






RESEARCH ARTICLE

Machine learning to discover factors predicting volume of white matter hyperintensities: Insights from the UK Biobank

Yigizie Yeshaw^{1,2,3,4}  | Iqbal Madakkattel^{1,2,3}  | Anwar Mulugeta^{1,2,3,5}  |
 Amanda Lumsden^{1,2,3}  | Elina Hypponen^{1,2,3} 

¹Australian Centre for Precision Health, University of South Australia, Adelaide, South Australia, Australia

²UniSA Clinical and Health Sciences, University of South Australia, Adelaide, South Australia, Australia

³South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia

⁴Department of Epidemiology and Biostatistics, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

⁵Department of Pharmacology and Clinical Pharmacy, College of Health Sciences, Addis Ababa, Ethiopia

Correspondence

Elina Hypponen, Australian Centre for Precision Health, University of South Australia, South Australian, Health & Medical Research Institute (SAHMRI) Level 8, GPO Box 2471, Adelaide, SA 5001, Australia.
 Email: elina.hypponen@unisa.edu.au

Funding information

Medical Research Future Fund, Grant/Award Number: MRF2007431; National Health and Medical Research Council Leadership Investigator Award, Grant/Award Number: GNT2025349

Abstract

INTRODUCTION: Brain white matter hyperintensities (WMHs) reflect the risks of stroke, dementia, and overall mortality.

METHODS: We used a hypothesis-free gradient boosting decision tree (GBDT) approach and conventional statistical methods to discover risk factors associated with volume of WMHs. The GBDT models considered data on 2891 input features, collected ~10 years prior to volume of WMH measurements from 44,053 participants. Top 3% of features, ranked by Shapley values, were taken forward to epidemiological analyses using linear regression.

RESULTS: Adiposity, lung function, and indicators of metabolic health (eg, glycated hemoglobin, hypertension, alkaline phosphatase, microalbumin, and urate) contribute to WMH prediction. Of lifestyle factors, smoking had the strongest association. Time spent outdoors, creatinine, and several red blood cell indices were among the identified less-known predictors of WMHs.

CONCLUSIONS: Obesity, high blood pressure, lung function, metabolic abnormalities, and lifestyle are key contributors to WMHs, providing opportunities to prevent or reduce their development.

KEYWORDS

machine learning, risk factors, UK Biobank, white matter hyperintensities

Highlights

- Obesity and related metabolic abnormalities were linked with WMHs.
- Associations with time spent outdoors, creatinine, some red blood cell indices and height were among the less-known risk factors identified.
- Action on blood pressure, metabolic abnormalities, and adequate oxygenation may help to prevent WMHs.
- Biomarker links may suggest simple blood tests could aid in early dementia prediction.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* published by Wiley Periodicals, LLC on behalf of Alzheimer's Association.

1 | BACKGROUND

White matter hyperintensities (WMHs) are regions within white matter that exhibit brighter signal intensity than that of the surrounding white matter on a fluid-attenuated inversion recovery (FLAIR) MRI.¹ WMHs are one of the common imaging markers of cerebral small vessel disease.² They have been the focus of many recent neuroimaging studies due to their increasing prevalence with age and evidence supporting significant predictive implications for diseases such as stroke, dementia, and overall mortality.³ Though the precise mechanisms underlying the development of WMHs are not fully understood, they may be linked to various factors, such as reduced cerebral blood flow, neuroinflammation, gliosis, and/or loss of axons and myelin.⁴ Alongside age, modifiable cerebrovascular risk factors, such as obesity, hypertension, and abnormal glucose metabolism, and lifestyle factors, including insufficient physical activity, alcohol intake, and smoking, have been associated with WMHs.⁵

Given the role of WMHs as an intermediate biomarker for cerebrovascular health, understanding risk factors affecting their development can provide important opportunities for the prevention of cognitive impairment, dementia, and stroke.⁶ With access to large-scale neuroimaging data and extensive risk factor information from the UK Biobank (UKB), we now have a unique opportunity to search for risk factors affecting the development of WMHs. In this study we use machine learning (ML), specifically the gradient boosting decision tree (GBDT) algorithm, to construct a more comprehensive understanding of the risk factors that are potentially linked with volume of WMHs.

Unlike traditional statistical approaches, which can be limited in their ability to model complex relationships, the GBDT method excels in capturing intricate interactions, non-linear relationships, and hierarchical effects among predictors. This is particularly advantageous in large-scale epidemiological studies, where challenges such as data missingness, multicollinearity, and high dimensionality often impede the utility of conventional methods. By leveraging GBDT's ability to iteratively refine predictive models, we can effectively screen thousands of potential predictors and unravel patterns that might remain obscured using traditional techniques. Thus, our approach not only addresses key methodological challenges but also provides novel insights into the multifactorial nature of WMHs, advancing the field beyond the scope of standard epidemiological frameworks.⁷

2 | MATERIALS AND METHODS

2.1 | Data source and study participants

We used information from the UKB, which is a large-scale prospective epidemiological cohort containing in-depth genetic and health information from over half a million participants residing in England, Wales, and Scotland. All were aged 37 to 73 years at the time of recruitment, which was conducted between March 2006 and July 2010.⁸ Data collection took place across 22 assessment centers and included physical measurements, blood and urine sampling, touchscreen questionnaires,

RESEARCH IN CONTEXT

- 1. Systematic review:** We reviewed the literature on risk factors for WMHs using PubMed. Previous studies have typically focused on a limited number of exposures, often involving small to moderate sample sizes and conventional epidemiological methods. This study aimed to discover risk factors for WMHs from a wide range of exposures included in the UK Biobank.
- 2. Interpretation:** Our analysis support and extend from characteristics previously linked to WMHs, with related risks observed for older age, greater weight, smoking, and indicators related to frailty and metabolic health. Associations with several blood cell parameters and lung function support a role of oxygenation.
- 3. Future directions:** Many of the features associated with WMHs can be modified or treated, offering opportunities to prevention. Several blood biomarker associations suggest that even with relatively simple biomarker assessments, it may be possible to identify those at risk of WMH.

and interviews with the participants. Since the introduction of an imaging substudy in 2014, more than 60,000 participants have undergone MRI scans of brain, heart, and abdomen.⁹ This study was restricted to 44,053 participants who had valid information on volume of WMHs and excludes those who had dementia at baseline ($N = 19$) or who had outlier WMH volume measurements (outside the range of ± 3 standard deviations from the mean, $N = 941$) (Figure 1).

2.2 | Volume of WMHs

Brain imaging data were acquired using a Siemens Skyra 3T scanner with a standard Siemens 32-channel RF receive head coil (VD13A SP4). After acquisition, the MRI data underwent central processing at the UKB to eliminate artifacts, align images across modalities and individuals, and generate valuable image-derived phenotypes (IDPs). Volume of WMHs, one of the IDPs, was automatically segmented using both T1 and T2 FLAIR as input in the Brain Intensity Abnormality Classification Algorithm (BIANCA) tool. Details of imaging and further image processing-related information can be found in previous studies.¹⁰

To account for variation in participant head size, we normalized volume of WMHs using the volumetric scaling factor derived from T1 head images to standard space (normalized volume of WMHs = volume of WMHs \times head size scaling factor).¹¹ Log-transformed volume of normalized WMHs (logWMH, in cubic centimeters [cm^3]) was used as an outcome to approximate a normal distribution for data. Data on volume of WMHs available up to April 2023 were considered in the analyses.

