What predicts individual brain health?: a machine learning study spanning the exposome

Mostafa Mahdipour

m.mahdipour@fz-juelich.de

Heinrich Heine University Düsseldorf

Somayeh Maleki Balajoo

Heinrich Heine University Düsseldorf

Federico Raimondo

Forschungszentrum Jülich https://orcid.org/0000-0003-4087-8259

Jianxiao Wu

Heinrich Heine University Düsseldorf

Eliana Nicolaisen-Sobesky

Heinrich Heine University Düsseldorf

Shammi More

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

Felix Hoffstaedter

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

Masoud Tahmasian

Research Centre Jülich https://orcid.org/0000-0003-3999-3807

Simon Eickhoff

Research Center Juelich

Sarah Genon

Research Centre Jülich

Article

Keywords:

Posted Date: June 9th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6410523/v1

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

Abstract

Promoting brain health is vital for well-being and reducing healthcare burdens. Individual brain health as measured with the Brain Age Gap (BAG) - the difference between chronological and predicted brain age-relates to many factors. However, an holistic view, integrating the range of factors an individual brain is exposed to, is missing for understanding how the exposome shapes brain health. After computing BAG as an indicator of individual grey matter (GM) health, we predicted it using machine learning based on 261 exposome variables (spanning biomedical, environmental, lifestyle, socio-affective, and early life domains) in UK Biobank participants. Exposome data can predict GM health with factors pertaining to cardiovascular and bone health, along with alcohol and smoking, nutrition and diabetes showing greater contribution to the prediction. In such domains, life period and duration of exposure appeared crucial. This calls for early prevention in cardiovascular and metabolic health to promote life-long brain health.

Introduction

Healthy brain aging is a critical societal challenge, as it underpins motor and cognitive abilities while also contributing to the reduction of neurodegenerative disease incidence. Promoting brain health is therefore essential not only for enhancing individual well-being and quality of life but also for alleviating the growing burden of aging on healthcare systems worldwide. Addressing this issue requires an urgent focus on global strategies to promote individual brain health as acknowledged by several recent national initiatives around the world (e.g., the European CSA BrainHealth https://www.brainhealth-partnership.eu/, Healthy Brain Initiative Collaborative https://www.alz.org/hbi-collaborative#about or Brain Health Diplomacy[1]).

At the scientific level, several initiatives have also been settled to identify key factors that influence neurocognitive health. One example among the most rigorous and influential initiatives is the Lancet Commission on Dementia which regularly publishes reports on risk factors, prevention strategies, and interventions for dementia. In its most recent report, the commission highlighted 14 modifiable risk factors—including depression, diabetes, smoking, hypertension, excessive alcohol consumption, air pollution, and visual impairment—that collectively account for an estimated 45% of potentially modifiable dementia risk factors[2]. While this report provides invaluable insights into dementia prevention, these factors are drawn from disparate studies with varying methodologies, making it unclear how they interact with one another, with additional minor risk factors, and with potential protective factors at the *individual level*.

To provide a normative, whole brain, indicator of brain health at the individual level, many studies now capitalize on a Brain Age framework. With this approach, a gap ("'Brain Age Gap'", BAG) between the apparent age of the brain (or "biological'" age) and the true age (or "'chronological'" age) is used as an estimator of the brain health of any individual[3]. This indicator is seductive for population studies because it offers a normative estimator, integrating whole-brain structural patterns while minimizing the influence of traditional covariates/confounders. Furthermore, this framework is particularly insightful in aging populations in which a wide diversity of aging paces can be observed with greater paces being generally associated with brain pathology, such as dementia[4].

Further understanding and promoting neurocognitive heath at the individual level in a precision brain health perspective requires a holistic approach taking into account a myriad of factors and considering how all these factors taken altogether explain neurocognitive health of individuals from the population. A relevant conceptual framework in that context is the *exposome* which encompasses the cumulative lifelong exposure to environmental, social, and biological factors affecting brain health[5]. Thus, all factors, collectively constitute the exposome. The contemporary definition of the exposome encompasses the external exposome with factors such as education, socioeconomical deprivation, social support, stress, and environmental features (air pollution, area density or greenness for instances)[5], but also the "internal exposome" with biomedical aspects such as inflammation. We further propose to use the term *expotype* to refer a person's unique exposome profile. This perspective acknowledges that an individual's expotype can combine risk factors for diseases (e.g., smoking since adolescence) with more protective factors (e.g., regular physical activity since childhood) that may interact with, or mitigate the effect of the former factor on brain health.

Using a BAG-based grey matter health estimator, some studies have emphasized that grey matter health is tightly linked to several aspects of body health in aging populations[6–8], others have highlighted relationships with early life factors[9] and socio-affective/mental health-related factors (e.g., long-term depressive symptoms[10]). However, relationships between many other factors can also be expected, particularly with lifestyle factors, given their associations with multimodal BAG estimators[11]. The relationship between brain health and lifestyle factors may be challenging to model. While some factors are protective (e.g., nutrition diet[12]), others may accelerate brain aging (such as smoking and alcohol consumption[13,14]), and they often interact with other factors. For example, the detrimental effect of diabetes on brain health can be mitigated by an optimal lifestyle[15]. Accordingly, the grey matter health of an individual should be seen as the outcome of his/her internal phenotype formed by interaction between different organ systems, but which also relates or interacts in turn with early life factors, socio-affective factors, as well as lifestyle factors. Yet, currently, a holistic view is missing to better understand how individual grey matter health is shaped by the expotype. Capturing the complex interplay between a range of internal and external factors together forming the exposome and relating it to individual grey matter health requires multivariate approaches and predictive modeling.

In this study, we employed machine learning approaches to examine the extent to which an individual's expotype can predict grey matter health, as reflected by grey matter BAG, in an aging population. We then investigated how the set of 261 distinct exposome variables spanning across biomedical domains (such as blood pressure, diabetes, illness and cancers, hearing loss, hips circumference, arterial stiffness), mental health/socio-affective domain (such as work/job satisfaction, family relationship satisfaction), socioeconomical factors (such as ethnicity, qualification), early life factors (such as birth weights, maternal smoking around birth), adversity in childhood factors (such as felt loved as a child, sexually molested as a child), but also lifestyle factors (such as nutrition diet, alcohol intake, smoking, …) and specific external environmental exposure (such as sun exposure, noisy workplace, population density in home area) contributes to the prediction. This was possible by leveraging the extensive assessment provided for a large population cohort in the UK Biobank project. Such a machine learning framework with additional replication analyses enables us to draw solid conclusions on the factors that contribute to predicting individual grey matter health in an aging population.

Results

Brain Age Gap indicator of grey matter health

We focused on grey matter health as a key aspect of brain health (see Methods for the rationale). In this aim, we first designed a high-performing Brain Age prediction model from grey matter volume after assessing different algorithms and parcellation schemes (see Supplementary Table 1). To build an estimator that would later be sensitive to deviance from a healthy reference norm, the models were trained and explicitly tested in a subset of 5025 healthy participants from the UK Biobank (age 62.12 ± 7.16, 2579 females, see Methods). The best model was the Ridge regression algorithm with 1054 grey matter features. When this model was applied to the remaining population data (n = 34365, age range: 44–82 years, mean 63.86 ± 7.57 years, 18128 females), it demonstrated an overall very good performance in capturing the brain aging process across the UK Biobank whole population with a high correlation between the chronological and predicted age (r = 0.76) and a Mean Absolute Error (MAE) of 3.93 years. The gap (BAG) between the predicted and chronological (real) age was then computed to provide an estimator of individual grey matter health. As could be expected this estimator was normally distributed (see Methods, Fig. 5) with most participants showing an almost null gap (i.e. an apparent brain that corresponds to their chronological age) while some participants show a positive gap indicating that their brain is estimated older than their chronological age and some other participants show a negative gap, hence reflecting relatively preserved grey matter compared to their chronological age. In other words, a wide range of variations could be observed in grey matter health within the UK Biobank populations. Next, we sought to predict these variations at the individual level based on the individual expotype.

