RESEARCH ARTICLE

Mathematical modeling for the prediction of cerebral white matter lesions based on clinical examination data

Yuya Shinkawa[1], **Takashi Yoshida**[2¤], **Yohei Onaka**[2], **Makoto Ichinose**[2], **Kazuo Ishii**[3*]

1 Kurume University Graduate School of Medicine, Kurume, Fukuoka, Japan, **2** Shin Takeo Hospital, Takeo, Saga, Japan, **3** Biostatistics Center, Kurume University, Kurume, Fukuoka, Japan

¤ Current address: Tokyo Shinagawa Hospital, Tokyo, Japan
* ishii_kazuo@med.kurume-u.ac.jp, kazuoishii2014@gmail.com

OPEN ACCESS

Citation: Shinkawa Y, Yoshida T, Onaka Y, Ichinose M, Ishii K (2019) Mathematical modeling for the prediction of cerebral white matter lesions based on clinical examination data. PLoS ONE 14(4): e0215142. https://doi.org/10.1371/journal.pone.0215142

Editor: Antonio Palazón-Bru, Universidad Miguel Hernandez de Elche, SPAIN

Received: January 19, 2019

Accepted: March 26, 2019

Published: April 16, 2019

Data Availability Statement: All relevant data are available from the Dryad repository doi:10.5061/dryad.73bh2q8 as Data files: WM_data.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Cerebral white matter lesions are ischemic symptoms caused mainly by microangiopathy; they are diagnosed by MRI because they show up as abnormalities in MRI images. Because patients with white matter lesions do not have any symptoms, MRI often detects the lesions for the first time. Generally, head MRI for the diagnosis and grading of cerebral white matter lesions is performed as an option during medical checkups in Japan. In this study, we develop a mathematical model for the prediction of white matter lesions using data from routine medical evaluations that do not include a head MRI. Linear discriminant analysis, logistic discrimination, Naive Bayes classifier, support vector machine, and random forest were investigated and evaluated by ten-fold cross-validation, using clinical data for 1,904 examinees (988 males and 916 females) from medical checkups that did include the head MRI. The logistic regression model was selected based on a comparison of accuracy and interpretability. The model variables consisted of age, gender, plaque score (PS), LDL, systolic blood pressure (SBP), and administration of antihypertensive medication (odds ratios: 2.99, 1.57, 1.18, 1.06, 1.12, and 1.52, respectively) and showed Areas Under the ROC Curve (AUC) 0.805, the model displayed sensitivity of 72.0%, and specificity 75.1% when the most appropriate cutoff value was used, 0.579 as given by the Youden Index. This model has shown to be useful to identify patients with a high-risk of cerebral white matter lesions, who can then be diagnosed with a head MRI examination in order to prevent dementia, cerebral infarction, and stroke.

Introduction

In Japan, it is generally recognized that the increase in national health expenditure that accompanies aging is a serious social problem [1]. Although the number of deaths by stroke has decreased drastically due to preventive treatment, stroke still ranks at the top of the health care expenditures in Japan [2]. It is estimated that 2 million people are currently bedridden, and this number will increase to 3 million by 2025 [3]. Therefore, the early detection and prevention of cerebrovascular diseases is important for the reduction of health care expenditures [4].

In recent years, cerebral white matter lesions, which are typically found during MRI tests, are drawing attention as a factor to predict dementia and stroke [5]. It is considered that cerebral white matter lesions are risk factors for dementia and stroke [5], which show progressive symptoms. Cerebral white matter lesions are mainly ischemic symptoms with microangiopathy [6], which show white spots scattered in the cerebral white matter in the MRI image, and gradually progress to a large lump [7].

The surface of the cerebrum is called gray matter, which is composed of gathering nerve cells. There is cerebral white matter at the central side of the gray matter, composed of bundles of nerve fibers that transmit signals from nerve cells. Cerebral white matter is enclosed by cerebral arterioles, which are nutritive vessels. If arteriosclerosis occurs in the cerebral arteriole, the vascular wall loses its elasticity and narrows the vascular lumen, which shows up as white spots on an MRI due to leaked liquids. Because arteriosclerosis is irreversible, if it occurs in the cerebral arterioles, the blood flow to the brain cells chronically decreases, leading to cell necrosis [8].

Most of the patients with cerebral white matter lesions usually engage in work and experience daily life without any problems. For patients without subjective symptoms, the comprehensive medical checkup that includes a head MRI scan, which is called "brain dock," is necessary and is the first choice to detect and diagnose white matter lesions in Japan. Generally, the brain dock is an added option to a comprehensive medical checkup that focuses on brain diseases, and includes diagnostics such as the head MRI. The first brain dock service in Japan was launched in the Shinsapporo Neurosurgical Hospital in March of 1988 [9], after which the brain dock services were expanded to over 600 facilities. As of November 30, 2018, 289 facilities were certified by the facility certification system for brain dock, which is administered by the Japan Brain Dock Society and has been in effect since 2009 [10]. At present, over 670 facilities provide the brain dock course, including those facilities certified by the Japan Brain Dock Society. The brain dock course is a medical evaluation at a patient’s own expense and is not covered by medical insurance. The pricing of brain docks is set independently by each medical facility, and it is said that it generally costs from 50 to 100 thousand yen (approx. 500 to 1000 USD).

According to “Guidelines of brain dock,” by the Japan Brain Dock Society, though head MRI imaging is performed in the brain dock course to detect cerebral white matter lesions and to grade the lesions, because there is no influence and no subjective symptom in the daily life, it is necessary to not to engender unnecessary anxiety. However, for patients with cerebral white matter lesions, it is necessary to pay attention to risk factors and care for adult diseases in order to prevent dementia, cerebral infarction, and stroke from occurring in the future [11].

The goal of this study was to construct a discriminant model to predict cerebral white matter lesions using clinical examination data from a routine medical checkup, toward the development of a screening test based on scientific evidence aiming at the prevention of dementia, cerebral infarction, and stroke.