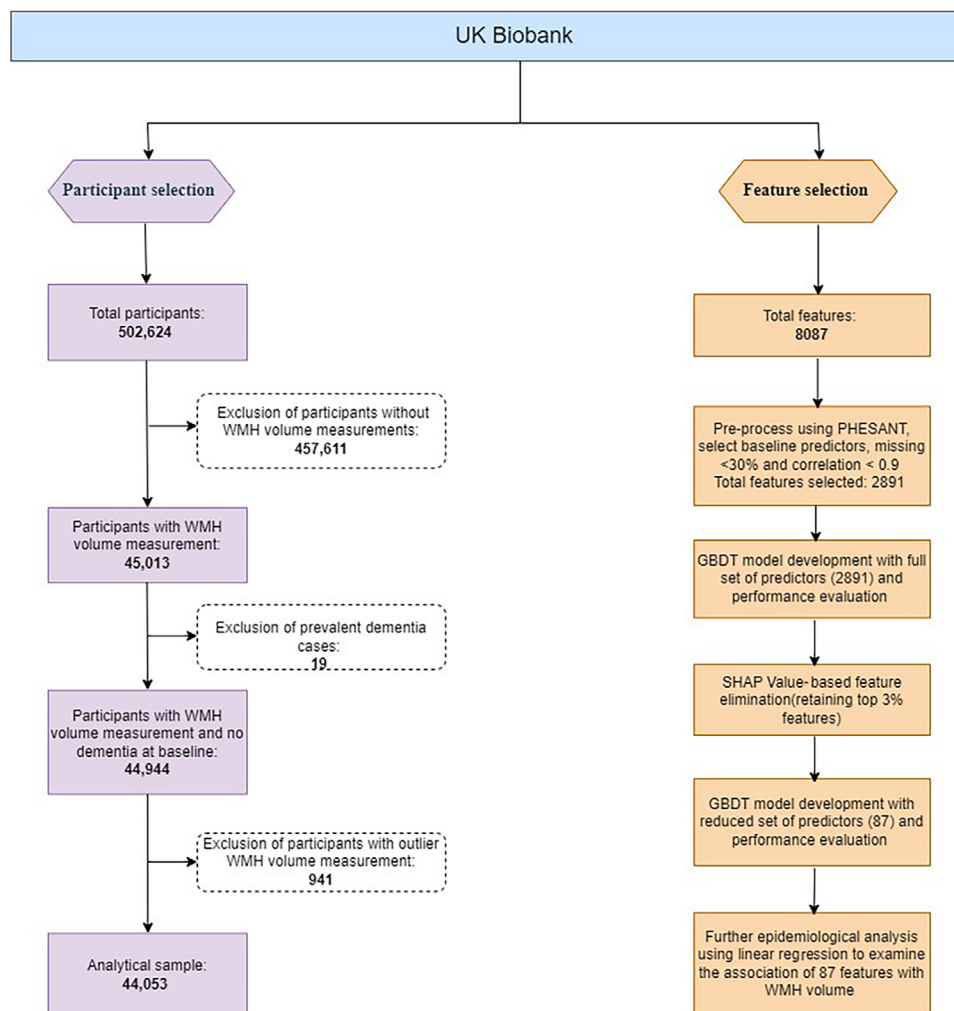


FIGURE 1 Analytical sample size determination and GBDT-SHAP machine learning pipeline for the study. GBDT, gradient boosting decision trees; PHESANT, PHENome scan analysis; SHAP, SHapley Additive exPlanation.

2.3 | Potential predictors

Baseline data encompassing comprehensive information from the touchscreen questionnaires, interviews, clinical assessments, and blood and urine samples were included in the screening for potential predictors of volume of WMHs. The data were pre-processed using the PHENome Scan Analysis Tool (PHESANT) to create dummy variables and remove negative numerical values that indicate missing data from specific categories.¹² A total of 2891 baseline phenotypic features were considered for our GBDT analyses after excluding closely correlated features ($|r| \geq 0.9$) and features with data available for less than 70% of the participants. Among the small sets of correlated features, the feature with the least missing values was considered for GBDT development. We categorized features into six broad categories: “baseline, personal, and sociodemographic characteristics,” “lifestyle and environment,” “physical measures,” “cognitive function and psychosocial factors,” “health and medical history,” and “biomarkers” for reporting purposes.

2.4 | Statistical analyses

2.4.1 | Identifying potential risk factors using GBDT-SHAP pipeline

The GBDT with SHapley Additive exPlanations (SHAP) ML pipeline⁷ was used to identify potential factors associated with volume of WMHs. We split the data into training, development, and test sets at a ratio of 60:20:20. The development set was used to mitigate overfitting.¹³ The test set was employed to report the performance metrics of the ML models. Using different random splits of the dataset into training, development, and test sets, we developed GBDT models and calculated feature importance (SHAP values) for a total of 150 times. SHAP values were calculated for each feature and each participant individually. To estimate the relative contribution of each feature to model prediction, we calculated the global feature importance by aggregating the mean of absolute SHAP values across all participants. To reduce variations arising from individual splits of data into training,

development, and test sets, we calculated the average of the feature importance. Finally, irrelevant predictors were eliminated by ranking features in the order of their SHAP values and by considering only the top 3% of features,¹⁴ yielding a reduced set of “important features” for our epidemiological analyses (Figure 1). Further details on GBDT and related data management are provided in the supplementary notes section of the [Supplementary Material](#). We used CatBoost version 0.21 and the SHAP package version 0.39.0 implemented in Python 3.10 for GBDT model development.

2.4.2 | Epidemiological analyses

Linear regression was used to examine the association between 87 potential predictors identified by the GBDT-SHAP pipeline and log volume of WMHs for 44,053 participants. The analyses were adjusted for basic confounders including age, sex, assessment center, ethnicity, education, employment, Townsend deprivation index, and duration until imaging (in years). All continuous features were divided into quintiles for the purpose of presenting the results. We also tested for the interaction by age (<65 years vs ≥65 years) and sex with other features by including relevant interaction terms in the models of the features and volume of WMHs. We checked for non-linearity of the association using quadratic terms of each continuous feature. To address the issue of multiple testing, we employed false discovery rate (FDR) correction.¹⁵ Effect estimates were presented as beta values along with their corresponding 95% confidence intervals (95% CI). The epidemiological analyses were conducted using STATA (version 17, StataCorp, College Station, TX, USA).

3 | RESULTS

3.1 | Participants

A total of 44,053 study participants were included in the analyses. The median follow-up time from the baseline assessment to volume of WMHs measurement was 10 years (interquartile range [IQR]: 8.30 to 12.12). Compared to others, volume of WMHs was higher among males, older people, participants with lower levels of education, and those from highly deprived areas (Table 1).

3.2 | Identifying potential risk factors

The average R^2 value of 150 iterations was 0.24 both for the model with all 2891 baseline features included and for the model limited to only the top 3% of features with the highest SHAP values, which means the additional variables included in the full model had minimal further explanatory power compared to the reduced model. This also reaffirmed the validity of our 3% cut-off point for selecting features for further epidemiological analyses.

TABLE 1 Volume of white matter hyperintensities (log cm³) by baseline characteristics.