Prediction of grey matter health from the expotype

The prediction of grey matter health from the exposome variables was performed in subsets of participants with data available for a wide range of exposome variables. As illustrated in Fig. 1, for our main analysis, we identified a main subset of 3706 participants with data for 261 distinct exposome variables (see Supplementary Table 2) within the UK Biobank population. We additionally identified a bigger subset of participants for replication who only missed the left and right bone density measurements (“replication subset”', 4292 participants, 259 distinct exposome variables; Supplementary Table 2). Third, we also identified a very large subset of participants but with data for only 201 distinct exposome variables (“variables-restricted subset” mainly missing mental health and socio-affective related variables; Supplementary Table 2). These three subsets enable us to examine the stability of our main findings across different subsamples, varying in sample size. The prediction of individual grey matter heath in these subsets was performed using a random forest algorithm, which has the advantage of accounting for non-linear relationships between the exposome variables and grey matter health. However, our results were also replicated using two other popular algorithms: Support Vector Regression (SVR) and Ridge Regression.

We could significantly predict individual grey matter health based on the expotype in the main subset (r = 0.23; p = 0.002, one-sided, t-test, FDR-corrected; Supplementary Table 3). We could replicate this achievement in the replication and variables-restricted subsets (which had less exposome variables but a higher number of participants) with similar accuracies (replication subset: r = 0.23; p = 0.004 and variables-restricted subset: r = 0.24; p = 0.002, both one-sided, t-test, FDR-corrected; Supplementary Table 3) confirming the robustness of our results across subsamples. Furthermore, individual grey matter health could also be predicted using alternative algorithms, although achieving numerically slightly lower prediction accuracy (with r = 0.17; p = 0.002 for Ridge Regression and r = 0.16; p = 0.002 for SVR; both one-sided, t-test, FDR-corrected; Supplementary Table 3).

Exposome factors’ contribution to the prediction

In order to get more insight into how different exposome factors contribute to the prediction of individual grey matter health, we used the SHAP explainer. To examine the stability across subsets and algorithms, we ran the explainer on replication and variables-restricted subsets (in addition to the main subset), as well as for additional algorithms in the main subset (see Supplementary Fig. 1–5 and Supplementary Table 4–5). The predictive model makes use of the broad spectrum of variables with no specific variable or set of variables showing a disproportionate contribution to the model (see feature importance ranking in Supplementary Tables 4 and 5). In other words, the prediction does not appear to be disproportionately driven by a few specific variables, but instead, it relies on the combination of information across a wide spectrum of variables. However, when looking at the top 30 variables (Fig. 2), some exposome factors appear particularly important for predicting individual grey matter health, and this is consistent across prediction algorithms. This group of factors mainly combines cardiovascular factors (such as variables pertaining to blood pressure, but also to smoking) with nutrition, diet, alcohol, and diabetes. Additionally, bone density and hip circumference also appear in the top contributing variables consistently for the main algorithm and one of the alternative algorithms.

Additional insight into the nature of the association between these most contributing variables and grey matter health can be obtained by examining the distribution of SHAP value in Fig. 3. This figure illustrates how the value (from low to high) of the exposome variable is associated with the BAG prediction (from negative which corresponds to better grey matter heath to positive which corresponds to worst grey matter health). Overall, most of the associations follow an expected pattern.

Concretely, for factors that are consensually considered as risk factors for brain diseases, higher exposure is associated with worst grey matter health (i.e. higher BAG). These include high blood pressure, smoking and alcohol, as well as diabetes. It should be noted here that for these domains, the variables that show the highest impact on the predicted grey matter health pertain to the duration of the exposure and the life periods in which the exposure happens. Hence, for smoking, duration, age at start and age at stop appear as complementary information instead of redundant variables for the prediction. Longer smoking duration, younger age when started smoking and older age when stopped smoking are all associated with worst predicted grey matter health. Similarly for high blood pressure, a longer duration and an older age at diagnosis (likely reflecting an older age for treatment) both lead to worst predicted grey matter health. Along the same line, our results show that when diabetes diagnosis happens at an older age, a worst grey matter health is predicted. This is also the case for other non-cancer illness: diagnosis (and likely treatment) at an older age is associated with younger. In contrast, the relationship between operations (be it first or second operation) and the predicted grey matter health goes in the opposite direction: operations performed at younger age appear to relate to a worst predicted grey matter health.

For the remaining top contributing variables different patterns can be observed. In particular, for variables related to nutrition diet, while high coffee intake is associated with worst predicted grey matter health, low dried fruit and cereal intake leads also to worst predicted grey matter health. Finally, low hip circumference and low bone density are both associated with worst grey matter health.

The variables that are lower in the contribution ranking can be found in Supplementary Table 4. It can be noted that some variables do pertain to the general factors listed above, but do not importantly contribute to the model. Examples of such variables are current smoking status (ranked 207 with random forest), vascular/heart problems diagnosed by doctor (ranked 154 with random forest) with the main algorithm, diabetes diagnosed by doctor (ranked 231 with random forest), history of injury caused by alcohol consumption ("Ever been injured or injured someone else through drinking alcohol"", ranked 240 with random forest). In contrast to variables from the same domain appearing in the top contributing group, these variables do not contain information about the life period or duration of the exposure. This pattern suggests that the most important aspect of cardiovascular, metabolic, and lifestyle factors on individual grey matter heath is the life period and chronicity of the exposure.

We can also note that a range of other variables show a relatively low contribution to the prediction of individual grey matter heath. These include mental health related variables (such as history of unusual and psychotic experience, ever had period of mania/excitability and depression-related problems) and socio-affective factors (such as friendships satisfaction), early life factors (such as maternal smoking around birth and "someone to take to physician when needed as a child"), ethnicity, but also specific exposure such as cannabis use and sun exposure. Although these factors can potentially contribute to predicting individual grey matter health, their influence on the model is relatively negligible.

Finally, these patterns do not importantly influenced by sex/gender. Although in the main analysis, we treated sex/gender as a covariate, in a supplementary analysis, we instead included it as a predictor variable in the model. We observed overall similar results. Furthermore, the contribution (SHAP value) of sex/gender was very low, indicating minimal influence on the predictions. This suggests that the model's performance remains stable and generalizable, regardless of whether sex/gender is accounted for as a covariate or included as a predictor variable.

Taken together, our results suggest that factors pertaining to the internal expotype including cardiovascular, metabolic and musculoskeletal aspects, together with smoking, alcohol and diet are the strongest predictors of individual grey matter health. Importantly, variables informing about the chronicity of life period and the duration of the exposure are the most important for individual prediction.