Materials and methods

Subjects

This study is an examination of 1,904 subjects, including 988 males and 916 females, who underwent head MRI and blood tests during the brain dock course of a comprehensive medical checkup sometime between April 1, 2016 and October 31, 2017 at Shin Takeo Hospital. During the investigation period, there were no hospital policy changes for the brain dock course of a comprehensive medical checkup. The average age of the examinees was 56.4 ± 11.5

[range: 20–85] (mean ± S.D.), and 1,044 subjects were diagnosed from their MRI results as having cerebral white matter lesions. In the head MRI, T1 weighted images (T1WI), T2 weighted images (T2WI), and Fluid Attenuated Inversion Recovery (FLAIR) images were obtained by using MRI scanners, MAGNETOM Symphony (Siemens Healthineers Japan, Tokyo, Japan) and MAGNETOM ESSENZA (Siemens Healthineers Japan, Tokyo, Japan).

Ethical consideration

This study was conducted based on the approval of the ethical review committee of Shin Takeo Hospital. For the protection of patient privacy, the patient data was collected with unlinkable anonymization by a third party, and was saved in a password-protected storage medium for research use only. The clinical data used in this study will be available upon request from readers to the corresponding author based on the data usage agreement and considering the patients' right of privacy.

Clinical data

General inspection and blood and biochemical tests

In the general inspection, six characteristics were recorded for the study: age, gender, systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), and presence of visceral steatosis (for the determination of metabolic syndrome) [12]. The blood and biochemical tests were conducted with laboratory test systems, C8000 (Canon Medical Systems Corporation, Tochigi, Japan) and Acute (Canon Medical Systems Corporation, Tochigi, Japan), respectively. HbA1c was determined with a glycohemoglobin analyzer HA8181 (Arkray inc., Kyoto, Japan). In the blood and biochemical tests, which were conducted as a part of the comprehensive medical checkup, LDL cholesterol (LDL), HDL cholesterol (HDL), LH ratio (quotient of LDL and HDL), Triglyceride (TG), hemoglobin A1c (HbA1c), and blood glucose level (BS) were recorded.

Ultrasonic testing

The ultrasonic testing was conducted with ultrasonic diagnostic equipments, LOGIQ S7 Expert (GE Healthcare Japan, Tokyo) and Aplio 400 (Canon Medical systems, Tochigi, Japan). Two characteristics were recorded: carotid plaque score (PS) [13] and plaque number (n-plaque). The PS was calculated as follows. The carotid artery was divided into four 15 mm long sections: the central side of the common carotid artery (CCA), the peripheral side of the CCA, the bifurcation of the CCA and the central side of the internal carotid artery. Then, the sum of the maximum values of intima-media thickness exceeding 1.1 mm was calculated.

Questionnaire in the specific health examination

In the questionnaire in the specific health examination (shown in S1 Doc), six questions were answered by examinees when receiving a comprehensive medical checkup, regarding their experience with medications to reduce blood pressure, medications to reduce blood sugar or insulin injection, medications to decrease the level of cholesterol or of neutral fat, as well as drinking habits (everyday, sometimes, or rarely drink (cannot drink)), drinking volume (less than 180 ml, 180–360 ml, 360–540 ml or more than 540 ml), and smoking habits.

Additional data to replicate all of the figures, graphs, tables, statistics, and other values in this study is available at doi:10.5061/dryad.007467q as Data files: WM_data.

Assessment of each clinical examination data and questionnaire

R version 3.4.4 and its suitable packages were used to perform all statistical analysis and statistical modeling in this study [14,15]. The relationships between each examination item or each answer of the specific medical examination questionnaire and the probability of the presence of cerebral white matter lesions were investigated. Student's *t* tests [16] were performed for continuous variables and Fisher's exact tests [17] were performed for categorical variables.

Comparison among various models based on different algorithms

Four kinds of models including linear models, nonlinear models, and a stochastic model were created, and an accuracy comparison among the models was conducted. The normalization of the variables was performed to adjust the factor levels. In the linear modeling, a logistic regression modeling (LogReg) [18] were employed, and in the nonlinear modeling, support vector machine (SVM) [19] or random forest (RF) [20] models were constructed. Furthermore, a Naive Bayes classifior (NB) [21] was constructed as a stochastic model. The probability of the presence or absence of white matter lesions was used as the dependent variable. In each model, the Youden Index [22, 23] was used as the most appropriate cutoff value to compare the performance among models.

The discrete variables (or categorical variables) in this study were treated as follows:

Regarding binary variables of models, 0 or 1 was assigned for absence or presence, respectively, e.g., sex: $X = 0$ for male and $X = 1$ for female; medication to reduce blood pressure: $X = 0$ for "No" and $X = 1$ for "Yes"; medication to reduce blood sugar or insulin injection: $X = 0$ for "No" and $X = 1$ for "Yes"; medication to reduce blood pressure: $X = 0$ for "No" and $X = 1$ for "Yes".

Regarding ternary variables of models, three kinds of vectors (0,0), (0,1), and (1,1) were assigned for the corresponding combinations with two variables (X_1,X_2), respectively, e.g. the determination of metabolic syndrome: $X_1 = 0$ and $X_2 = 0$ for non-metabolic syndrome, $X_1 = 1$ and $X_2 = 0$ for the reserve of metabolic syndrome, and $X_1 = 1$ and $X_2 = 1$ for metabolic syndrome; the drinking habits: $X_1 = 0$ and $X_2 = 0$ for "rarely drink (cannot drink)", $X_1 = 1$ and $X_2 = 0$ for "sometimes" and $X_1 = 1$ and $X_2 = 1$ for "everyday".

For multivariable modeling, the appropriate R packages were used as follows: LogReg was performed with glm() with a family ="binominal" argument in the "stats" package; SVM was performed with ksvm() in the "kernlab" package (kernlab v0.9–27) [24]; RF was performed with randomForest() in the "randomForest" package (randomForest, Version 4.6–14) [20]; NB was performed with NaiveBayes() in the "klaR" package (klaR, Version 0.6–14) [25].