Characteristics	N (%)	Volume of white matter hyperintensities, in log cm ³ Median (IQR)
Age		
<65 years	39,546 (89.77)	1.23 (0.63, 1.87)
≥65 years	4507 (10.23)	2.09 (1.52, 2.66)
p value*		1E-307
Sex		
Male	20,722 (47.04)	1.35 (0.72, 2.05)
Female	23,331 (52.96)	1.27 (0.67, 1.94)
p value*		5.420E-06
Ethnic background		
White European	42,606 (96.72)	1.32 (0.70, 2.00)
Asian	591 (1.34)	1.12 (0.60, 1.65)
Black African	284 (0.64)	1.21 (0.70, 1.75)
Other/mixed/unknown	572 (1.30)	1.18 (0.59, 1.83)
p value*		5.66E-05
Education		
None	2830 (6.42)	1.69 (1.11, 2.31)
NVQ/CSE/A levels	13,684 (31.06)	1.30 (0.69, 1.96)
Degree/professional	27,401 (62.20)	1.27 (0.66, 1.96)
Missing	138 (0.31)	1.58 (0.94, 2.20)
p value*		2.53E-133
Townsend deprivation index		
Highly deprived	25,307 (57.45)	1.33 (0.71, 2.01)
Less deprived	18,707 (42.46)	1.28 (0.68, 1.96)
Missing	39 (0.09)	1.01 (0.61, 1.33)
p value*		1.04E-25
Employment		
None	2654 (6.02)	1.20 (0.61, 1.86)
Retired	10,689 (24.26)	1.82 (1.23, 2.43)
1st quartile (lowest working hour)	6863 (15.58)	1.23 (0.62, 1.92)
2nd quartile	5057 (11.48)	1.13 (0.54, 1.75)
3rd quartile	9713 (22.05)	1.11 (0.53, 1.72)
4th quartile (highest working hour)	8607 (19.54)	1.12 (0.54, 1.76)
Missing	470 (1.07)	1.40 (0.74, 2.13)
p value*		1E-307

*, p values are from likelihood ratio test with all linear regression analyses including age, sex, and assessment center as covariates.

Of the top 3% of features with the highest SHAP values, “baseline, personal, and sociodemographic characteristics” accounted for 45%, with age being the major contributor (~33%). “Lifestyle and environment” contributed 7%, “physical measures” 19%, “biomarkers” 14%, “medical and family history” 14%, and the remaining 1% by “psychosocial and cognitive function” features. Of 509 self-reported non-cancer illnesses considered in the analyses, only multiple sclerosis, hypertension, and “number of non-cancer illnesses” were identified as potentially important predictors for WMHs. The list of features with their SHAP values are presented in [Supplementary Material, Tables S1 and S2](#)).

3.3 | Risk factors associated with volume of WMHs

Linear regression analyses of important features adjusted for age, sex, assessment center, ethnicity, education, employment, Townsend deprivation index, and duration until imaging are presented in [Supplementary Material, Table S3](#). Where sex interactions were detected, results stratified by sex are also presented. We also conducted sensitivity analyses excluding the MS cases ($n = 88$), finding identical results compared to those presented in [Supplementary Material, Table S3](#) ([Supplementary Material, Table S4](#)).

3.4 | Associations with physical measures

Several measures reflecting body size and composition, including shorter height, greater weight, and higher whole-body fat mass, were associated with higher volume of WMHs. Higher impedance of leg and arm was associated with lower volume of WMHs, as was greater hand grip strength. Indicators reflecting better lung function (peak expiratory flow rate, forced expiratory volume in 1 s [FEV1]) were associated with lower volume of WMHs. Volume of WMHs was higher with greater pulse rate and systolic and diastolic blood pressure (Figure 2).

3.5 | Associations with cognition- and medical condition-related features

Individuals with multiple sclerosis had volumes of WMHs that were 112% higher (β 1.12, 95% CI = 0.94 to 1.29) compared to others. A higher number of incorrect matches in cognitive tests, stress in last 2 years, long-standing illness/disability, and chronic diseases including hypertension were also associated with higher volume of WMHs (Figure 3). In contrast, volume of WMHs was lower in participants whose mothers and fathers remained alive at the time of the baseline survey, with further associations with family history of disease and volume of WMHs picked up by the participants' mother or siblings not having selected chronic illnesses.

3.6 | Associations with lifestyle and environment-related features

Of the lifestyle factors, smoking had the strongest associations (Figure 4). Current smokers (β 0.17, 95% CI = 0.14 to 0.21) and those who smoke on most/all days (β 0.20, 95% CI = 0.16 to 0.24) had higher volumes of WMHs compared to non-smokers. Participants in the highest quintile of coffee intake had higher volumes of WMHs compared to those in the lowest quintile. Conversely, there was some evidence that higher cereal intake was associated with lower volume of WMHs, although this potential benefit was not seen for participants in the highest intake group (Figure 4).

3.7 | Associations with biomarkers

Higher volume of WMHs was also associated with several biomarkers (Figure 5). Higher alkaline phosphatase, glycated hemoglobin (HbA1c), and glucose levels were associated with higher volume of WMHs. From cardiovascular biomarkers, triglycerides were directly linked with volume of WMHs, whereas HDL cholesterol had an inverse association. Higher levels of kidney biomarkers, including urinary microalbumin, total protein, urate, and urinary sodium, were associated with higher volume of WMHs. In contrast, blood levels of creatine and urea were associated with slower volume of WMHs. Having greater gamma glutamyltransferase, aspartate aminotransferase, and albumin was associated with higher volumes of WMHs. Insulin-like growth factor-1 and testosterone were inversely associated with volume of WMHs. Of the hematology-related biomarkers, high light scatter reticulocyte percentage, immature reticulocyte fraction, hematocrit percentage, and monocyte count were directly associated with volume of WMHs, whereas a U-shaped association was seen for mean spheroid cell volume and volume of WMHs.

4 | DISCUSSION

WMHs are linked with an increased risk of dementia, and they can be detected many years prior to the occurrence of clinical dementia symptoms. Hence, WMHs could serve as a proxy for further risk of dementia, providing potential to prioritize individuals for intensive screening or prevention strategies, including selection of participants to dementia prevention trials.¹⁶ Furthermore, identification of risk factors for this surrogate biomarker may provide new insight to help inform dementia prevention strategies by identifying modifiable risk factors that can be acted upon before the onset of the disease. This prospective study, which to our knowledge is based on the world's largest brain imaging data resource, was conducted to identify, from a comprehensive list of factors, those associated with volume of WMHs using a novel hypothesis-free ML method. Our ML analyses showed that some lifestyle factors, many measures of body composition, and general