Discussion

By considering the exposome reflected in the combination of more than 200 variables spanning different body systems, socio-affective, early life, as well as life style factors, individual grey matter health can hence be predicted in an aging population. Importantly, the predictive model appears to make use of the wide spectrum of variables. In other words, the contribution to the individual prediction is distributed across variables. This confirms that multifaceted and multivariate views on the exposome are needed to explain interindividual variability in brain health. In that view, the heterogeneity of factors to be considered can be appreciated from our findings on the top contributing variables. These include measurements of the cardiovascular system (pertaining to blood pressure), but also specific metabolic conditions (namely diabetes), common risk factors or body diseases (alcohol and smoking), but also musculoskeletal factors (namely bone density), nutrition diet (namely coffee, dried fruits and cereals intake) and also medical history of operations/illnesses.

Hence, our results reinforce recent evidence of associations in large aging population cohorts between both the metabolic system, and the cardiovascular system, with brain health[6–8]. In that context, our study more specifically suggests that alcohol consumption, smoking, high blood pressure and diabetes are the strongest risk factors for accelerated brain structural aging. It thus adds to the broad evidence that alcohol consumption negatively impacts brain structure, by affecting a range of structural features and being associated with several aging patterns[16–18]. Furthermore, alcohol consumption likely interacts with other risk factors[19] in that context making it one of the top relevant variable when aiming to promote individual brain health.

Our results also show that the top contributing variables for the key other risk factors that are smoking, high blood pressure and diabetes pertains to the life period at which the exposure has taken place and the duration of the exposure. It should be noted that both variables pertaining to life period and variables pertaining to duration for smoking and higher blood pressure importantly contribute to the prediction. This demonstrates that life period and chronicity of the exposure are complementary (rather than redundant) information for determining individual brain health. Overall, the sooner these detrimental exposures are discarded (by stopping smoking and by diagnosing and treating high blood pressure), the lower the negative impact on brain health. Thus our study crucially highlights the need of early prevention campaigns[20] targeting diabetes, smoking and blood pressure monitoring for public health strategies.

Another important factor from the internal exposome that recently appeared to be associated with brain health in the UK Biobank is the musculoskeletal system [8]. In the current study, we found that bone mineral density, specifically, was often one of the most important contributors for predicting individual grey matter health. Although the relationship between bone health and brain health has received little attention in the past, evidence for a skeletal-muscle-brain axis exists[21]. For example, low bone mineral density is a risk factor for dementia[22]. In that context, some shared underlying pathophysiological mechanisms have been proposed[23]. In particular, an important neuroprotective role of irisin (a myokine induced during physical exercise by the musculoskeletal system) has been often pointed out as a potential key mechanism in linking bone heath to brain heath [24]. Thus, the effects on neurocognitive health of multifaceted Interventions (like physical exercise, vitamin, calcium, hormonal therapy and other medications) that contribute to preserve bone health across aging should be further investigated in future studies.

Beyond the critical effect of body health, our multivariate model also revealed the contribution of nutrition and diet on grey matter health. In particular, coffee, dried fruit, and cereals intake appear in the top most contributing factors consistently across algorithms. A higher coffee intake leads to a worst predicted grey matter health. Although a neuroprotective role for coffee is often discussed in the literature, our results demonstrate that a daily consumption above the moderate range (around 2 cups/day, the mean of the sample, see summary statistics in Supplement) leads to a worst predicted grey matter health. This finding is consistent with a recent report of positive association between coffee intake and several brain atrophy patterns in the UK Biobank[18]. The detrimental effect of high coffee consumption on brain health can potentially be explained by interaction with other risk factors discussed above, as well as by gene-diet interaction. It has for example been showed that in UK Biobank participants with genetic predisposition to elevated intraocular pressure, greater caffeine consumption was associated with higher intraocular pressure and higher glaucoma prevalence [25]. Altogether all these results point towards recommending relatively low coffee consumption (not more than 2 cups/day) for life-long brain health considering that high coffee consumption can detrimentally interact with factors impacting brain health (in particular cardiovascular risk factors) leading ultimately to poorer grey matter health.

In contrast, high cereals and high dried fruit intakes generally lead to better predicted brain health. Our findings are aligned with previous evidence that diets in which cereals and dried fruits (such as nuts) are predominant (such as the MIND/DASH diets) have a beneficial effect on brain structure, cognition, and dementia risk[26,27]. It should be noted here that cereal intake mostly reflects whole-grain cereal intake based on the type of cereals that were reported by the participants (see Supplement Table 6). Several elements (in particular fibers) in cereals and dried fruits are known to have health-improving properties and may even mitigate the effects of diabetes on brain health[28]. Although further studies are needed in the future to better understand how diet and other factors interact with metabolic aspects (in particular diabetes) to influence individual brain health, our current study suggests that high (whole-grain) cereal and dried fruits intake are important factors contributing to promote grey matter health.

In our machine learning models, hip circumference also appears to contribute to the prediction of individual grey matter health. More concretely, lower (than the mean of 102 cm, see Supplement) hip circumference is associated with worst individual grey matter health in the prediction model. However, additional analyses show that, when taken in isolation, hip circumference is not associated with grey matter health in the main subset (null correlation with BAG, see Supplementary Fig. 6). This suggests that hip circumference plays a role as a moderator variable in the individual prediction of grey matter health. In other words, lower hip circumference on its own does not result in worst grey matter health, but in interaction with other factors (such as low bone density), it can lead to worst predicted grey matter heath. It can for example be assumed that a combination of low hip circumference and low bone density reflects physical frailty, a body condition associated with neurocognitive impairment[29]. Although further studies are needed to characterize in which individual profiles, low hip circumference is associated with low grey matter heath, our study indicates that this morphological information may be highly relevant for individual prediction of impaired grey matter health.

Finally, some other factors included in the predictive model, show relatively null or negligible contributions in the prediction of grey matter health. These factors include mental health-related variables (such as having potentially already experienced unusual and psychotic experiences, depression-related items, and socio-affective factors), early life factors (such as maternal smoking and adversity in childhood), ethnicity, specific exposure such as cannabis use, sun exposure, and home area population density. Although these external factors can be found to relate to grey matter in some studies [30–32], our results show that other factors, in particular those more directly related to body health or the internal exposome, have a greater contribution when it comes to predicting the global grey matter health at the individual level. Thus, when using a multivariate approach that offers a holistic view on the exposome, we can observe that some exposome factors have a relatively minor contribution while others, pertaining to individual body health and related lifestyle, play a more important role in explaining interindividual variability in grey matter in aging.

In sum, by using a machine learning approach that allows a holistic view, our study has paved the way towards individual prediction of brain phenotype from the expotype. Several limitations should be noted, however, at this stage. First, the prediction accuracy remains relatively moderate. This suggests that additional factors should be taken into account. These should include more fine biological measurements (such as those based on biofluids) beyond the basic biological information on which we focused in this study for the sake of optimizing the sample size. Despite this limitation in the investigated factors, the current study set a primary framework for precision brain health and public health policies by revealing the spectrum of mostly modifiable factors explaining variability in grey matter health. In the future, another important step towards precision brain health would be to leverage individual genetic profiles to examine to which extent the combination of the individual genotype and the individual expotype can predict individual brain health in aging. It has indeed been shown that brain health as reflected by BAG relates to genetic factors in genome-wise association studies[33,34]. In the perspective of combining genetic and exposome factors, including more diverse cohorts with regard to ethnicity and geographic areas should also be considered. This would contribute in the future to a better understanding of how sociodemographic factors interact with others in explaining interindividual variability in brain health [5,35].