Variable selection

In the variable selection, graphical modeling [26] was performed so as not to select duplicate variables from the same cluster showing strong correlation by considering multicollinearity [27] with the R packages "corpcor (corpcor, Version 1.6.9)" [28] and "qgraph (qgraph, Version 1.5)." [29] The strength of correlation was calculated with a Pearson's product moment correlation coefficient. For the model performance comparison of the models, 10-fold cross-validation [30] was performed for each model.

The procedure of the 10-fold cross-validation was as follows. First, the dataset was divided into ten fractions, and the first 1/10 fraction was used as a holdout set. Model training was performed with the remaining 9/10 fractions, and then the 1/10 holdout set was used for evaluation of the trained model, from which the values of certain evaluation indices were recorded. After that, the first 1/10 holdout was returned to the original data and the next 1/10 holdout set was taken out. This training and testing of the model were performed in an iterative manner until each of the fractions had been used for as a holdout set.

Evaluation of discrimination performance and model selection

Accuracy, error rate, sensitivity [31], specificity [31], positive predictive value (PPV) [32], negative predictive value (NPV) [32], and Area Under the ROC Curve (AUC) [33] were used for as evaluation indices. The accuracy was calculated by (number of true positive + number of true negative)/total population. The error rate was calculated by (number of false positive + number of false negative)/total population. True positive rate (TPR, also called sensitivity) was calculated by number of true positive/number of positive. True negative rate (TNR, also called specificity) was calculated by number of true negative/number of negative. Meanwhile, PPV was calculated by number of true positive/number of predicted positive and NPV was calculated by number of true negative/number of predicted negative. The ROC curves were plotted using the R packages "plotROC (plotROC, Version 2.2.1)" [34] and "ggplot2 (ggplot2 3.0.0)" [35].

The average values of the 10-fold cross-validations of each index were compared. Finally, based on the accuracy comparison among the four models, the model showing good accuracy in addition to good clinical interpretation was selected and established as the discriminant model for identifying the patients with cerebral white matter lesions.

Results

Assessment of each clinical examination data and questionnaire

Using the head MRI, which is the gold standard for detection of white matter lesions, 1,044 out of 1,904 subjects (54.8%) were diagnosed with white matter lesions. Fig 1 shows some typical head MRI examples of the presence or absence of white matter lesions. In Fig 1, the association between the presence or absence of white matter lesions and other results from the clinical investigations, i.e., the general inspection, blood and biochemical tests, ultrasonic testing, and specific health examination questionnaire, were investigated.

Table 1 shows the assessment of each clinical examination data and questionnaire. In the general inspection, significant differences in age, gender, SBP, DBP, and the presence of metabolic syndrome were found between groups. In the blood tests, significant differences were seen in HDL, HbA1c, and BS. In the ultrasonic tests, significant differences were found in PS and n-plaque. In the answers to the specific health examination questionnaire, significant differences were seen in the patients' experience with medications to reduce blood pressure and those to reduce blood sugar or insulin injection, as well as drinking habits, drinking volume, and smoking habits.

Variable selection for multivariate analysis

Graphical modeling was performed to avoid the multicollinearity problem that can result from the interactions among explanatory variables when creating a mathematical model. Fig 2 shows the results of the graphical modeling for variable selection. Four clusters of variables with strong correlation were created. The first cluster contained PS and n-plaque; therefore, PS was selected from this cluster because it could capture the effects of both plaque number and plaque thickness [36]. The second cluster contained SBP and DBP; SBP, which is generally considered to be a risk factor of atherosclerosis [37], was selected from this cluster. The third cluster contained HbA1c and BS; HbA1c, which is thought to reflect the blood sugar level for the most recent several months, was selected from this cluster. The fourth cluster contained LDL, HDL, and LH ratio; LDL, also called "bad" cholesterol, which is regarded to be a risk