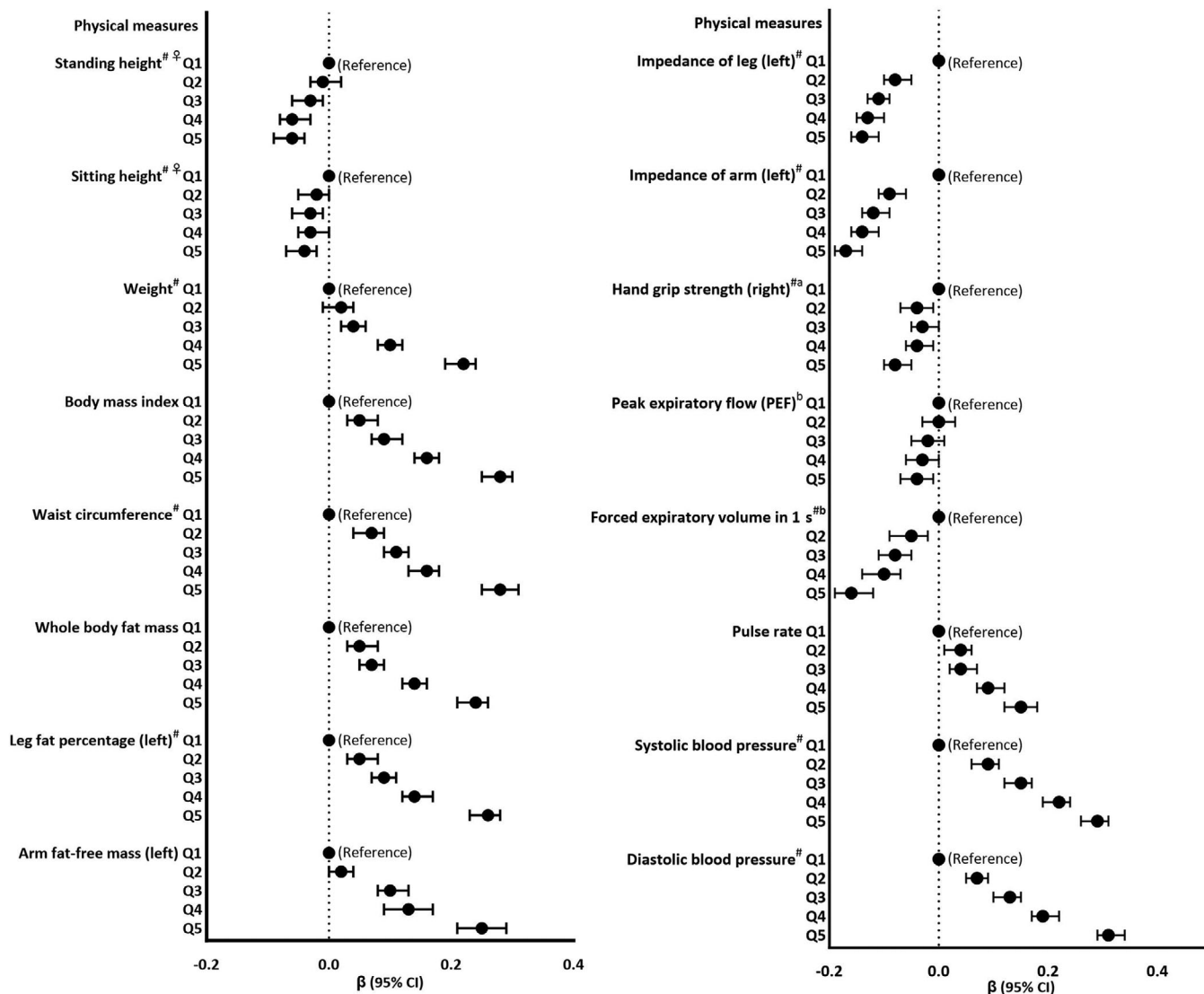


FIGURE 2 Forest plots for associations of SHAP important physical measures with volume of WMHs, β , and their 95% confidence interval (reference group as indicated) determined by linear regression. Analyses were adjusted for sex, age, assessment center, Townsend deprivation index, education, ethnicity, employment, and duration until imaging. Q1 to Q5 represent quintiles from lowest to highest values. a: Right-hand grip further adjustment for height and weight. b: FEV1 and PEF were further adjusted for height. #: the variable had interaction with sex but with directionally consistent estimates unless presented with a gender sign, ♀: the association only in females. β , beta; CI, confidence interval; FEV1, forced expiratory volume in 1 s; PEF, peak expiratory flow; SHAP, SHapley Additive explanation; WMH, white matter hyperintensity.

fitness were associated with volume of WMHs. While it may be more difficult to modify some of the strong disease associations, such as that with MS, an action to effectively manage hypertension and metabolic imbalances, reflected by the many associations with biomarkers, may provide opportunities to prevent or reduce the development of WMHs.

Our analyses strongly suggest that obesity and related metabolic abnormalities, including hypertension, hyperlipemia, higher HbA1c and glucose, and non-alcoholic fatty liver disease,¹⁷ may play a key role in the development of WMHs. The links between hypertension and other metabolic factors with WMHs may be explained by vascular remodeling and arteriosclerosis, leading to brain tissue ischemia, oxidative stress, and neuroinflammation.¹⁸ It is also possible that the association between urinary sodium and volume of WMHs is related to hyperten-

sion, either through a link with related medications (diuretics) or a high salt diet, which may increase blood pressure.¹⁹ We also observed a link between shorter body height and greater WMH volume in females but not in males. Short stature has been linked to a greater risk of cardiovascular diseases,²⁰ particularly in females, which could contribute to a higher burden of WMHs in this group. For example, a systematic review and meta-analysis demonstrated an association between short stature and an increased risk of type 2 diabetes in females.²¹

Unsurprisingly, indicators related to frailty and longstanding illness were observed to associate with greater volume of WMHs. For example, consistent with a previous UKB study,²² we observed an association between greater hand grip strength (a predictive biomarker of whole-body muscle strength and better nutritional status)²³ and

