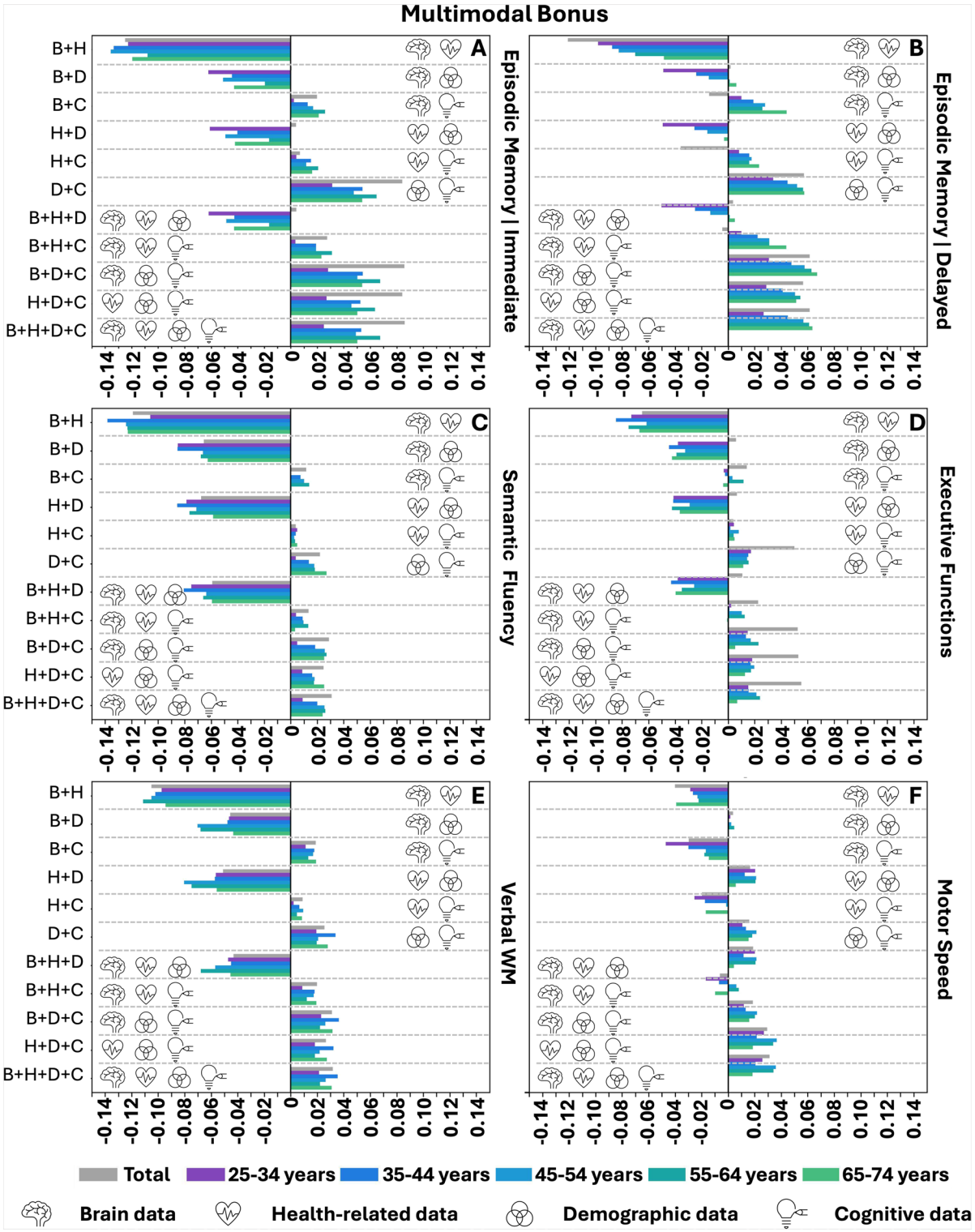
In general, immediate episodic memory was best predicted (R^2 range 0–0.32, in 67–100% of folds $R^2 >$ dummy regressor; Fig. 2, Suppl. Tables 2–3), while motor speed and verbal working memory showed the lowest predictability (R^2 range −0.03–0.18, in 52–100% of folds $R^2 >$ dummy regressor; Fig. 2, Suppl. Tables 2–3). Predictability of the other cognitive tests (executive functions, verbal fluency, and delayed episodic memory) was intermediate, falling between low and high levels of R^2 (R^2 range −0.01–0.3, in 46–100% of folds $R^2 >$ dummy regressor; Fig. 2, Suppl. Tables 2–3). When comparing prediction performance across age decades for specific cognitive test scores, differences in predictability emerged. Lower prediction performance was found for executive functions and motor speed in the oldest age group compared to the other age decades (Decade 5; R^2 range −0.01–0.08, in 46–100% of folds $R^2 >$ dummy regressor; Decades 1–4: R^2 range 0–0.14, in 51–100% of folds $R^2 >$ dummy regressor). ML outcomes for delayed episodic memory performance and semantic fluency, in turn, were found to be lower in the youngest participants (Decade 1: R^2 range 0–0.14, in 59–100% of folds $R^2 >$ dummy regressor) compared to all others (Decades 2–5: R^2 range 0–0.18, in 60–100% of folds $R^2 >$ dummy regressor).