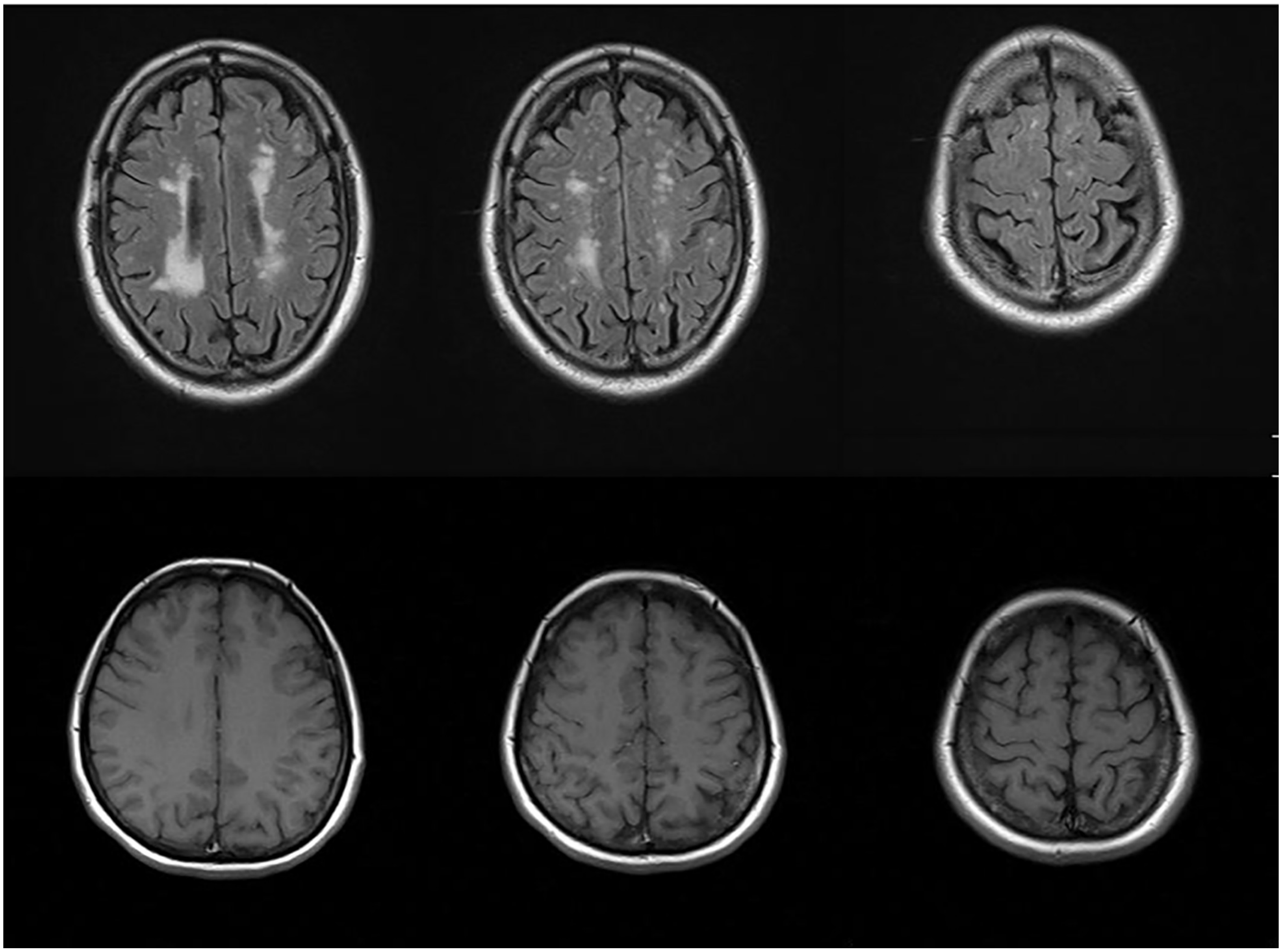

Fig 1. Typical examples of the presence or absence of cerebral white matter lesions. The upper three images are those of a subject with cerebral white matter lesions, while the lower three images are those of a subject without cerebral white matter lesions. T1 weighted images (T1WI), T2 weighted images (T2WI), and Fluid Attenuated Inversion Recovery (FLAIR) images were obtained using MRI equipment (MAGNETOM Symphony and MAGNETOM ESSENZA; Siemens Healthineers Global), but only the FLAIR images are shown here.

https://doi.org/10.1371/journal.pone.0215142.g001

factor of arteriosclerosis [38], was selected from this cluster. From the specific health examination questionnaire, experience with medications to reduce blood pressure, experience with medications to reduce blood sugar or insulin injection, experience with medications to decrease the level of cholesterol or of neutral fat, and drinking habits [39] were selected.

Performance comparison among mathematical models

Table 2 shows the results of the performance comparison among models by 10-fold cross-validation. The Youden Index [22, 23] was used as the most appropriate cutoff value to evaluate the performance of each model. NB showed the highest average accuracy of 72.0% in the hold-out datasets; RF showed the highest average sensitivity of 83.1%; in specificity and PPV, logistic regression analysis (LogReg) showed the highest averages, 79.4% and 79.1%, respectively; in NPV, Naive Bayes classifier (NB) showed the highest average of 69.8%; and in AUC, LogReg showed the highest average of 0.799. Fig 3 shows the ROC curves with each model using the same test data set, and they are almost overlapping with all the models.

Table 1. Assessment of each clinical examination data and questionnaire.

		non white matter		white matter		
n		**860**		**1044**		
factor						***p*-value**
PS, mean (sd)		0.59	(1.3)	1.49	(2.3)	<0.001
age, mean (sd)		49.96	(10.8)	61.73	(9.2)	<0.001
LDL, mean (sd)		119.64	(31.6)	121.88	(29.4)	0.109
HDL, mean (sd)		59.47	(14.7)	62.49	(15.8)	<0.001
LH, mean (sd)		2.14	(0.8)	2.08	(0.7)	0.068
TG, mean (sd)		115.59	(124.4)	108.71	(73.6)	0.134
HbA1c, mean (sd)		5.66	(0.6)	5.86	(0.7)	<0.001
BS, mean (sd)		102.41	(16.8)	105.45	(19.2)	<0.001
SBP, mean (sd)		119.96	(16.6)	127.1	(19.3)	<0.001
DBP, mean (sd)		72.43	(11.7)	75.05	(12.5)	<0.001
the number of plaque, mean (sd)		0.34	(0.7)	0.84	(1.2)	<0.001
BMI, mean (sd)		23.13	(3.4)	23.17	(3.4)	0.774
gender, n (%)	male	497	(57.8)	491	(47.0)	<0.001
	female	363	(47.2)	553	(53.0)	
metabolic syndrome, n (%)	no	687	(79.9)	742	(71.1)	<0.001
	reserve	79	(9.2)	117	(11.2)	
	yes	94	(10.9)	185	(17.7)	
smoking habit, n (%)	yes	197	(22.9)	139	(13.3)	<0.001
	no	663	(77.1)	905	(86.7)	
medication to reduce blood pressure, n (%)	yes	109	(12.7)	353	(33.8)	<0.001
	no	751	(87.3)	691	(66.2)	
medication to reduce blood sugar or insulin injection, n (%)	yes	32	(3.7)	107	(10.2)	<0.001
	no	828	(96.3)	937	(89.8)	
medication to reduce a level of cholesterol, n (%)	yes	82	(9.5)	233	(22.3)	<0.001
	no	778	(90.5)	811	(77.7)	
amount of drinking per day, n (%)	less tha 180mL	504	(58.6)	720	(69.0)	<0.001
(in terms of Sake)	180-360mL	237	(27.6)	230	(22.0)	
	360mL-540mL	89	(10.3)	69	(6.6)	
	more than 540mL	30	(3.5)	25	(2.4)	
drink habit, n (%)	rarely drink	331	(38.5)	465	(44.5)	0.028
	sometimes	269	(31.3)	296	(28.4)	
	everyday	260	(30.2)	283	(27.1)	

p-values were calculated by the Student's *t* test for continuous variables and by the Fisher's exact test for categorical variables. "LH" shows the LH ratio. The other abbreviations are shown in "the Materials and Methods" section.

https://doi.org/10.1371/journal.pone.0215142.t001

Creation of a clinical predictive model

Finally, the LogReg model was selected as a clinical predictive model because it showed the highest score in three of the six indices (specificity, PPV and AUC) by 10-fold cross-validation, and because of its clinical interpretability (e.g., odds ratio), usability, and model simplicity.