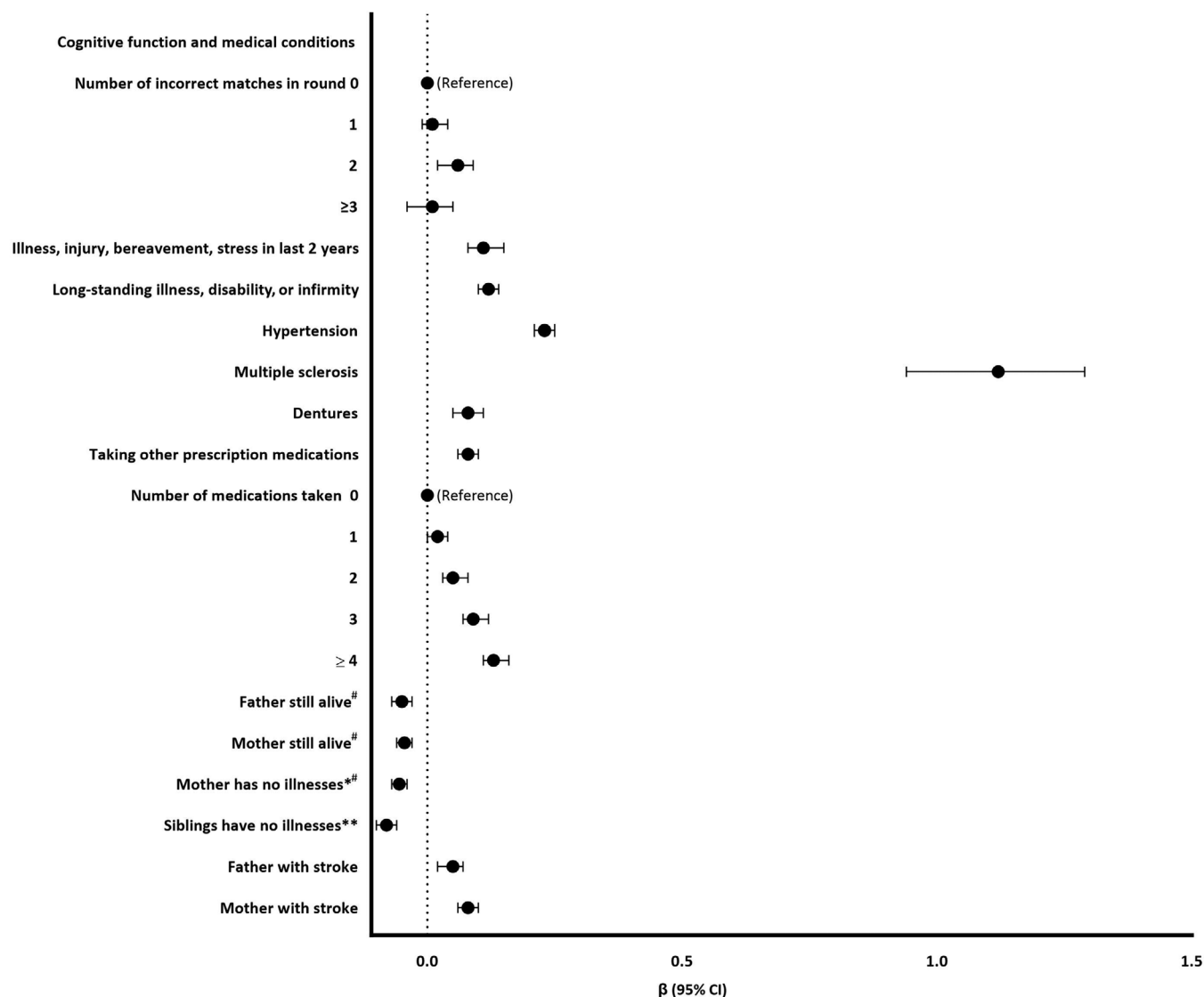


FIGURE 3 Forest plots for associations of SHAP important cognitive and medical condition related features with volume of white matter hyperintensities, β , and their 95% confidence interval (reference group as indicated) determined by linear regression. Analyses were adjusted for sex, age, assessment center, Townsend deprivation index, education, ethnicity, employment, and duration until imaging. Q1 to Q5 represent quintiles from lowest to highest values. *Mother has no Parkinson's disease, severe depression, lung cancer, bowel cancer, or breast cancer. **Siblings have no heart disease, stroke, hypertension, chronic bronchitis/emphysema, dementia, or diabetes. #: the variable has interaction with sex but with directionally consistent estimates. β , beta; CI, confidence interval; SHAP, SHapley Additive explanation; WMH, white matter hyperintensity.

lower volume of WMHs. This may seem to contradict the association between greater fat-free mass in the arm (ie, muscle mass) and higher volume of WMHs. However, as fat-free mass represents highly respiratory tissue, it is possible that greater muscle mass increases volume of WMHs due to oxidants produced as a byproduct of respiration. Interestingly, creatine, a compound derived both from diet and endogenous synthesis that helps provide energy to muscles, has been linked both to greater grip strength²⁴ and lower oxidative stress.²⁵ While we did not have a direct measure of creatine in our study, a higher serum level of creatinine – the waste product of creatine usage – was associated with lower volume of WMHs, consistent with the inverse association between grip strength and volume of WMHs.

Several kidney-related biomarkers were also associated the development of WMHs. Higher microalbumin was markedly positively associated with volume of WMHs. Microalbuminuria, a symptom of kidney dysfunction, is the aberrant leakage of albumin from the blood into the urine, usually due to damage to nephron microvasculature. Observationally, bidirectional associations have been observed between WMHs and poor renal function,²⁶ suggesting that the detection of microalbumin by the GBDT approach may signify poor microvascular integrity that may apply also to the brain. Renal function is also important for regulating blood pressure, offering a possible volume of a WMH-increasing mechanism downstream of kidney dysfunction. Having higher urate in the blood, often associated with kidney dysfunction

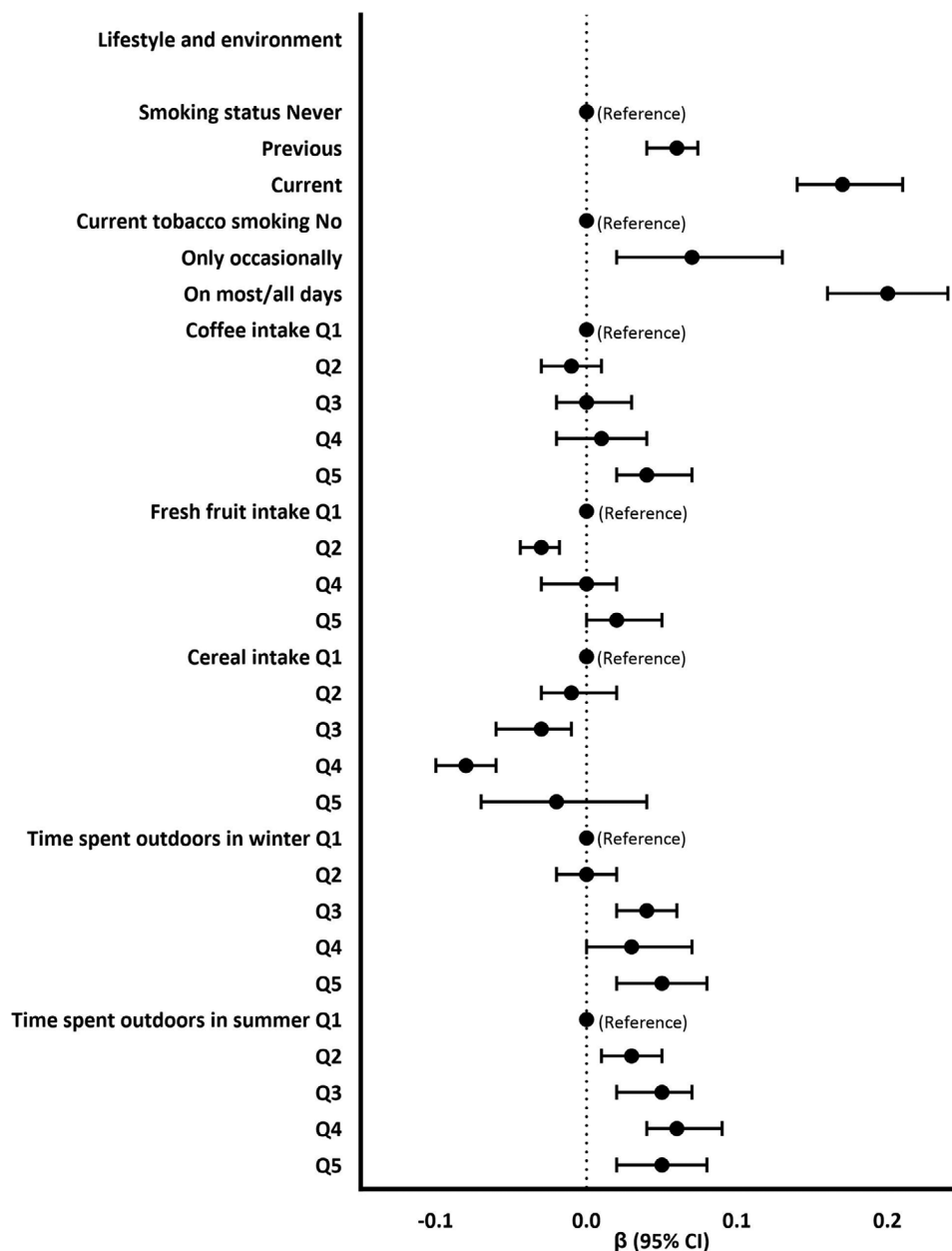


FIGURE 4 Forest plots for associations of SHAP important lifestyle and environment-related features with volume of WMHs, β , and their 95% confidence interval (reference group as indicated) determined by linear regression. Analyses were adjusted for sex, age, assessment center, Townsend deprivation index, education, ethnicity, employment, and duration until imaging. Q1 to Q5 represent quintiles from lowest to highest values. β , beta; CI, confidence interval; SHAP, SHapley Additive explanation; WMH, white matter hyperintensity.

and metabolic diseases,²⁷ was also associated with a greater volume of WMHs.