Methods

General workflow

Machine learning models were developed in two main steps in this study. First, a brain age prediction model based on grey matter data was developed. This model was needed to derive a grey matter health indicator (BAG) based on the discrepancy between the predicted and the chronological age. Second, a grey matter health (BAG) prediction model was developed based on exposome data. Both steps were carried on in the UK Biobank but the two models were developed in strictly separated subsets of participants (to avoid data leakage) as illustrated in Fig. 4 below.

To develop a brain age prediction model, we identified a large subset of 5025 (cognitively) healthy participants (“‘healthy sample’”) within UK Biobank. Within the remaining population of the UK Biobank, we identified one main subset, a replication subset and a “‘variables-restricted’” subset of participants (see Supplementary Table 2) to which the previously developed brain age model was applied to compute a BAG for each and every participant. These subsets were defined based on the availability of exposome variables in the participants (see Fig. 1). In the second step, machine learning models were developed to predict individual grey matter health (i.e. BAG) in the participants of these three subsets.

Participants

Detail explanation of data collection in UK Biobank cohort can be found in https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf. This study focus on participants with available imaging data (39390 participants, aged 44–82 years, mean 63.64 ± 7.54 years, n = 20707 females). Written informed consent was obtained from all participants. The present analyses were conducted under data application number 41655 and were approved by the ethical committee of the Heinrich Heine University Düsseldorf.

Cognitively healthy participants (“‘healthy sample”) had no self-reported long-standing illness disability or infirmity (UK Biobank data field #2188), no self-reported diabetes (field #2443), no stroke history (field #4056), no ICD-10 diagnosis and good or excellent self-reported health (field #2178). These criteria were defined in line with a previous study[11] and led to a sample of 5025 healthy participants (age range 46–

82 years, mean 62.12 years ± std 7.16 years, 2579 females), while leaving 34365 participants (age range: 44–82 years, mean 63.86 ± 7.57 years, 18128 females) to define subsets for exposome based prediction of grey matter health (see Supplementary Table 2).

Grey Matter Features

To obtain an estimator of grey matter health, we used grey matter volume features. It should be noted that when using a Brain Age Gap Indicator, depending on the neuroimaging features which are used, different neurobiological aspect of brain health are probed. Many recent studies in large neuroimaging databases have capitalized on multimodal neuroimaging data (typically including functional, white matter, and grey matter features) to derive a Brain Age Gap estimator. This combination of different features was generally for the sake of optimizing the model's accuracy[3], but it results in a global brain health indicator (combining different aspects of brain structure and function) that may lack specificity from a neurobiological standpoint. In contrast, unimodal estimators, such as those explicitly based on grey matter, can be derived with close accuracies to indicator derived a multimodal model[3,11] while providing a more specific insight into the healthiness of grey matter and its relationship to different factors.

Comprehensive information regarding the neuroimaging data from the UK Biobank is available here: https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf. T1-weighted MRI images were acquired by 3 Tesla Siemens Skyra scanners with 32 channel head coils and used an MPRAGE sequence with 1-mm isotropic resolution, with Field-of-view: 208×256×256. For Imaging data preprocessing, we used an in-house developed framework for computationally reproducible processing of large-scale data (FAIRly big[36]). For this purpose, a singularity container with a pipeline to perform voxel-based morphometry (VBM)[37] on individual T1-weighted MRI images based on the Computational Anatomy Toolbox (CAT)[38] was created. All T1-w anatomical were processed with the CAT version 12.7. After normalisation and segmentation, the grey matter volume segments were modulated for non-linear transformations and smoothed. Grey matter was parcellated using a combination of the Schaefer atlas for 200, 400, 600, 800 and 1000 cortical regions[39] and the Melbourne subcortex atlas for 32 and 54 subcortical regions[40] leading to five levels of representations of grey matter (grey matter volume for either 232, 454, 654, 854 and 1054 regions). After evaluation of brain age prediction performance with grey matter features at these different levels of granularity (see Supplementary Table 1), the atlas of 1054 regions was selected as the optimal grey matter representation for BAG computation in the UK Biobank population.

Brain Age Prediction Model

Four predictive algorithms were evaluated to design a Brain Age Prediction Model. These included Linear Regression, Ridge Regression[41], Support Vector Regression (SVR)[42,43], and Random Forest (RF)[44] as implemented in scikit-learn[45] and Julearn packages[46]. They were all trained to predict an individual's chronological age using the five different sets of grey matter features (with different levels of granularity) in healthy sample (n = 5025). Predictive brain age models were trained on 80% of the healthy sample (n = 4020) using 10-fold nested cross-validation with 5 repeats for estimating the chronological age based on the five different sets of individual's grey matter features. Furthermore, a 10-folds inner cross-validation loop was implemented for hyperparameter tuning using a grid search approach. The optimized brain age models were then validated on held-out set made by the remaining 20% of the healthy sample (n = 1005). Model performance was quantified using the Pearson correlation coefficient (r) and Mean Absolute Error (MAE) between predicted and chronological age in the held-out set.

Finally, the model with the minimum prediction error in held-out set was selected and fitted on the entire healthy sample (training + held-out) and used to estimate the chronological age in population set (n = 34365, age range: 44–82 years, mean 63.86 ± 7.57 years, 18128 females) in the UK Biobank dataset. According to the results presented in the Supplementary Table 1, this best model was Ridge regression with 1054 grey matter features.

Importantly, several studies have brought attention to an age bias/age dependency in brain age prediction requiring a so-called age-bias correction[3]. This was done here by regressing out the effects of age on the predicted age[47]. To do so, the slope and intercept of the regression line between chronological age and the predicted age were estimated in the training set and used to adjust the predicted age in the test and the population samples.

Brain Age Gap (BAG) as an indicator of grey matter health

The predicted age was then used to compute an indicator of grey matter health for each individual participant based on the discrepancy between his/her (true) chronological age and the "'apparent'" (i.e. predicted) age of his/her brain grey matter. This was done by calculating the so-called "'brain age gap'" (BAG) by subtracting the actual chronological age from the adjusted predicted brain age. Accordingly, it provided a normalized measure of the extent to which an individual's brain appeared older (BAG > 0) or younger (BAG < 0) than same-aged peers. The BAG offers a distinctive advantage as it is a personalized measure by nature. Moreover, it is cross-validated and does not rely on chronological age, allowing it to directly assess deviations from population norms. In agreement, with this view, this estimator was normally distributed in the UK Biobank whole population, as well as within our main subset, the replication subset and the variable-restricted subset (see Fig. 5). In other words, across all samples, most participants showed an almost null gap (i.e. an apparent brain that corresponds to their chronological age) while some participants show a positive gap indicating that their brain is estimated older than their chronological age and some other participants show a negative gap hence reflecting relatively preserved grey matter compared to their chronological age.