The final discriminant model created in this study is as follows:

$$\log\frac{\Pr(Y_i = 1|X_i = x_i)}{\Pr(Y_i = 0|X_i = x_i)} = -0.28+1.1x_{i1}+0.45x_{i2}+0.16x_{i3}+0.06x_{i4}+0.12x_{i5}-0.04x_{i6}$$
$$+0.43x_{i7}+0.15x_{i8}+0.42x_{i9}+0.37x_{i10}+0.15x_{i11}+0.24x_{i12}+0.04x_{i13}$$

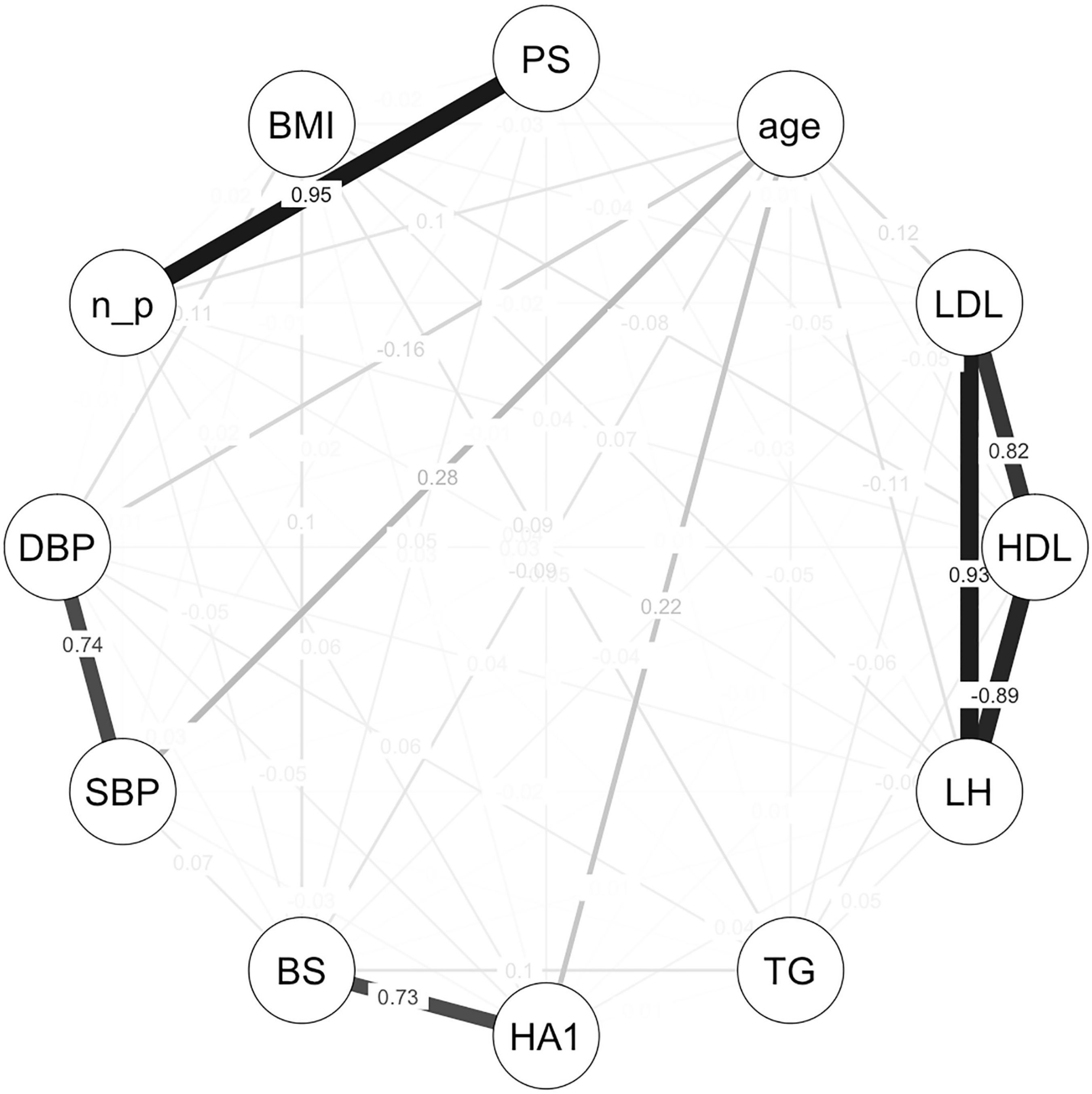

Fig 2. Variable selection by graphical modeling. The graphical modeling was performed using the R packages "corpcor" [26] and "qgraph" [27].

https://doi.org/10.1371/journal.pone.0215142.g002

where the presence or absence of white matter lesions was used as the dependent variable Y ($Y = 1$ when white matter lesions were present, and $Y = 0$ when white matter lesions were absent). The explanatory variables were defined as follows: X_1 showed age; X_2 showed sex ($X_2 = 0$ for male and $X_2 = 1$ for female); X_3 showed PS; X_4 showed LDL; X_5 showed SBP; X_6 showed HbA1c; X_7 and X_8 showed the determination of metabolic syndrome ($X_7 = 0$ and $X_8 = 0$ for non-metabolic syndrome; $X_7 = 1$ and $X_8 = 0$ for the reserve of metabolic syndrome; $X_7 = 1$ and

Table 2. Performance comparison of each model by 10-fold cross-validation.

model	AUC	Cut off point	accuracy	error	TPR	TFR	PPV	NPV
LogReg	0.799	0.566	71.1%	28.9%	64.4%	79.4%	79.1%	64.8%
NB	0.776	0.382	72.0%	28.0%	76.5%	66.6%	73.8%	69.8%
SVM	0.787	0.679	70.7%	29.3%	64.5%	78.7%	48.8%	28.0%
RF	0.790	0.428	71.7%	28.3%	83.1%	58.0%	39.8%	30.6%