Several features related to smoking and lung function were deemed important predictors of volume of WMHs. The observed association of smoking with greater volume of WMHs is congruent with a previous study²⁸ and could be due to its links with ischemic brain injury, hypoperfusion, endothelial dysfunction, blood-brain barrier breakdown, inflammation, oxidative stress, hypoxia, and neuronal and glial degeneration.²⁹ Cigarette smoking is also a risk factor for chronic diseases such as diabetes and non-alcoholic fatty liver disease^{30,31} that could further contribute to risk of WMHs. Consistent with the detri-

mental effects of smoking on lung health, indicators reflecting poorer lung function were also picked up by the GBDT method and were associated with greater volume of WMHs, most notably lower FEV1, which was linked to 16% greater volume of WMHs. Such associations were reported in a previous study³² and may reflect the benefit of well-functioning lungs in providing adequate oxygen and preventing ischemic brain injury.⁴ Coffee intake was deemed an important predictive trait, with high intake associating with a greater volume of WMHs, consistent with previous findings.³³ Prolonged heavy consumption of caffeinated coffee reduces cerebral blood flow, resulting in chronic cerebral hypoperfusion, which in turn could contribute to

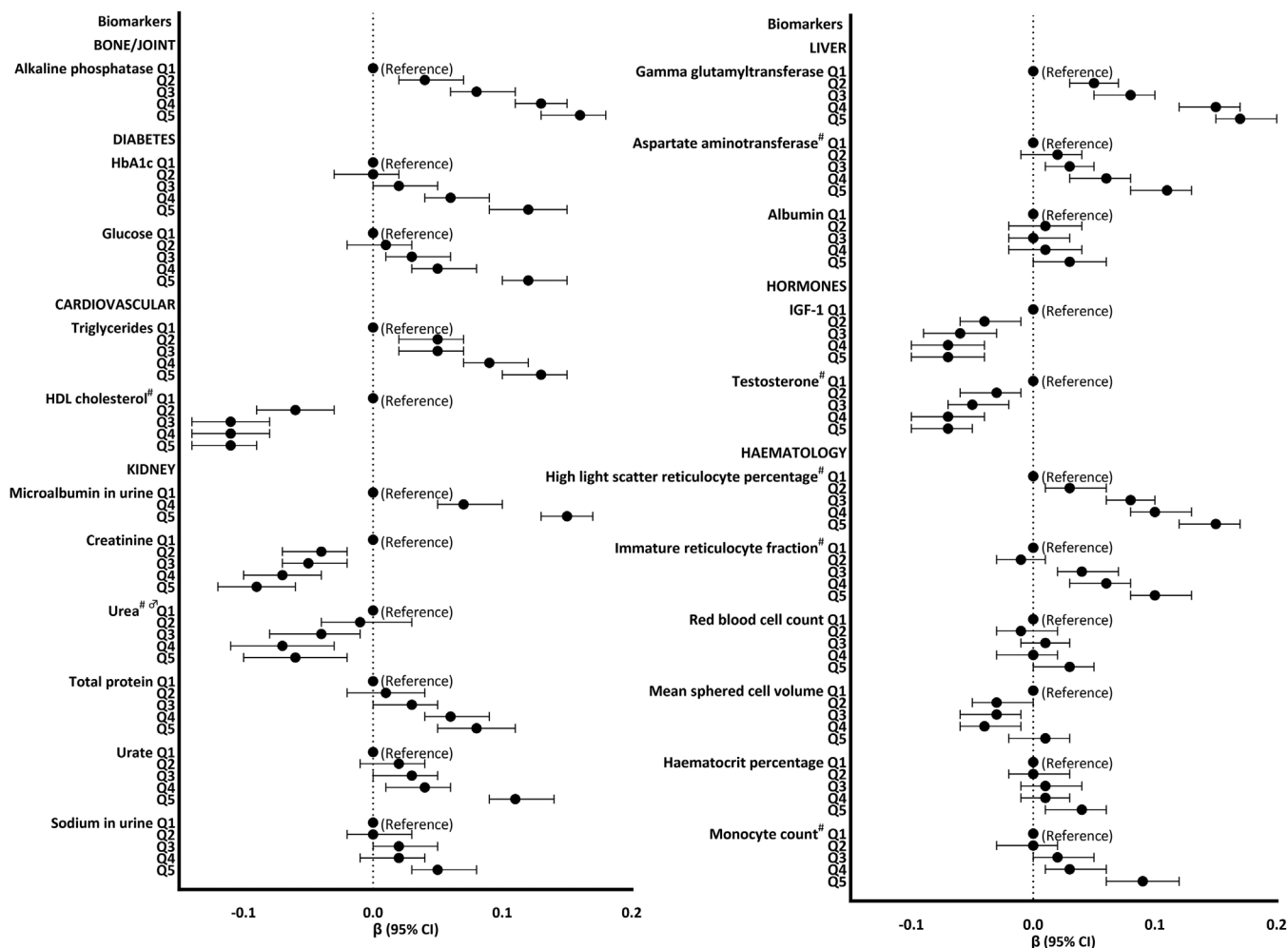


FIGURE 5 Forest plots for associations of SHAP important biomarkers with volume of WMHs, β , and their 95% confidence interval (reference group as indicated) determined by linear regression. Analyses were adjusted for sex, age, assessment center, Townsend deprivation index, education, ethnicity, employment, and duration until imaging. Q1 to Q5 represent quintiles from lowest to highest values. Since the levels of microalbumin in urine are not evenly distributed into five quintiles (68% had the lowest measurement), only Q1 (68%), Q4 (12%), and Q5 (20%) were reported. #: the variable had interaction with sex but with directionally consistent estimates unless presented with a gender sign, δ : the association only in males. β , beta; CI, confidence interval; HbA1c, glycated hemoglobin, IGF-1, insulin-like growth factor 1; SHAP, SHapley Additive explanation; WMH, white matter hyperintensity.

greater volume of WMHs.³⁴ We also found some evidence that higher cereal intake may be associated with lower volume of WMHs, consistent with an earlier study.³⁵ While the association did not follow a dose–response pattern, it is possible that the association between cereal intake and WMHs reflects the benefits of consuming a healthy diet.³⁶ There is also evidence suggesting that celiac disease, an autoimmune disorder triggered by an abnormal adaptive immune response against gluten-containing grains (which may be reflected by lower cereal intake), is associated with greater focal white-matter lesions,³⁷ higher risk of vascular dementia,³⁸ and cognitive deficit.³⁹

Interestingly, spending more time outdoors (in summer and in winter) was found to be associated with a greater volume of WMHs. While more work is necessary to understand this relationship, it is in line with another UKB study reporting increased risk of myocardial infarction among individuals that spend more time outdoors.⁴⁰ There is also evidence indicating that individuals who work outdoors often

show elevated blood pressure levels related to greater exposure to air pollution and noise levels resulting from urbanization.^{41,42} Hence, the indirect roles of environmental attributes associated with outdoor activities could explain the observed association between time spent outdoors and volume of WMHs in this study.