In conclusion, the best predictability was achieved in the total sample compared to the individual age decades. Cognitive tests appeared to be differentially predictable, with the most pronounced differences found between immediate episodic memory and motor speed. Closer examination of the individual cognitive tests emphasized the occurrence of predictability differences in the youngest (25–34 years) and oldest (65–74 years) age groups.

Predictability differences between input features

Predictability was further compared between analyses including either MRI-derived brain structural, health-related, demographic, or cognitive performance data as input features (Fig. 1A). Results revealed highest predictability from demographic and cognitive data (R^2 range 0.02–0.24, in 82–100% of folds $R^2 >$ dummy regressor) outperforming brain structural and health-related data (R^2 range −0.03–0.09, in 46–100% of folds $R^2 >$ dummy regressor; Fig. 2, Suppl. Tables 2–3). Despite the fact that brain structure and health-related data generally did not exceed levels of predictability of demographic and cognitive data, they appeared to be marginally helpful for the prediction of delayed episodic memory (R^2 range 0.01–0.09, in 72–100% of folds $R^2 >$ dummy regressor; Fig. 2, Suppl. Tables 2–3).

Following up on this, we next zoomed into the feature modalities and identified single feature importances in the prediction of cognitive test scores. In general, the most influential features for the prediction of cognitive test scores were identified as those derived from the demographic and cognitive data (Suppl. Table 5). Nevertheless, slight differences in the composition of relevant features emerged for different cognitive tests (Fig. 3). For an overview of feature importances across samples, tests, and features, see Suppl. Table 5. In summary, age was found to be most relevant in the prediction of executive functions, immediate and delayed episodic memory, but only in the total sample. For all other decades and cognitive tests, age was found to be of less importance. Across decades, episodic memory performance appeared to be particularly important in the prediction of semantic fluency performance and vice versa. The combination of sex, body mass index, and smoking status was


Multimodal Bonus
A
B
C
D
E
F
Episodic Memory | Immediate
Episodic Memory | Delayed
Semantic Fluency
Executive Functions
Verbal WM
Motor Speed
B+H
B+D
B+C
H+D
H+C
D+C
B+H+D
B+H+C
B+D+C
H+D+C
B+H+D+C
-0.14
-0.12
-0.10
-0.08
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Total
25-34 years
35-44 years
45-54 years
55-64 years
65-74 years
Brain data
Health-related data
Demographic data
Cognitive data

◂**Fig. 4** Multimodal bonuses for each cognitive test score [**A**–**F**] across the total sample and different age decades. Positive bonus indicates multimodal combination outperforms best single modality; negative bonus indicates multimodal combination performs worse than best single modality. B = brain; C = cognition; D = demographics; H = health; Verbal WM = verbal working memory

found exclusively among the top ranked features for motor speed. Educational level and socioeconomic status were found to be relevant particularly for performance in semantic fluency, verbal working memory, and episodic memory. Cortical and subcortical grey matter volume appeared to be partially important for predicting performance in semantic fluency, motor speed, and verbal working memory. Hence, the results indicate considerable heterogeneity in the factors predicting cognitive performance across tests and age decades, with multiple input features contributing to the prediction of a single cognitive test score.

Moreover, we addressed whether the combination of different input feature modalities (i.e. brain structural, health-related, demographic, or cognitive data) may be beneficial for prediction of cognitive test scores (two-way, three-way and four-way combinations) (Fig. 1A). Across cognitive tests, samples, and algorithms, there was a general tendency for multimodal combinations (R^2 range −0.01–0.32, in 54–100% of folds $R^2 >$ dummy regressor) outperforming single modalities (R^2 range −0.03–0.24, in 46–100% of folds $R^2 >$ dummy regressor; Figs. 2 and 4, Suppl. Tables 2–4). The strongest effect was observed for the prediction of immediate and delayed episodic memory based on a combination of demographic and cognitive data as input features (unimodal: R^2 range 0–0.24, in 67–100% of folds $R^2 >$ dummy regressor; multimodal: R^2 range 0.02–0.32, in 81–100% of folds $R^2 >$ dummy regressor; Figs. 2 and 4, Suppl. Tables 2–4). This led to the highest prediction increments of up to 9% gain in R^2 compared to the best single modality (i.e. cognitive data). In contrast, the combination of brain structure and health-related data performed markedly worse (up to 14% lower R^2) than the best single modality across analytic choices (Fig. 4, Suppl. Table 5).

Overall, our results indicate that demographic and cognitive data clearly outperformed brain structural and health-related data across samples, tests, and algorithms with an additional benefit when combining demographic and cognitive data as input features.