The four models, logistic regression (LogLeg), Naive Bayes classifier (NB), support vector machine (SVM), and random forest (RF), were compared with 6 indices: accuracy, error rate (error), true positive rate (TPR, also called sensitivity), true negative rate (TNR, also called specificity), positive predictive value (PPV), negative predictive value (NPV), and Area Under the ROC Curve (AUC).

https://doi.org/10.1371/journal.pone.0215142.t002

$X_8 = 1$ for metabolic syndrome); X_9 showed the experience with medication to reduce blood pressure ($X_9 = 0$ for "No"; $X_9 = 1$ for "Yes"); X_{10} showed the experience with medication to reduce blood sugar or insulin injection ($X_{10} = 0$ for "No"; $X_{10} = 1$ for "Yes"); X_{11} showed the experience with medication to reduce the level of cholesterol or of neutral fat ($X_{11} = 0$ for "No"; $X_{11} = 1$ for "Yes"); X_{12} and X_{13} showed the drinking habits ($X_{12} = 0, X_{13} = 0$ for "rarely drink (cannot drink)"; $X_{12} = 1, X_{13} = 0$ for "sometimes"; $X_{12} = 1, X_{13} = 1$ for "everyday"). For the i-th patient's data x_i, $\Pr(Y_i = 1|X_i = x_i)$ showed the probability that the i-th patient had the

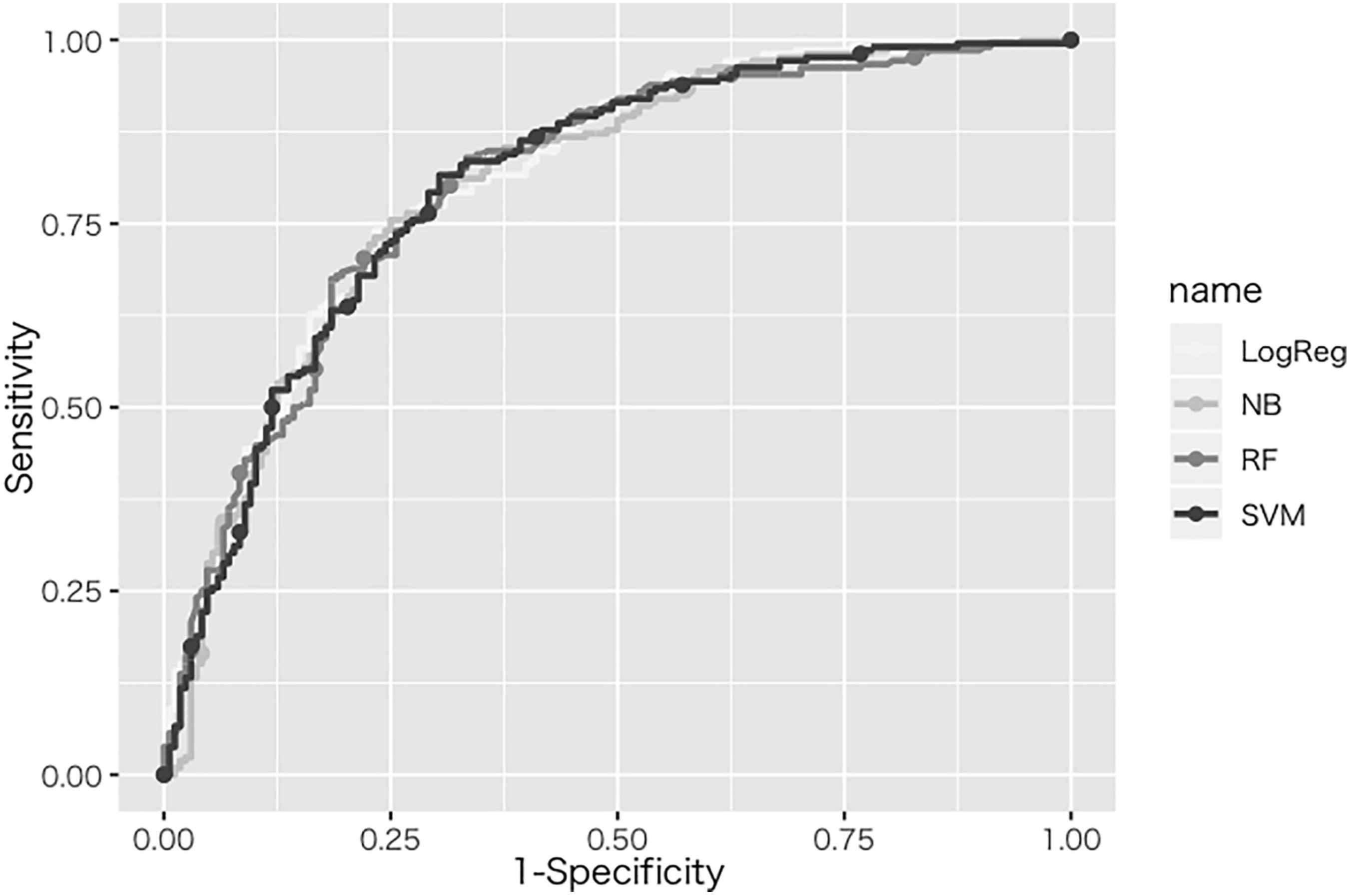

Fig 3. ROC curves using all the investigated models. LogReg: logistic regression analysis. NB: Naive Bayes classifier. RF: random forest. SVM: support vector machine. The ROC curves were plotted using the R packages "plotROC" [32] and "ggplot2" [33].

https://doi.org/10.1371/journal.pone.0215142.g003

Table 3. Odds ratio of each variable estimated by the logistic regression model.