WMHs are commonly observed in patients with MS,⁴³ and, in line with this, MS was associated with substantially greater volume of WMHs in our study. In MS, the immune system non-selectively attacks the body's own myelin antigens, leading to infiltration of lymphocytes, microglial activation, demyelination, and axonal degeneration, mechanisms that are believed to relate to the development of WMHs in MS.⁴⁴ Interestingly, higher total serum protein was associated with greater volume of WMHs, increasing gradually across quintiles. As the predominant serum protein, albumin, was only marginally associated with higher volume of WMHs, it is possible that the total protein association reflects high serum immunoglobulin levels, consistent with the

associations of autoimmune diseases such as MS with greater WMH lesions.⁴⁵

Several blood cell markers, including many red blood cell-related traits, were also picked up as important features for volume of WMH prediction in our study. The presence of high levels of reticulocytes may indicate chronic hypoxia or systemic inflammation, both of which can contribute to endothelial dysfunction.⁴⁶ Higher monocyte count, a marker of inflammation, was also associated with higher volume of WMHs in this study, in line with a previously reported association between monocyte count and cerebral small vessel disease.⁴⁷

An important strength of our study is the use of a novel ML pipeline, which was able to select important features from among thousands of included predictors and in the context of interactions and non-linear associations, where traditional models often face limitations. The use of UKB, the most comprehensive large-scale dataset of its kind, enabled us to explore a wide range of potential predictors to comprehensively investigate risk factors linked to volume of WMHs. However, the observational nature of the study did not allow us to establish causality of association, confirm clinical relevance, or exclude presence of residual confounding or reverse causality. Though objective ways of data collection methods such as clinical examinations and blood sampling were applied to collect many of the data items, some of the data were collected through interviews and questionnaires, which may be affected by reporting or recall bias.

Compared to the general population, participants included in the UKB are healthier,⁴⁸ which may introduce healthy volunteer bias. Due to the low response rate of UKB, generalization of our findings to the whole UK population may not be possible. However, a meta-analysis comparing UKB with 18 other studies of acceptable response rates highlighted that, despite a very low response rate, risk factor associations in the UKB appear to be generalizable and consistent with these comparator studies.⁴⁹

In conclusion, our analyses showed that some lifestyle factors, biomarkers and many measures of body composition, metabolic health, and general fitness are associated with volume of WMHs. Many of the modifiable influences on volume of WMHs and dementia risk are shared with other chronic conditions, highlighting the importance of healthy lifestyles and effective management of metabolic health. Further research is needed to establish causal relationships and develop effective intervention strategies based on these insights.

ACKNOWLEDGMENTS

We thank all UK Biobank participants and people involved in the establishment and managing of the cohort resource. This project was conducted under project number 89630. The study was in part supported by a Research Training Program international scholarship from the Government of Australia and the Medical Research Future Fund (MRF2007431), Australia. EH was funded by a National Health and Medical Research Council Leadership Investigator Award (GNT2025349). The funders were not involved in the study's design or conduct, data handling, analysis, manuscript writing, or decision to submit for publication.

CONFLICT OF INTEREST STATEMENT

The authors report no competing interests. Author disclosures are available in the [Supporting Information](#).

CONSENT STATEMENT

The UKB obtained ethical approval from the National Information Governance Board for Health and Social Care and the Northwest Multicentre Research Ethics Committee under reference 11/NW/0382. Participants provided electronic consent during data collection to utilize their anonymous data and to access their medical records for health-related research purposes.⁵⁰

DATA AVAILABILITY STATEMENT

Approved users of the UK Biobank can access the data used for this research upon application.

ORCID

Yigizie Yeshaw  <https://orcid.org/0000-0003-4587-7925>

Iqbal Madakkattel  <https://orcid.org/0000-0003-2339-5917>

Anwar Mulugeta  <https://orcid.org/0000-0002-8018-3454>

Amanda Lumsden  <https://orcid.org/0000-0002-0214-6498>

Elina Hyppönen  <https://orcid.org/0000-0003-3670-9399>

REFERENCES

- Yoshita M, Fletcher E, Harvey D, et al. Extent and distribution of white matter hyperintensities in normal aging, MCI, and AD. *Neurology*. 2006;67(12):2192. doi:10.1212/01.wnl.0000249119.95747.1f
- Pantoni L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *Lancet Neurol*. 2010;9(7):689-701. doi:10.1016/s1474-4422(10)70104-6
- Debette S, Schilling S, Duperron MG, Larsson SC, Markus HS. Clinical significance of magnetic resonance imaging markers of vascular brain injury: a systematic review and meta-analysis. *JAMA Neurol*. 2019;76(1):81-94. doi:10.1001/jamaneurol.2018.3122
- Humphreys CA, Smith C, Wardlaw JM. Correlations in post-mortem imaging-histopathology studies of sporadic human cerebral small vessel disease: a systematic review. *Neuropathol Appl Neurobiol*. 2021;47(7):910-930. doi:10.1111/nan.12737
- Lin K, Wen W, Lipnicki DM, et al. Risk factors and cognitive correlates of white matter hyperintensities in ethnically diverse populations without dementia: the COSMIC consortium. *Alzheimers Dement*. 2024;16(1):e12567. doi:10.1002/dad2.12567
- Wardlaw JM, Muñoz-Maniega S. What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. *J Am Heart Assoc*. 2015;4(6):001140. doi:10.1161/jaha.114.001140
- Madakkattel I, Zhou A, McDonnell MD, Hyppönen E. Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. *Sci Rep*. 2021;11(1):22997. doi:10.1038/s41598-021-02476-9
- Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z
- Help us to complete the world's largest imaging study. UK Biobank; 2024. Accessed February 11, 2025. <https://www.ukbiobank.ac.uk/explore-your-participation/contribute-further/imaging-study>
- Miller KL, Alfaro-Almagro F, Bangerter NK, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*. 2016;19(11):1523-1536. doi:10.1038/nn.4393