ML validation analyses

Importantly, to ensure robustness and generalizability of the results, the current study performed several validation analyses across different samples with different characteristics (e.g. sample size, geographical differences, educational differences).

We first discerned whether the established ML results here are independent of sample size. To do so, prediction performance of all cognitive tests was examined across larger age groups (younger [Decades 1 and 2] vs. middle [Decade 3] vs. older [Decades 4 and 5]) as well as a randomly drawn subsample from the total sample ($N = 4375$) (Fig. 1E). Results revealed similar ML outcomes as in the main analyses (younger: R^2 range 0.01–0.22, in 71–100% of folds $R^2 >$ dummy regressor; middle: R^2 range 0–0.22, in 70–100% of folds $R^2 >$ dummy regressor; older: R^2 range −0.01–0.23, in 68–100% of folds $R^2 >$ dummy regressor; random total: R^2 range 0.01–0.31, in 68–100% of folds $R^2 >$ dummy regressor; Suppl. Tables 6–11, Suppl. Fig. 1-2). Differences across targets, modalities, approaches, and the lifespan were mostly preserved, providing sustenance to the overall trends in the main analyses. Results also supported the notion that the performance gain of using the total sample instead of age-specific ones was not merely driven by sample size.

Afterwards, we performed a hold-out and external validation to assess the out-of-distribution generalization ability. Since the NAKO was set up as a multicentre cohort study, the hold-out validation analysis examined whether results generalize well across the five different MRI study sites (Fig. 1E). Results confirmed a good generalizability across MRI sites and mirrored patterns found in the main analyses (main analyses: R^2 range across test 0.02–0.32; hold-out validation: R^2 range across tests −0.03–0.33; Suppl. Tables 12–13, Suppl. Fig. 3-4).

In the external validation analyses, models were trained on NAKO data and tested out-of-distribution using data from the population-based 1000BRAINS study, which aims at gaining a greater understanding of the sources for the high inter-individual variability in brain aging [33]. In comparison to the

hold-out validation, prediction accuracies markedly decreased in the external validation. Nevertheless, the general patterns observed in the main analyses were replicated. As such, immediate episodic memory (NAKO: R^2 range 0.06–0.32; 1000BRAINS: R^2 range < -1 to 0.26; Suppl. Table 14, Suppl. Fig. 3-4) yielded the best external prediction performance compared to other cognitive tests (NAKO: R^2 range 0.01–0.3; 1000BRAINS: $R^2 < 0.07$; Suppl. Table 14, Suppl. Fig. 5). Results for executive functions and delayed episodic memory could not be generalized to 1000BRAINS (NAKO: R^2 range 0.05–0.3; 1000BRAINS: $R^2 < 0$; Suppl. Table 14, Suppl. Fig. 5). In terms of modalities, particularly demographic data seemed to be informative for prediction (Suppl. Table 14, Suppl. Fig. 5). Thus, results supported the general trends from the main analyses albeit lower prediction performance (particularly external validation) and emphasized that results may generalize to other cohorts under certain circumstances (i.e. cognitive tests, modality).

Discussion

The current study aimed at a foundational investigation into the predictability of cognitive test scores based on MRI-derived brain structural measures together with health-related, demographic and cognitive data in a large cross-sectional sample ($N = 21{,}877$) from the NAKO. Overall, results showed mixed prediction outcomes across the lifespan (age range 25–74), with differences across age groups, modalities, approaches, and cognitive tests. Specifically, results demonstrated (1) higher predictability in the total sample compared to single age decades, (2) higher predictive power of demographic and cognitive data with up to 17% more variance explained, compared to brain structural and health-related data, (3) a superadditive effect with up to 9% increase in R^2 when joining demographic and cognitive data indicating a multimodal benefit, and finally (4) better predictability for episodic memory performance compared to other cognitive test scores. External validation results underscore the prediction advantage of demographic data and the higher predictability of episodic memory. Overall, results emphasize that analyzing the entire adult lifespan in conjunction with age-specific subgroups may provide important insight into previously unseen trends in the data.

Predictability differences between the total sample and age decades

In our systematic ML assessment, higher prediction performance was achieved in the total sample (R^2 up to 0.32) compared to different age decades (R^2 up to 0.22). This was consistent across different cognitive tests, modalities, approaches, and algorithms, independent of the sample size. Based on prior findings and various (cognitive) changes that individuals undergo over the lifespan, we assumed that prediction levels for cognitive functioning would differ in an age-dependent manner [11, 28–31]. A potential explanation for the discrepancy in the current results could be that the effect of aging from young to old was stronger than the influence of all other risk factors. This was most evident for episodic memory and executive functions, which typically show the strongest age effects across cognitive domains [52–54]. Important to note, sample size did not seem to explain the differences in predictability, as shown by the validation analyses with varying sample sizes for the different predictions reflecting the whole age range in a smaller random sample. Rather, differences in the homogeneity of the total sample versus single age decades might explain these differences. Both age and cognitive test scores vary substantially across the lifespan, but to a considerably lower extent when the age span is reduced. Thus, more homogenous study participants, with respect to the variables of interest, i.e. input and targets, limit the predictive power for explaining differences in cognition due to the reduced variability [55]. Thus, our results indicate that including the whole adult age span contributes highly relevant and more variable information for the prediction of cognitive abilities, particularly for the prediction of episodic memory performance. The factor age, in contrast, is less influential when focusing on specific age groups.