		odds ratio	95%CI		
age		2.99	2.61	3.45	*
gender	male	reference			
	female	1.57	1.22	2.03	*
PS		1.18	1.03	1.35	*
LDL		1.06	0.95	1.19	
SBP		1.12	1.00	1.26	*
HbA1c		0.97	0.84	1.11	
metabolic syndrome	no	reference			
	reserve	1.53	1.06	2.23	*
	yes	1.16	0.81	1.67	
medication to reduce blood pressure	no	reference			
	yes	1.52	1.13	2.06	*
medication to reduce blood sugar or insulin injection	no	reference			
	yes	1.45	0.83	2.61	
medication to reduce a level of cholesterol	no	reference			
	yes	1.16	0.83	1.64	
drink habit	rarely drink	reference			
	sometimes	1.28	0.98	1.67	
	everyday	1.04	0.78	1.40	

(*) indicates that the confidence interval does not include 1.

https://doi.org/10.1371/journal.pone.0215142.t003

white matter lesions, and $\Pr(Y_i = 0|X_i = x_i)$ showed the probability that the i-th patient did not have white matter lesions.

Table 3 shows the odds ratios estimated by the logistic regression model. Each variable was normalized and the estimate was adjusted to the same order. Table 4 shows the results of the

Table 4. Summary of the logistic regression model using selected variables.

parameter	coefficients	std	*p*-value	
(Intercept)	-0.28	0.14	0.04	*
age	1.10	0.07	<0.001	*
Gender:female	0.45	0.13	<0.001	*
PS	0.16	0.07	0.02	*
LDL	0.06	0.06	0.30	
SBP	0.12	0.06	0.05	*
HbA1c	-0.04	0.07	0.32	
metabolic syndrome:reserve	0.43	0.19	0.03	*
metabolic syndrome:yes	0.15	0.18	0.41	
medication to reduce blood pressure:yes	0.42	0.15	0.01	*
medication to reduce blood sugar or insulin injection : yes	0.37	0.29	0.20	
medication to reduce a level of cholesterol : yes	0.15	0.17	0.38	
drink habit : sometimes	0.24	0.14	0.08	
drink habiit : everyday	0.04	0.15	0.79	

"std" represents the standard error in the output of glm().

(*) indicates a significant variable ($p<0.05$).

https://doi.org/10.1371/journal.pone.0215142.t004

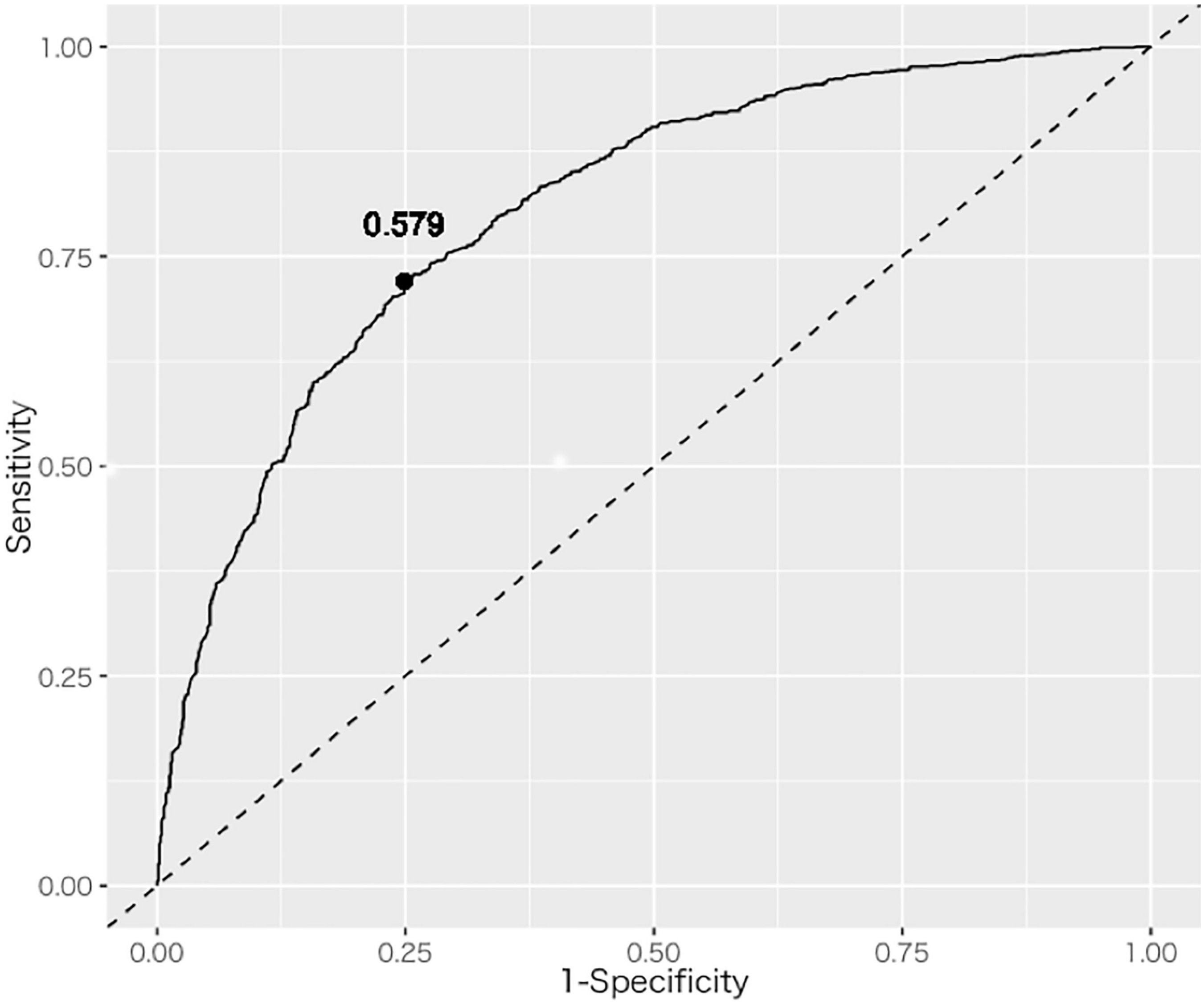

Fig 4. A ROC curves using the final logistic discriminant model. A point on the ROC curve shows the most appropriate cutoff value giving by the Youden Index.

https://doi.org/10.1371/journal.pone.0215142.g004

logistic regression analysis performed on all of the data, instead of the train/test split dataset or cross-validation data. The confidence intervals of the odds ratio of age, gender, PS, SBP, and experience with medications to reduce blood pressure showed significance, as they were all greater than one.