Exposome variables

To represent the exposome, we used a wide range of non-imaging variables encompassing biomedical, lifestyle, socio-economical and early life factors as illustrated in Fig. 1 and the full list of exposome variables are available in Supplementary Table 2. The selection of variables was guided by previous publications reporting either associations (with correlation or regression usually) with brain health as measured by the brain age gap[8,9,11,15] or associations with grey matter structure[48] in the UK Biobank. This choice was also constrained by data availability. To avoid potential biases associated with substantial missing values across participants, we focused on variables that were available in at least 2000 participants including both males and females. As illustrate in Fig. 1, this led us to define a main subset in which 261 distinct exposome variables were available for 3706 participants. By dropping out two variables (left and right heel bone density), we could create a bigger subset of 4292 participants that served as a replication subset. Finally, we also identified a bigger subset of 7736 participants with data for 201 distinct exposome variables. In this "'variables-restricted subset", mainly variables related to socio-affective and mental health domains were missing compared to the main and replication subsets. Standard preprocessing were performed on exposome variables including: handling non-informative value (e.g.: negative values indicating "'prefer not to answer'"), calculation of duration from age variables and variables standardization.

Grey Matter Health prediction model

To predict grey matter health at the individual level, we first implemented a random forest algorithm which used a decision tree-based approach and has the advantage of accounting for non-linear relationships between the exposome variables and grey matter health. Random Forest was trained and tested for each subset (i.e. the main subset, the replication subset and the "'variables-restricted'" subset separately), However, for the sake of replication, we also implemented two additional popular algorithms: ridge regression (which was also used for the brain age model) and Support Vector Regression (SVR). Each additional algorithm was trained and tested on the main subset. Each model was trained using BAG as the target variable (representing grey matter health) and exposome variables as input features within a nested cross-validated scheme. This ensured unbiased performance estimation, with 5 inner folds for hyperparameter tuning and 5 outer folds with 5 repeats to estimate generalization performance. Grid search and Optuna search[49] was employed in the inner folds for hyperparameter tuning.

To mitigate the disproportionate influence of larger-scale features, standardization was applied using the Standard Scaler function from scikit-learn package[45]. Controlling for covariates was performed on the exposome variables within the cross-validation framework (using estimated from the train set). The set of covariates/ confounds include age, square of age, sex, height, and volumetric scaling from T1 head image to standard space. Thus, all care was taken to neutralize the influence of usual confounds by using a normative indicator of grey matter health and by additionally controlling for covariates within the prediction analysis.

Model performances were evaluated using multiple metrics including the mean absolute error, mean squared error, root mean squared error, r-square, and Pearson correlation between predicted BAG and true BAG from the test sets for each outer fold. For each subset and each algorithm, permutation test was performed to assess significance against a null distribution by shuffling the scores of the Grey

Matter Health variable in 500 repeats of 5-fold cross-validation. Multiple comparisons across different metrics were corrected using false discovery rate (FDR; Benjamini and Hochberg 1995) of $p < 0.05$. Additionally, after selecting the best model, SHapley Additive exPlanations (SHAP)[50] were employed to provide an insight into how individual exposome variables contribute to the prediction. Indeed, SHAP offers a cohesive framework for interpreting predictions in explainable Artificial Intelligence, assigning importance to each variable contributing to a prediction. All analysis steps were conducted using Python, SHAP[50], scikit-learn[45] and the Julearn package[46].

Declarations

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data were obtained from the UK Biobank. Researchers can register to access all data used in this study via the UK Biobank Access Management System (https:// bbams.ndph.ox.ac.uk/ams/).

Code availability

The analysis code is publicly available through GitHub (https://github.com/MostafaMahdipour/Predicting_Brain_Age_Gap_BAG_using_UKB_exposome).

Acknowledgements

This research has been conducted using data from UK Biobank (https://www.ukbiobank.ac.uk/), a major biomedical database. We are grateful to UK Biobank for making the data available and to all study participants, who generously donated their time to make this resource possible. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/2–1, GE 2835/9-1, SFB 1451 – Project-ID 431549029). SMB is supported by the EBRAINS 2.0 Project funded from the European Union’s Horizon Europe Programme under the Specific Grant Agreement No. 101147319.

Author contributions

Study concept and design was by M.M., S.B.E., and S.G. Data preparation and preprocessing was carried out by M.M., E.N.S., F.H., and S.M.B. Model development was by M.M., S.M.B., F.R., and S.G. Data interpretation was by M.M., S.M.B., M.T., S.B.E., and S.G. Drafting of the manuscript was by M.M., S.M.B., and S.G. M.M. conducted the analyses, prepared the tables and figures. S.M., J.W., F.R., S.M.B., and S.G. contributed to the statistical/machine learning analyses. Critical revision of the manuscript for important intellectual content was by M.T., and S.B.E. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Dawson, W.D., *et al.* The Brain Health Diplomat's Toolkit: supporting brain health diplomacy leaders in Latin America and the Caribbean. *The Lancet Regional Health – Americas* **28**(2023).
2. Livingston, G., *et al.* Dementia prevention, intervention, and care: 2024 report of the <em>Lancet</em> standing Commission. *The Lancet* **404**, 572-628 (2024).
3. Gaser, C., Kalc, P. & Cole, J.H. A perspective on brain-age estimation and its clinical promise. *Nat Comput Sci* **4**, 744-751 (2024).
4. Varikuti, D.P., *et al.* Evaluation of non-negative matrix factorization of grey matter in age prediction. *Neuroimage* **173**, 394-410 (2018).
5. Ibanez, A., *et al.* Neuroecological links of the exposome and One Health. *Neuron* **112**, 1905-1910 (2024).
6. De Lange, A.-M.G., *et al.* Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study. *NeuroImage* **222**, 117292 (2020).
7. Wagen, A.Z., *et al.* Life course, genetic, and neuropathological associations with brain age in the 1946 British Birth Cohort: a population-based study. *The Lancet Healthy Longevity* **3**, e607-e616 (2022).
8. Tian, Y.E., *et al.* Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nature medicine* **29**, 1221-1231 (2023).
9. Vidal-Pineiro, D., *et al.* Individual variations in 'brain age'relate to early-life factors more than to longitudinal brain change. *elife* **10**, e69995 (2021).
10. Dintica, C.S., *et al.* Long-term depressive symptoms and midlife brain age. *Journal of affective disorders* **320**, 436-441 (2023).
11. Cole, J.H. Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. *Neurobiology of aging* **92**, 34-42 (2020).
12. Kapogiannis, D., *et al.* Brain responses to intermittent fasting and the healthy living diet in older adults. *Cell Metabolism* **36**, 1668-1678.e1665 (2024).
13. Bittner, N., *et al.* When your brain looks older than expected: combined lifestyle risk and BrainAGE. *Brain Structure and Function* **226**, 621-645 (2021).
14. Jawinski, P., *et al.* Linking brain age gap to mental and physical health in the Berlin aging study II. *Frontiers in Aging Neuroscience* **14**, 791222 (2022).
15. Dove, A., *et al.* Diabetes, Prediabetes, and Brain Aging: The Role of Healthy Lifestyle. *Diabetes Care* **47**, 1794-1802 (2024).