Following from this, the question arises how the predictability of cognition may vary between age decades. One main result of our data is that no differences between age decades were observed, which contradicts published prior results which reported predictability differences between younger and older groups [11, 28–31]. The majority of studies reporting predictability differences for specific cognitive test scores (e.g.

memory and executive functions) between age groups have used MRI-derived imaging data, particularly FC information and sample sizes $N < 500$ [11, 28–31]. In contrast, the current study investigated predictability differences in a sample of $N = 21{,}877$, with each age group containing information on $N > 2000$ based on MRI-derived brain structural metrics together with health-related, demographic, and cognitive data. Thus, a potential explanation for the disparity in results could be a sample size difference. Prior research suggests that larger sample sizes may allow ML outcomes to reach estimates closer to the true level of predictability [5, 27, 56]. Thus, it might have been the case that the large sample sizes in each age decade allowed us to better capture the actual predictability levels. In turn, it appeared that age decades did not differ considerably in their predictability once the global age effect has been accounted for in the total sample.

Despite the general trend, the youngest and oldest age groups were found to differ from other decades in the predictability of specific cognitive test scores. Predictability for executive functions and motor speed was lower in the oldest group (65–74 years), while predictability of immediate episodic memory and semantic fluency was lower for the youngest group (25–34 years). One potential explanation for these findings could be related to cohort effects. It is important to keep in mind that the data assessed here come from a cross-sectional study design. Consequently, it is possible that feature importances diverge most prominently in the youngest and oldest age groups. For example, sex was shown to be highly important for the prediction of motor speed in all decades but the oldest. A similar picture emerged for episodic memory and the feature of sex in the youngest group. Instead, it might be that other factors are of relevance for predicting cognitive performance in these specific age groups. These may, for instance, include sleep quality and depressive symptomatology for motor performance [57] and genetic data for executive functions [58] in older adults, hormonal and physical activity data for verbal fluency [25, 59], and personality factors for episodic memory performance [60, 61] in younger adults. Hence, it may be argued that age-specific particularities emerge, which may be grounded in shifts in the relevance of particular input features for prediction across the lifespan. As such, current results support the notion of different factors being relevant at different times throughout an individual's life [16]. The investigation of such additional factors may not only enhance individual-level predictability and foster the development of precision medicine through tailored intervention programs, but also inform public health strategies aimed at preserving cognitive performance across the lifespan.

Predictability differences across input modalities

The current results emphasize that demographic and cognitive data clearly outperform brain and health-related data. Our findings are in line with an ever-growing amount of studies showing that particularly demographic data may be highly predictive of cognitive performance [2, 3, 13, 20, 62, 63]. Beyond demographic variables, other cognitive variables were also found to carry fundamental information for the prediction of specific cognitive test scores, most likely due to their interdependencies as hypothesized. For instance, other cognitive test scores yielded the highest prediction performance among single modalities for specific cognitive targets, i.e. verbal memory and semantic fluency (up to 7% more explained variance compared to other modalities). In both cases, particularly, episodic memory performance appeared to be highly relevant for prediction. One potential explanation for these effects might lie in the underlying neural constructs, i.e. the involvement of the left frontal lobe in all three cognitive functions [64]. As has been demonstrated in previous studies focusing on particular cognitive functions, as well as in longitudinal studies and neurodegenerative diseases, there is a close interaction between cognitive functions [25, 26, 65, 66]. For example, executive functions have been found to be predictive of semantic fluency, and vice versa, baseline episodic memory to explain an additional ~ 20% of variance in memory decline on top of age and sex, and subjective cognitive decline to be the most important feature for the future conversion to mild cognitive impairment [25, 26, 65, 66]. Thus, it could be argued that different cognitive functions are highly informative of specific cognitive test performance across age groups, potentially due to their contribution to the same general ability [67]. In contrast to this, prediction performance from brain structural data was found to be rather limited not exceeding prediction levels of $R^2 = 0.09$.

The current results therewith support prior findings showing a rather weak link between brain structural data and cognitive performance at the individual level across the lifespan [2, 5, 30, 68]. At the same time, they are in contrast to studies that have demonstrated more favourable predictive levels based on different types of imaging data, i.e. FC and SC data, in samples of younger and older adults [8, 10, 69]. Reasons for these discrepancies in results may be manifold, e.g. differences in sample sizes, cognitive tests examined, and brain metrics used, and thus require further investigation. Overall, recent seminal studies have suggested that effect sizes for brain-behaviour relationships may be rather small and that even with larger sample sizes, prediction accuracies for brain-behaviour relationships remain modest [27, 56]. As such, the brain-cognition relationships may be highly complex across the lifespan, which should be explored further [5].