11. Smith SM, Alfaro-Almagro F, Miller KL. UK Biobank Brain Imaging Documentation. UK Biobank; 2024. Accessed February 11, 2025. https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf
12. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software application profile: pHEsANT: a tool for performing automated phenotype scans in UK Biobank. *Int J Epidemiol*. 2018;47(1):29-35. doi:10.1093/ije/dyx204
13. Maldonado S, Weber R, Famili F. Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Inf Sci*. 2014;286:228-246. doi:10.1016/j.ins.2014.07.015
14. Bolón-Canedo V, Sánchez-Marroño N, Alonso-Betanzos A. Feature selection for high-dimensional data. *Prog Artif Intell*. 2016;5:65-75.
15. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995;57(1):289-300.
16. d'Arbeloff T, Elliott ML, Knodt AR, et al. White matter hyperintensities are common in midlife and already associated with cognitive decline. *Brain Commun*. 2019;1(1). doi:10.1093/braincomms/fcz041
17. Bedogni G. Fatty Liver Index. MDCalc. Accessed January 11, 2025. <https://www.mdcalc.com/calc/10001/fatty-liver-index>
18. Evans LE, Taylor JL, Smith CJ, Pritchard HAT, Greenstein AS, Allan SM. Cardiovascular comorbidities, inflammation, and cerebral small vessel disease. *Cardiovasc Res*. 2021;117(13):2575-2588. doi:10.1093/cvr/cvab284
19. Kim KW, Seo H, Kwak MS, Kim D. Visceral obesity is associated with white matter hyperintensity and lacunar infarct. *Int J Obesity*. 2017;41(5):683-688. doi:10.1038/ijo.2017.13
20. Paajanen TA, Oksala NKJ, Kuukasjärvi P, Karhunen PJ. Short stature is associated with coronary heart disease: a systematic review of the literature and a meta-analysis. *Eur Heart J*. 2010;31(14):1802-1809. doi:10.1093/eurheartj/ehq155
21. Janghorbani M, Momeni F, Dehghani M. Hip circumference, height and risk of type 2 diabetes: systematic review and meta-analysis. *Obes Rev*. 2012;13(12):1172-1181. doi:10.1111/j.1467-789X.2012.01030.x
22. Firth JA, Smith L, Sarris J, et al. Handgrip strength is associated with hippocampal volume and white matter hyperintensities in major depression and healthy controls: a UK Biobank study. *Psychosom Med*. 2020;82(1):39-46. doi:10.1097/psy.0000000000000753
23. Bohannon RW. Grip strength: an indispensable biomarker for older adults. *Clin Interv Aging*. 2019;14:1681-1691. doi:10.2147/cia.S194543
24. Stout JR, Sue Graves B, Cramer JT, et al. Effects of creatine supplementation on the onset of neuromuscular fatigue threshold and muscle strength in elderly men and women (64 - 86 years). *J Nutr Health Aging*. 2007;11(6):459-464.
25. Arazi H, Eghbali E, Suzuki K. Creatine supplementation, physical exercise and oxidative stress markers: a review of the mechanisms and effectiveness. *Nutrients*. 2021;13(3):869. doi:10.3390/nu13030869
26. Wei CS, Yan CY, Yu XR, et al. Association between white matter hyperintensities and chronic kidney disease: a systematic review and meta-analysis. *Front Med*. 2022;9:770184. doi:10.3389/fmed.2022.770184
27. Soltani Z, Rasheed K, Kapusta DR, Reisin E. Potential role of uric acid in metabolic syndrome, hypertension, kidney injury, and cardiovascular diseases: is it time for reappraisal?. *Curr Hypertens Rep*. 2013;15(3):175-181. doi:10.1007/s11906-013-0344-5
28. Gray JC, Thompson M, Bachman C, Owens MM, Murphy M, Palmer R. Associations of cigarette smoking with gray and white matter in the UK Biobank. *Neuropsychopharmacology*. 2020;45(7):1215-1222. doi:10.1038/s41386-020-0630-2
29. Csordas A, Bernhard D. The biology behind the atherothrombotic effects of cigarette smoke. *Nat Rev Cardiol*. 2013;10(4):219-230. doi:10.1038/nrcardio.2013.8
30. Ng R, Sutradhar R, Yao Z, Wodchis WP, Rosella LC. Smoking, drinking, diet and physical activity—modifiable lifestyle risk factors and their associations with age to first chronic disease. *Int J Epidemiol*. 2019;49(1):113-130. doi:10.1093/ije/dyz078
31. Akhavan Rezaat A, Dadgar Moghadam M, Ghasemi Nour M, et al. Association between smoking and non-alcoholic fatty liver disease: a systematic review and meta-analysis. *SAGE Open Med*. 2018;6:2050312117745223. doi:10.1177/2050312117745223
32. Frenzel S, Bis JC, Gudmundsson EF, et al. Associations of pulmonary function with MRI brain volumes: a coordinated multi-study analysis. *J Alzheimers Dis*. 2022;90(3):1073-1083. doi:10.3233/jad-220667
33. Park J, Han JW, Lee JR, et al. Association between lifetime coffee consumption and late life cerebral white matter hyperintensities in cognitively normal elderly individuals. *Sci Rep*. 2020;10(1):421. doi:10.1038/s41598-019-57381-z
34. Addicott MA, Yang LL, Peiffer AM, et al. The effect of daily caffeine use on cerebral blood flow: how much caffeine can we tolerate? *Hum Brain Mapp*. 2009;30(10):3102-3114. doi:10.1002/hbm.20732
35. Song S, Gaynor AM, Cruz E, et al. Mediterranean diet and white matter hyperintensity change over time in cognitively intact adults. *Nutrients*. 2022;14(17). doi:10.3390/nu14173664
36. Dèdelé A, Bartkutė Ž, Chebotarova Y, Miškinytė A. The relationship between the healthy diet index, chronic diseases, obesity and lifestyle risk factors among adults in Kaunas City, Lithuania *Front Nutr*. 2021;8:599567. doi:10.3389/fnut.2021.599567
37. Kieslich M, Errázuriz G, Posselt HG, Moeller-Hartmann W, Zanella F, Boehles H. Brain white-matter lesions in celiac disease: a prospective study of 75 diet-treated patients. *Pediatrics*. 2001;108(2):E21. doi:10.1542/peds.108.2.e21
38. Lebowitz B, Luchsinger JA, Freedberg DE, Green PH, Ludvigsson JF. Risk of dementia in patients with celiac disease: a population-based cohort study. *J Alzheimers Dis*. 2016;49(1):179-185. doi:10.3233/jad-150388
39. Croall ID, Sanders DS, Hadjivassiliou M, Hoggard N. Cognitive deficit and white matter changes in persons with celiac disease: a population-based study. *Gastroenterology*. 2020;158(8):2112-2122. doi:10.1053/j.gastro.2020.02.028
40. Miguet M, Venetis S, Rukh G, Lind L, Schiöth HB. Time spent outdoors and risk of myocardial infarction and stroke in middle and old aged adults: results from the UK Biobank prospective cohort. *Environ Res*. 2021;199:111350. doi:10.1016/j.envres.2021.111350
41. Santos UP, Ferreira Braga AL, Bueno Garcia ML, et al. Exposure to fine particles increases blood pressure of hypertensive outdoor workers: a panel study. *Environ Res*. 2019;174:88-94. doi:10.1016/j.envres.2019.04.021
42. Tomei F, Ricci S, Giammichele G, et al. Blood pressure in indoor and outdoor workers. *Environ Toxicol Pharmacol*. 2017;55:127-136. doi:10.1016/j.etap.2017.06.022
43. Welton T, Kent D, Constantinescu CS, Auer DP, Dineen RA. Functionally relevant white matter degradation in multiple sclerosis: a tract-based spatial meta-analysis. *Radiology*. 2015;275(1):89-96. doi:10.1148/radiol.14140925
44. Rovira A, Barkhof F, Jäger R, Thurnher M. Multiple sclerosis and variants. *Clinical Neuroradiology*. 2018:1-41.
45. Currie S, Hadjivassiliou M, Clark MJ, et al. Should we be 'nervous' about coeliac disease? Brain abnormalities in patients with coeliac disease referred for neurological opinion. *J Neurol Neurosurg Psychiatry*. 2012;83(12):1216-1221. doi:10.1136/jnnp-2012-303281
46. Janaszak-Jasiecka A, Siekierzycka A, Płoska A, Dobrucki IT, Kalinowski L. Endothelial dysfunction driven by hypoxia-the influence of oxygen deficiency on bioavailability. *Biomolecules*. 2021;11(7). doi:10.3390/biom11070982
47. Noz MP, Wiegertjes K, et al. Trained immunity characteristics are associated with progressive cerebral small vessel disease. *Stroke*. 2018;49(12):2910-2917. doi:10.1161/strokeaha.118.023192
48. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with

- those of the general population. *Am J Epidemiol.* 2017;186(9):1026-1034. doi:[10.1093/aje/kwx246](https://doi.org/10.1093/aje/kwx246)
49. Batty GD, Gale CR, Kivimäki M, Deary IJ, Bell S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ.* 2020;368:m131. doi:[10.1136/bmj.m131](https://doi.org/10.1136/bmj.m131)
50. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12(3):e1001779. doi:[10.1371/journal.pmed.1001779](https://doi.org/10.1371/journal.pmed.1001779)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yeshaw Y, Madakkattel I, Mulugeta A, Lumsden A, Hypponen E. Machine learning to discover factors predicting volume of white matter hyperintensities: Insights from the UK Biobank. *Alzheimer's Dement.* 2025;17:e70090. <https://doi.org/10.1002/dad2.70090>