16. Daviet, R.*, et al.* Associations between alcohol consumption and gray and white matter volumes in the UK Biobank. *Nature communications* **13**, 1175 (2022).
17. Topiwala, A., Ebmeier, K.P., Maullin-Sapey, T. & Nichols, T.E. Alcohol consumption and MRI markers of brain structure and function: Cohort study of 25,378 UK Biobank participants. *NeuroImage: Clinical* **35**, 103066 (2022).
18. Yang, Z.*, et al.* Brain aging patterns in a large and diverse cohort of 49,482 individuals. *Nature Medicine* **30**, 3015-3026 (2024).
19. Thornton, V.*, et al.* Alcohol, smoking, and brain structure: common or substance specific associations. *medRxiv* (2024).
20. Farina, F.R.*, et al.* Next generation brain health: transforming global research and public health to promote prevention of dementia and reduce its risk in young adult populations. *The Lancet Healthy Longevity* (2024).
21. Kalc, P., Dahnke, R., Hoffstaedter, F. & Gaser, C. Low bone mineral density is associated with gray matter volume decrease in UK Biobank. *Frontiers in aging neuroscience* **15**, 1287304 (2023).
22. Xiao, T.*, et al.* Association of bone mineral density and dementia: the Rotterdam study. *Neurology* **100**, e2125-e2133 (2023).
23. Frame, G., Bretland, K.A. & Dengler-Crish, C.M. Mechanistic complexities of bone loss in Alzheimer's disease: a review. *Connective tissue research* **61**, 4-18 (2020).
24. Sadier, N.S.*, et al.* Irisin: An unveiled bridge between physical exercise and a healthy brain. *Life Sciences* **339**, 122393 (2024).
25. Kim, J.*, et al.* Intraocular Pressure, Glaucoma, and Dietary Caffeine Consumption: A Gene–Diet Interaction Study from the UK Biobank. *Ophthalmology* **128**, 866-876 (2021).
26. Zhang, J.*, et al.* Associations of midlife dietary patterns with incident dementia and brain structure: findings from the UK biobank study. *The American journal of clinical nutrition* **118**, 218-227 (2023).
27. Charisis, S., Yannakoulia, M. & Scarmeas, N. Diets to promote healthy brain ageing. *Nature Reviews Neurology*, 1-12 (2024).
28. Tufail, T.*, et al.* Cereals: An Overview. *Cereal Grains*, 1-13 (2023).
29. Jiang, R.*, et al.* Associations of physical frailty with health outcomes and brain structure in 483 033 middle-aged and older adults: a population-based study from the UK Biobank. *The Lancet Digital Health* (2023).
30. Vered, S., Sznitman, S. & Weinstein, G. The association between cannabis use and neuroimaging measures in older adults: findings from the UK biobank. *Age and Ageing* **53**(2024).
31. Madden, R.A.*, et al.* Structural brain correlates of childhood trauma with replication across two large, independent community-based samples. *European Psychiatry* **66**, e19 (2023).
32. Li, H., Cui, F., Wang, T., Wang, W. & Zhang, D. The impact of sunlight exposure on brain structural markers in the UK Biobank. *Scientific Reports* **14**, 10313 (2024).

33. Jónsson, B.A., *et al.* Brain age prediction using deep learning uncovers associated sequence variants. *Nature communications* **10**, 1-10 (2019).

34. Yi, F., *et al.* Genetically supported targets and drug repurposing for brain aging: A systematic study in the UK Biobank. *Science Advances* **11**, eadr3757 (2025).

35. Moguilner, S., *et al.* Brain clocks capture diversity and disparities in aging and dementia across geographically diverse populations. *Nature Medicine* **30**, 3646-3657 (2024).

36. Wagner, A.S., *et al.* FAIRly big: A framework for computationally reproducible processing of large-scale data. *Scientific data* **9**, 80 (2022).

37. Ashburner, J. & Friston, K.J. Voxel-based morphometry—the methods. *Neuroimage* **11**, 805-821 (2000).

38. Gaser, C., *et al.* CAT: a computational anatomy toolbox for the analysis of structural MRI data. *Gigascience* **13**, giae049 (2024).

39. Schaefer, A., *et al.* Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex* **28**, 3095-3114 (2018).

40. Tian, Y., Margulies, D.S., Breakspear, M. & Zalesky, A. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nature neuroscience* **23**, 1421-1432 (2020).

41. Rifkin, R.M. & Lippert, R.A. Notes on regularized least squares. (2007).

42. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **10**, 61-74 (1999).

43. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* **2**, 1-27 (2011).

44. Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001).

45. Pedregosa, F., *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).

46. Hamdan, S., *et al.* Julearn: an easy-to-use library for leakage-free evaluation and inspection of ML models. *arXiv preprint arXiv:2310.12568* (2023).

47. de Lange, A.-M.G. & Cole, J.H. Commentary: Correction procedures in brain-age prediction. *NeuroImage: Clinical* **26**(2020).

48. Miller, K.L., *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience* **19**, 1523-1536 (2016).

49. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* 2623-2631 (2019).

50. Lundberg, S.M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30**(2017).

Figures

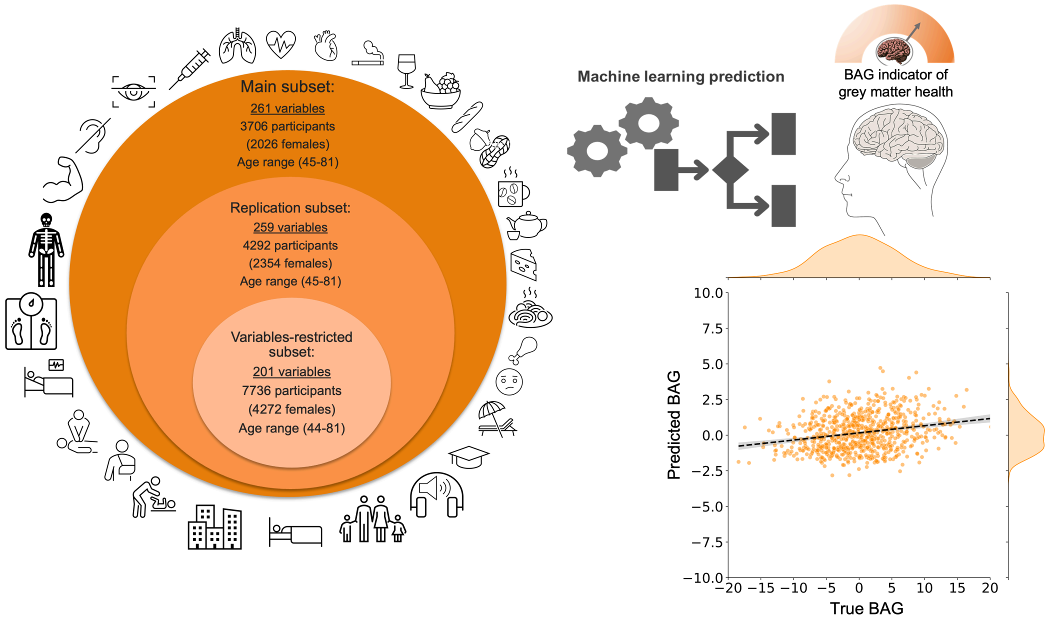

Figure 1

Prediction of the individual grey matter health (BAG) from the individual expotype. Left panel: Machine learning algorithms were trained and tested in three subsets of participants. More than 200 variables were included in each subset, spanning biomedical, lifestyle, socioaffective, early life, and environmental domains. Right panel: individual grey matter health could be significantly predicted from the individual expotype. A full list of all exposome variables in the main, replication, and variables-restricted subsets are available in Supplementary Table 2, and detailed prediction performances are available in Supplementary Table 3.