In terms of health-related data, prediction performance was found to be surprisingly low for different cognitive test scores with the exception of motor speed. While both prior univariate and multivariate studies indicated a strong association between health-related factors and cognitive performance [16, 23, 24], currently chosen health-related information did not seem to be predictive of cognitive test scores. One potential explanation for this could be that health-related behaviour may be strongly related to demographic data, found to be highly predictive in the current study [22, 70]. For example, research has shown higher SES and educational attainment to be related to a more health-promoting lifestyle across the lifespan [22, 70]. Hence, health-related information might have not added extra information to the prediction in the current study. This may be further aggravated by the fact that at the individual level complex relationships between health-related factors and cognition may challenge the identification of clear patterns in the data. As such, health-related data in the current study might have not added additional information to the prediction contrary to prior findings emphasizing their association with cognition. Overall, it may be argued that cognitive test scores may be predicted well from demographic and cognitive data, but only questionable benefits for prediction may arise for including brain structural and health-related information.

Beyond the examination of the predictive power of single modalities, we also focused on the investigation of the predictive potential of multimodal combinations. Present findings highlighted that multimodal combinations may outperform single modalities in predicting different cognitive tests in the total sample and across age decades. Demographic and cognitive data that already led to promising results in the unimodal setup appeared to work well together with a predictability increase of up to 9% in R^2 compared to the best single modality. Thus, they appeared to be superadditive—in their effect on prediction performance extending prior findings of a multimodal benefit to new combinations of data types [13, 63]. Nevertheless, it should be stressed that this multimodal benefit was found only for specific combinations. Multimodal combinations including brain and health-related data did not seem to boost prediction levels, which is in line with recent findings [71].

Limitations and methodological considerations

The current study focused on the prediction of cognitive performance from different input modalities across the lifespan. While overall predictability is comparable to other studies, it is still in its absolute value not reaching a level allowing the deployment across studies to reliably predict cognition across the lifespan [72]. In order to make progress towards the ultimate goal of accurately predicting future cognitive abilities, it is important to emphasize that current predictors only capture certain aspects of cognitive functioning throughout adulthood. However, cognitive performance should be regarded as the result of a complex interplay between various biological, psychological, and environmental factors with linear and non-linear dynamics over time that cannot be adequately captured by classical machine learning algorithms. It should also be noted that the included variables may not have provided sufficient detail to extract meaningful associations between predictors and targets. Specifically, brain metrics were included using global measures of grey and white matter. More fine-grained, regional brain metrics (e.g. those defined using cytoarchitectonic parcellations of the Julich Brain Atlas [25] or functional parcellations [Schaefer et al., 2017]) have previously shown that brain-cognition relationships are regionally specific

rather than global. Therefore, future studies should investigate additional variables and vary the granularity of measurements. In addition to that, it should be noted, however, that although distinct variables have been included as features to predict cognitive performance, i.e. paper–pencil based cognitive tests, blood parameters, MRI-derived brain metrics, or self-reported demographic data, these variables have been shown to be dependent on each other. For instance, blood pressure might be related to WML, cognitive performance might be (partially) associated with sex. While the different ML models applied in the current study can handle correlated input features (to varying degrees), multicollinearity may affect interpretability of the results, i.e. by affecting feature importances. Thus, we cannot rule out the possibility that the ML models did not uniquely attribute predictive relevance to any single variable, as the shared signal may be distributed arbitrarily across correlated features. Additionally, it should be kept in mind that while the current analyses were performed on a very large sample size, only a limited number of neuropsychological tests were investigated. To obtain a more fine-grained picture of predictability differences, it appears warranted to investigate effects in a larger neuropsychological battery and replicate findings in other cohorts. Similarly, it should be accentuated that the current study only looked at cross-sectional data; i.e. predictors and cognitive outcomes were assessed at the same time point, which identifies associations rather than longitudinal risk trajectories, thereby limiting any conclusion on the true prognostic value. With the ultimate goal of a prospective marker in mind, it becomes necessary to also investigate predictive power in a longitudinal setting, i.e. following cognitive trajectories over years. With more and more large cohorts collecting data over time, this endeavour becomes more feasible and holds the promise of further advancing the field [7].

Conclusions

The current study investigated the predictive power of MRI-derived brain structural, health-related, demographic, and cognitive data for cognitive test performance in a very large sample of adult participants from the NAKO. Current results demonstrate mixed prediction performance (R^2 range = −0.02–0.32) of cognitive test scores across the lifespan. Predictability differences emerged across samples, cognitive tests, modalities, and approaches. Closing the loop on the hypotheses, the current study shows (1) higher, not lower, predictability in the total sample compared to different age decades, (2) higher predictive power of demographic and cognitive data compared to MRI-derived brain structural and health-related data, (3) a multimodal benefit based on joining information from the top performing single modalities (demographic and cognitive data seem to be superadditive), and (4) superior predictability for episodic memory performance compared to other cognitive test scores. Overall, our results emphasize the importance of the selected demographic and cognitive data. Furthermore, they underscore the limitations of the chosen brain and health-related variables. The investigation of the whole age span alongside specific age decades might serve as a potential research area in future prediction studies. This may allow gaining insights into previously undiscovered patterns in the data and help disentangle relevant factors determining cognition across the lifespan.

Acknowledgements We thank all participants who took part in the NAKO study and the staff of this research initiative. We thank the Heinz Nixdorf Foundation (Germany) for the generous support of the Heinz Nixdorf Study. We also thank the scientists and the study staff of the Heinz Nixdorf Recall Study and 1000BRAINS. This research was supported by the Joint Lab "Supercomputing and Modeling for the Human Brain". We gratefully acknowledge the computing time granted through JARA-HPC on the supercomputer JURECA at Forschungszentrum Jülich.

Author contribution CMH wrote the manuscript. CMH and CJ provided the general idea of the study, which was further refined together with SC. FB, KB, PB, JAD, AF, KHG, MH, JK, TK, CJKT, LK, TK, ML, TN, AP, TP, OR, SR, CS, MBS, MOW, and KW were involved in data collection and /or contributed to the study design of the NAKO. NB, CJ, and SC were involved in data collection for 1000BRAINS. PD and TM revised the pipeline to process the image data. CMH calculated all analyses and interpreted the data with advice from CJ, NB, and SC. CJ gave advise towards the statistical methods used and interpretation. All authors critically evaluated the manuscript. CJ supervised the entire study.

Funding Open Access funding enabled and organized by Projekt DEAL. This project was conducted with data (NAKO-746) from the German National Cohort (NAKO Gesundheitsstudie, NAKO) (www.nako.de). The NAKO is funded by the Federal Ministry of Research, Technology and Space (BMFTR) [project funding reference numbers: 01ER1301A/B/C, 01ER1511D, 01ER1801A/B/C/D and 01ER2301A/B/C], federal states of Germany and the Helmholtz Association,

the participating universities, and the institutes of the Leibniz Association. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101147319 (EBRAINS 2.0 Project; SC).

Data Availability The data that support the findings of this study are part of the German National Cohort (NAKO Gesundheitsstudie). NAKO data are subject to the EU/EEA General Data Protection Regulation and to the NAKO Terms of Use. Qualified researchers affiliated to EU or EEA institutions may apply for access through the NAKO TransferHub (https://transfer.nako.de). Applications are evaluated by the NAKO Use & Access Committee; successful applicants must sign a Data-Use Agreement and cover associated handling fees. This study used data from the 1000BRAINS study, a population-based cohort (Caspers et al. 2014). Access to raw neuroimaging and phenotypic data is restricted due to ethical and data protection regulations. Data is available to qualified researchers upon reasonable request to the 1000BRAINS study organizers.

Declarations

Competing interests The authors declare no competing interests.