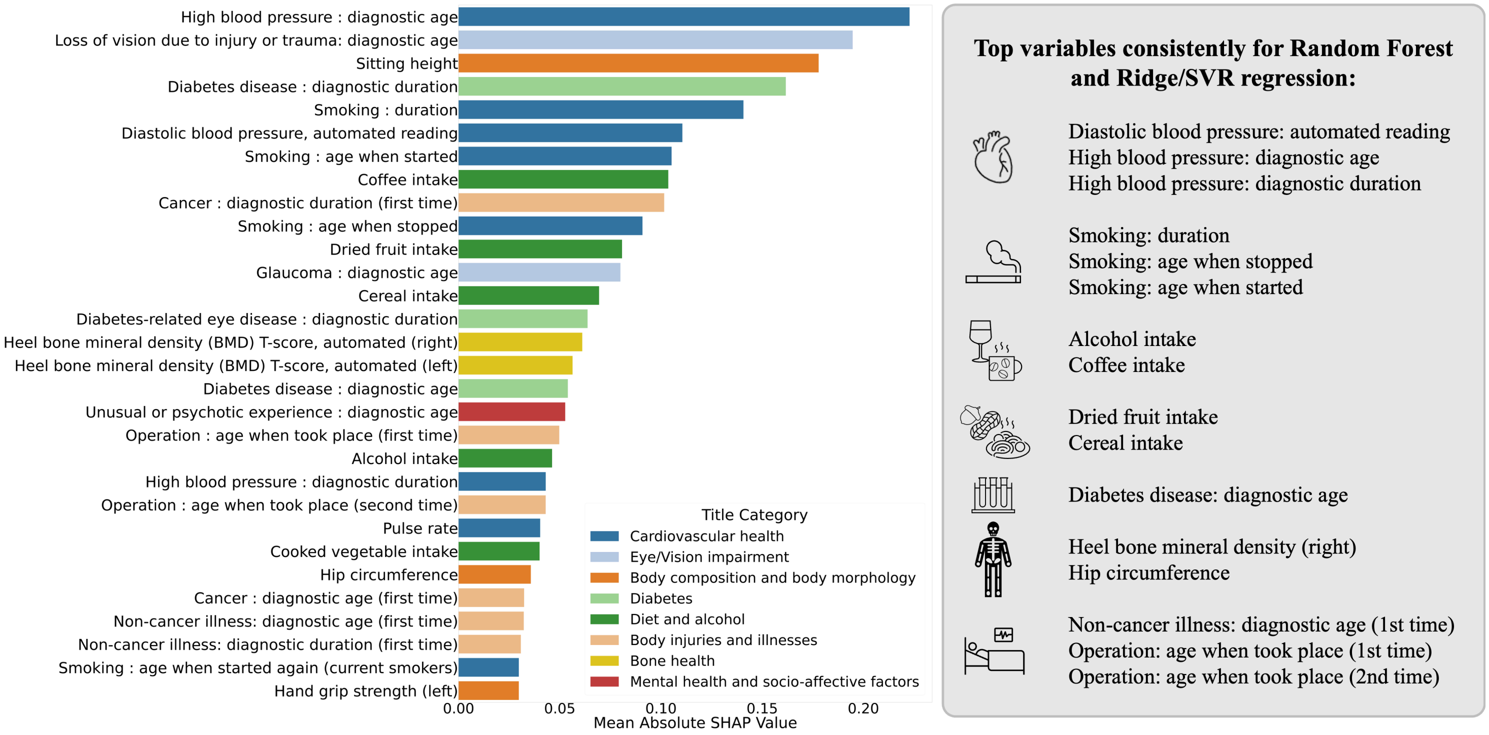

Figure 2

Thirty top contributing variables in the main subset (n = 3706, 261 exposome variables). Left Panel: Exposome variables contribution (as reflected by absolute SHAP value) for prediction of grey matter health based on random forest algorithm. Right panel: summary of the most contributing variables consistently across random forest and at least one other algorithm (see Supplementary Figures 1 and 2).

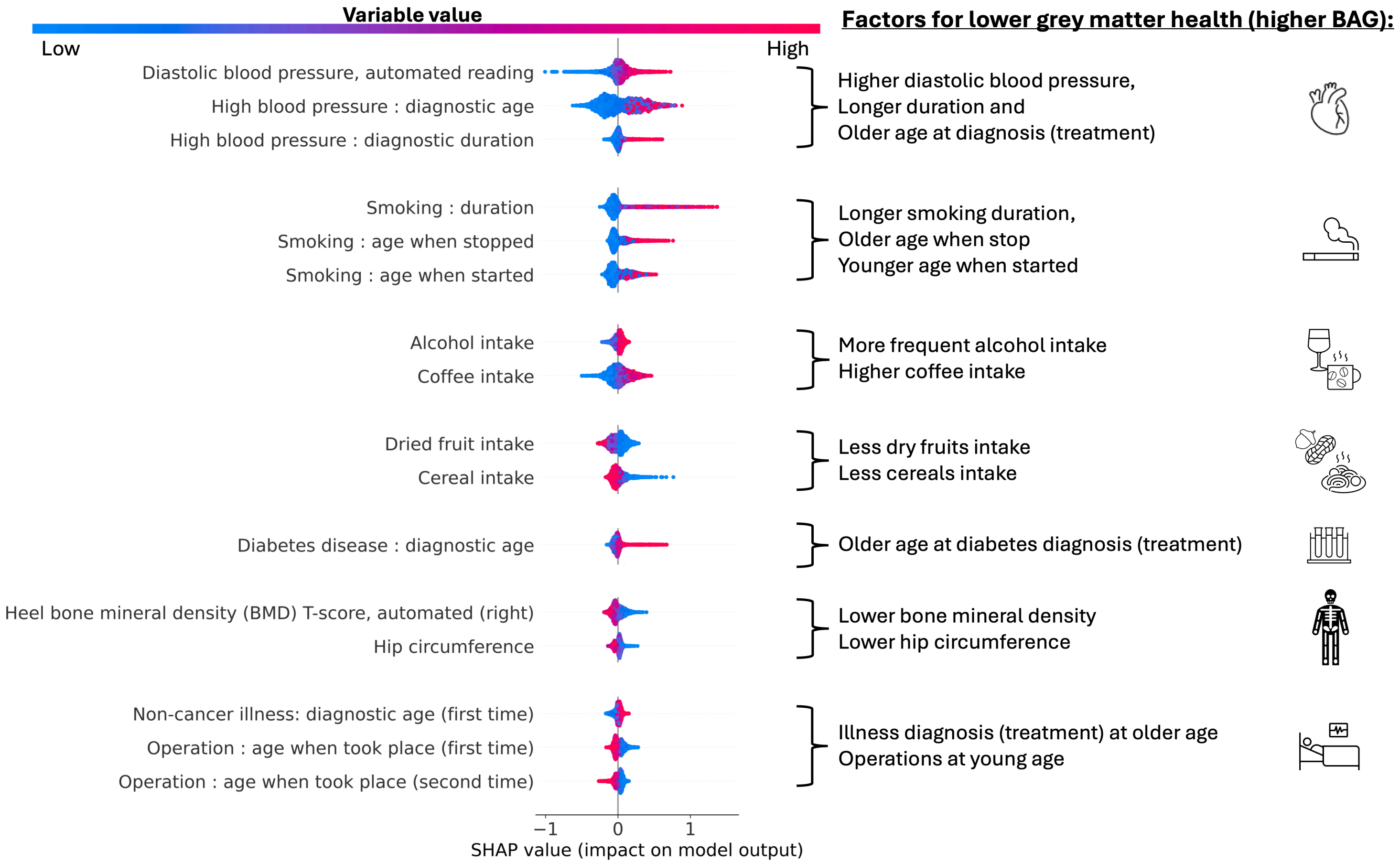

Figure 3

SHAP value for the common most contributing variables when using random forest in the main subset (n = 3706, 261 exposome variables). The x axis represents the impact on the prediction: zero could be considered as the default prediction, while a shift to the right reflects a higher predicted BAG (i.e., worst predicted grey matter health) and a shift to the left represents a lower predicted BAG (i.e., a better predicted grey matter health). Blue indicates a lower value for the variable, while pink/red indicates a higher value. The right panel summarizes which direction of the variable leads to worst grey matter health (i.e. to higher BAG).

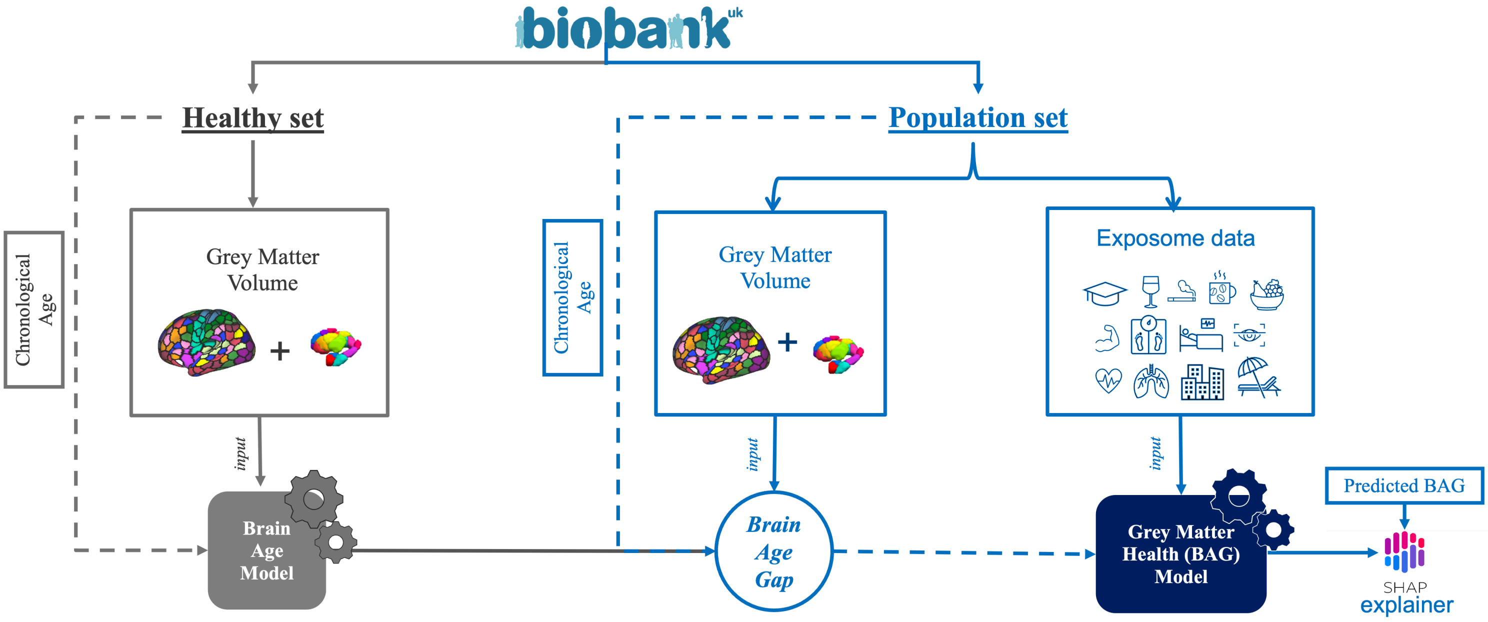

Figure 4

General workflow in the UK Biobank. Two machine learning models were developed: a Brain Age Prediction Model was developed from an healthy set of UK Biobank participants (n = 5025, 2579 females) while a Grey Matter Health (BAG) Prediction Model based on exposome data was developed in participants who have not been included in the healthy set to avoid any data leakage. Continuous lines indicates features/predictors fed into the model while dashed are used for target variables. Three subsets were further created from the population set to evaluate grey matter health (i.e. BAG) prediction models.

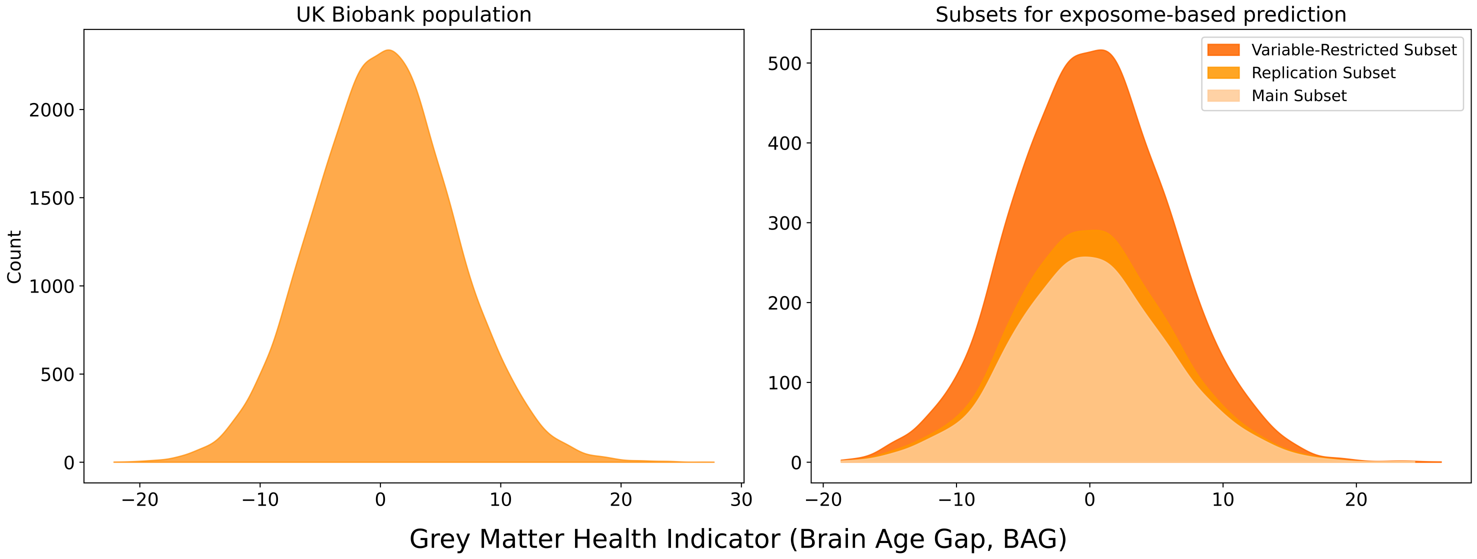

Figure 5

Distribution of the Brain Age Gap in the remaining UK Biobank population (i.e. excluding the healthy subset in which the Brain Age prediction model was developed; n=34365, age range: 44-82 years, mean

63.86 ± 7.57 years, 18128 females) and in the specific subsets of participants (main subset, n = 3706; replication subset, n = 4202; variable-restricted subset, n = 7736) that were used for exposome-based prediction of grey matter health (see below).

Supplementary Files

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

- SupplementaryMaterialsMahdipourGenon.docx
- SupplemantaryTable4.docx
- SupplemantaryTable5.docx