References

1. United Nations, Department of Economic and Social Affairs, Population Division. World population prospects 2019: highlights (ST/ESA/SER.A/423). 2019.
2. Krämer C, et al. Prediction of cognitive performance differences in older age from multimodal neuroimaging data. GeroScience. 2024;46:283–308.
3. Vieira BH, et al. Predicting future cognitive decline from non-brain and multimodal brain imaging data in healthy and pathological aging. Neurobiol Aging. 2022;118:55–65.
4. Kwak S, Kim H, Kim H, Youm Y, Chey J. Distributed functional connectivity predicts neuropsychological test performance among older adults. Hum Brain Mapp. 2021;42:6495–507.
5. Genon S, Eickhoff SB, Kharabian S. Linking interindividual variability in brain structure to behaviour. Nat Rev Neurosci. 2022;23:307–18.
6. Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci Biobehav Rev. 2012;36:1140–52.
7. Wu J, Li J, Eickhoff SB, Scheinost D, Genon S. The challenges and prospects of brain-based prediction of behaviour. Nat Hum Behav. 2023;7:1255–64.
8. Dhamala E, Jamison KW, Jaywant A, Dennis S, Kuceyeski A. Distinct functional and structural connections predict crystallised and fluid cognition in healthy adults. Human Brain Mapping. 2021;25420. https://doi.org/10.1002/hbm.25420.
9. Dubois J, Galdi P, Paul LK, Adolphs R. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philos Trans R Soc Lond B Biol Sci. 2018;373:20170284.
10. Ooi LQR, et al. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI; 2022. http://biorxiv.org/lookup/doi/10.1101/2022.03.08.483564.
11. Heckner MK, et al. Predicting executive functioning from functional brain connectivity: network specificity and age effects. Cereb Cortex. 2023;11:6495–507.
12. Omidvarnia A, et al. Individual characteristics outperform resting-state fMRI for the prediction of behavioral phenotypes. Commun Biol. 2024;7:771.
13. Rasero J, Sentis AI, Yeh F-C, Verstynen T. Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. PLoS Comput Biol. 2021;17:e1008347.
14. Jaarsma E, et al. Modifiable risk factors for accelerated decline in processing speed: results from three Dutch population cohorts. J Prev Alzheimers Dis. 2024;11:108–16.
15. Jin Y, Liang J, Hong C, Liang R, Luo Y. Cardiometabolic multimorbidity, lifestyle behaviours, and cognitive function: a multicohort study. Lancet Healthy Longev. 2023;4:e265-73.
16. Livingston G, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing commission. Lancet. 2024;404:572–628.
17. Canavan M, O'Donnell MJ. Hypertension and cognitive impairment: a review of mechanisms and key concepts. Front Neurol. 2022;13:821135.
18. Dove A, et al. The impact of diabetes on cognitive impairment and its progression to dementia. Alzheimers Dement. 2021;17:1769–78.
19. Kunutsor SK, Isiozor NM, Voutilainen A, Laukkanen JA. Handgrip strength and risk of cognitive outcomes: new prospective study and meta-analysis of 16 observational cohort studies. Geroscience. 2022;44:2007–24.
20. Schrempft S, et al. Life-course socioeconomic conditions and cognitive performance in older adults: a cross-cohort comparison. Aging Ment Health. 2023;27:745–54.
21. Song R, et al. Associations between cardiovascular risk, structural brain changes, and cognitive decline. J Am Coll Cardiol. 2020;75:2525–34.
22. Wang A-Y, et al. Socioeconomic status and risks of cognitive impairment and dementia: a systematic review and

meta-analysis of 39 prospective studies. J Prev Alzheimers Dis. 2022. https://doi.org/10.14283/jpad.2022.81.
23. Zhao L, et al. Identifying a group of factors predicting cognitive impairment among older adults. PLoS One. 2024;19:e0301979.
24. Poudel GR, et al. Machine learning for prediction of cognitive health in adults using sociodemographic, neighbourhood environmental, and lifestyle factors. Int J Environ Res Public Health. 2022;19:10977.
25. Amunts J, Camilleri JA, Eickhoff SB, Heim S, Weis S. Executive functions predict verbal fluency scores in healthy participants. Sci Rep. 2020;10:11141.
26. Amunts J, et al. Comprehensive verbal fluency features predict executive function performance. Sci Rep. 2021;11:6929.
27. Marek S, et al. Reproducible brain-wide association studies require thousands of individuals. Nature. 2022;603:654–60.
28. Pläschke RN, et al. Age differences in predicting working memory performance from network-based functional connectivity. Cortex. 2020;132:441–59.
29. Kandaleft D, Murayama K, Roesch E, Sakaki M. Resting-state functional connectivity does not predict individual differences in the effects of emotion on memory. Sci Rep. 2022;12:14481.
30. Soch J, et al. Structural and functional MRI data differentially predict chronological age and behavioral memory performance. Eneuro. 2022;9:ENEURO.0212-22. https://doi.org/10.1523/ENEURO.0212-22.2022.
31. Yu J, Fischer NL. Age-specificity and generalization of behavior-associated structural and functional networks and their relevance to behavioral domains. Hum Brain Mapp. 2022;43:2405–18.
32. Peters A, et al. Framework and baseline examination of the German National Cohort (NAKO). Eur J Epidemiol. 2022;37:1107–24.
33. Caspers S, et al. Studying variability in human brain aging in a population-based German cohort- rationale and design of 1000BRAINS. Front Aging Neurosci. 2014;6:149.
34. German National Cohort (GNC) Consortium. The german National Cohort: aims, study design and organization. Eur J Epidemiol. 2014;29:371–82.
35. Bamberg F, et al. Baseline MRI examination in the NAKO Health Study—findings on feasibility, participation and dropout rates, comfort, and image quality. Dtsch Ärztebl Int. 2024. https://doi.org/10.3238/arztebl.m2024.0151.
36. Buhrmann A, et al. Cerebellar grey matter volume in older persons is associated with worse cognitive functioning. Cerebellum. 2021;20:9–20.
37. Garnier-Crussard A, et al. White matter hyperintensities across the adult lifespan: relation to age, Aβ load, and cognition. Alzheimers Res Ther. 2020;12:127.
38. Janacsek K, et al. Subcortical cognition: the fruit below the rind. Annu Rev Neurosci. 2022;45:361–86.
39. Bamberg F, et al. Whole-body MR imaging in the german national cohort: rationale, design, and technical background. Radiology. 2015;277:206–20.
40. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. Neuroimage. 1999;9:179–94.
41. Fischl B, et al. Whole brain segmentation. Neuron. 2002;33:341–55.
42. Griffanti L, et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): a new tool for automated segmentation of white matter hyperintensities. Neuroimage. 2016;141:191–205.
43. Gaser C, Kalc P, Cole JH. A perspective on brain-age estimation and its clinical promise. Nat Comput Sci. 2024. https://doi.org/10.1038/s43588-024-00659-8.
44. Tustison NJ, Cook PA, Holbrook AJ et al. The ANTsX ecosystem for quantitative biological and medical imaging. Sci Rep 2021;11:9068. https://doi.org/10.1038/s41598-021-87564-6
45. Miller T, Bittner N, Moebus S, Caspers S. Identifying sources of bias when testing three available algorithms for quantifying white matter lesions: BIANCA, LPA and LGA. GeroScience. 2024. https://doi.org/10.1007/s11357-024-01306-w.
46. Sundaresan V, et al. Automated lesion segmentation with BIANCA: impact of population-level features, classification algorithm and locally adaptive thresholding. Neuroimage. 2019;202:116056.
47. Stein MJ, et al. Differences in anthropometric measures based on sex, age, and health status: findings from the German National Cohort (NAKO). Dtsch Ärztebl Int. 2024.https://doi.org/10.3238/arztebl.m2024.0016.
48. OECD. Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries. Paris: OECD; 1999.
49. Ganzeboom HBG, De Graaf PM, Treiman DJ. A standard international socio-economic index of occupational status. Soc Sci Res. 1992;21:1–56.
50. Kleineidam L, et al. The assessment of cognitive function in the German National Cohort (NAKO) – associations of demographics and psychiatric symptoms with cognitive test performance. World J Biol Psychiatry. 2023;24:909–23.
51. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
52. Grady C. The cognitive neuroscience of ageing. Nat Rev Neurosci. 2012;13:491–505.
53. Hedden T, Gabrieli JDE. Insights into the ageing mind: a view from cognitive neuroscience. Nat Rev Neurosci. 2004;5:87–96.
54. Park DC, Reuter-Lorenz P. The adaptive brain: aging and neurocognitive scaffolding. Annu Rev Psychol. 2009;60:173–96.
55. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172:971–80.
56. Schulz M-A, Bzdok D, Haufe S, Haynes J-D, Ritter K. Performance reserves in brain-imaging-based phenotype prediction. Cell Rep. 2024;43:113597.
57. Küppers V, et al. Lower motor performance is linked with poor sleep quality, depressive symptoms, and grey matter volume alterations. Preprint at; 2024.https://doi.org/10.1101/2024.06.07.597666.
58. Friedman NP, et al. Individual differences in executive functions are almost entirely genetic in origin. J Exp Psychol Gen. 2008;137:201–25.

59. Passarello N, et al. Boosting effect of regular sport practice in young adults: preliminary results on cognitive and emotional abilities. Front Psychol. 2022;13:957281.
60. Kang W. Associations between Big Five personality traits and episodic memory performance in young, middle-aged, and older people: evidence from the immediate and delayed word recall tasks. Pers Individ Differ. 2023;202:111967.
61. Schmiedek F, Lövdén M, Lindenberger U. Keeping it steady: older adults perform more consistently on cognitive tasks than younger adults. Psychol Sci. 2013;24:1747–54.
62. Ahmadzadeh M, Cosco TD, Best JR, Christie GJ, DiPaola S. Predictors of the rate of cognitive decline in older adults using machine learning. PLoS One. 2023;18:e0280029.
63. Dadi K, et al. Population modeling with machine learning can enhance measures of mental health. Gigascience. 2021;10:giab071.
64. Nyberg L, et al. Common prefrontal activations during working memory, episodic memory, and semantic memory. Neuropsychologia. 2003;41:371–7.
65. Schaeverbeke JM, et al. Baseline cognition is the best predictor of 4-year cognitive change in cognitively intact older adults. Alzheimers Res Ther. 2021;13:75.
66. Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MÁ. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. Sci Rep. 2020;10:20630.
67. Tucker-Drob EM, Reynolds CA, Finkel D, Pedersen NL. Shared and unique genetic and environmental influences on aging-related changes in multiple cognitive abilities. Dev Psychol. 2014;50:152–66.
68. Hilger K, et al. Predicting intelligence from brain gray matter volume. Brain Struct Funct. 2020;225:2111–29.
69. Xiao Y, et al. Predicting visual working memory with multimodal magnetic resonance imaging. Hum Brain Mapp. 2021;42:1446–62.
70. Vassilaki M, et al. Association of neighborhood socioeconomic disadvantage and cognitive impairment. Alzheimers Dement. 2023;19:761–70.
71. Caunca MR, et al. Machine learning-based estimation of cognitive performance using regional brain MRI markers: the Northern Manhattan Study. Brain Imaging Behav. 2021;15:1270–8.
72. Eickhoff SB, Langner R. Neuroimaging-based prediction of mental traits: road to utopia or Orwell? PLoS Biol. 2019;17:e3000497.