The ROC curve with logistic discrimination is shown in Fig 4. In the ROC curve, for the most appropriate cutoff value, 0.579, given by the Youden Index [22, 23], the specificity was 0.249 and sensitivity was 0.720, which is the farthest point on the ROC curve from the diagonal line of AUC = 0.5.

Discussion

In the assessment of questionnaire in Table 1, though smoking habits showed significance ($p<0.001$) in the between-groups comparison, the ratio of non-smoking patients with white matter lesions (86.7%) was higher than that of smokers with white matter lesions with smoking habits (13.3%). Generally, it is considered that smoking is a risk factor for cerebrovascular disorders [37], but this study showed the opposite result. Although it is said that smoking is a risk factor of cerebrovascular disorder [36], smoking itself is not a direct cause of ischemic change due to microangiopathy, that is, white matter lesions. It is possible that other indirect and complex factors including changes in blood constituents, such as a decrease of "good" HDL cholesterol due to smoking habits, could lead to their expression. Therefore, it is probably necessary to investigate the detailed temporal changes in smoking amounts in order to complete an effective evaluation of smoking habits. In the specific health examination questionnaire in this study, the question regarding smoking habits only asked whether the participant was a heavy smoker, defined as having smoked a total of over 100 cigarettes or have smoked over a period of 6 months, not considering the past history or cigarettes smoked or/and the number of cigarettes smoked in a day. It was considered that the index could become a more meaningful variable regarding smoking, if it included detailed information such as the smoking index represented by the product of the number of cigarettes smoked per day and the smoking history (years). However, in Japan, specific health examination questionnaires are utilized in medical checkups to check the patient's past medical history and lifestyle and to aid in patient consultation by using the similar format of questionnaires at most of the hospitals. Therefore, it is considered an important part of the patient consultation in medical checkups. Therefore, in this modeling, the specific health examination questionnaires were regarded as an important factor from the viewpoint of its practical usage.

From the assessment results of the 10-fold cross-validation, it was determined that the appropriate variable selection had been performed using the variables obtained from the limited clinical examination data so that there was not much difference in the accuracy among created models. The models used in this study include linear (LogReg) and nonlinear (SVM and RF) models as well as a stochastic model (NB). SVM [40] and RF [41] are often reported as showing better discrimination performance than linear models. Furthermore, it is considered that Bayesian modeling, which is stochastic, shows less overfitting and better generalization ability than conventional machine learning [42]. However, some contradictory reports [43–45] have appeared recently; therefore, so it was necessary to clarify the effect on the discrimination performance due to the difference in algorithms. In this study, it was demonstrated that the model algorithm did not significantly influence the discrimination performance if the appropriate variable selection had been conducted when constructing the predictive model.

Considering both the discrimination performance and the interpretability of the model, the linear model is better than the other models for the prediction of the presence of white matter lesions from clinical examination data. Using a logistic regression model, the presence of white matter lesions can be represented with a probability, and the odds ratio of a given risk factor can also be calculated. Therefore, it is considered to be suitable for clinical use among models in this study.

In this study, logistic discrimination was selected for the construction of the clinical model. There were no significant differences in accuracy, error rate, and AUC among all algorithms, but significant differences were observed in sensitivity and specificity. Generally, in screening testing before the major invasive testing, sensitivity is important in efforts to reduce the occurrence of false negatives. However, in this study, cerebral white matter lesions were predicted, a diagnosis that does not show subjective symptoms, does not lead to lethal symptoms, and does not require direct care. When diseases that cause cerebrovascular disorders are discovered, such as hypertension, hyperlipidemia, and atrial fibrillation, treatment is conducted to combat them. Considering the practical usage of the model, the model is to be used for the people getting general health examinations, by predicting the presence or absence of cerebral white matter lesions, thus allowing for the recommendation that the predicted high risk subjects receive the brain dock. Therefore, in this modeling, specificity is more important than sensitivity. By using the logistic discriminant model showing high specificity, the odds ratio of risk factors can be calculated, making the clinical evaluation seems easy.

Regarding the odds ratio of age (2.99 in Table 3), since age was standardized, the odds ratio for 1 year change in age was calculated to be 1.10, which is consistent with reports that slight white matter lesions in the elderly are an age-related phenomenon [46]. Advancing white matter lesions indicate a high risk of dementia and stroke [6], and it is important to prevent their progression.

The odds ratio of PS was significant at 1.12, which is consistent with the fact that PS is commonly used as an indicator of arteriosclerosis [47]. In the facilities investigated in this study, a PS test is performed only in the brain dock course. Given this demonstration of an association of PS with cerebral white matter lesions, if PS were to be added as an optional testing item in the comprehensive medical checkup, it would be useful as a variable for cerebral white matter lesion prediction.

Variables such as the medication to reduce blood pressure, the medication to reduce blood sugar or insulin injection, and the medication to decrease the level of cholesterol or of neutral fat are data derived from the specific health examination questionnaire and are used as a substitute for variables from the past medical history. Naturally, not all hospitals conduct a comprehensive medical checkup and there are many comprehensive medical checkups performed by private companies in Japan; therefore, it seems that there are not so many comprehensive medical checkups performed in the primary care hospitals. Therefore, because the medical history data must be obtained from the questionnaire, it is difficult to identify the exact disease name and diagnostic results without a common questionnaire format. If information can be collected in a common format, such as using hierarchical categories regarding the medical history with the questionnaire, a past history can also be used as variables, which would very likely improve the discrimination performance.

Approximately 54.8% of the patients whose data were used in this study have been diagnosed as having white matter lesions and so the prevalence is high. The brain dock course in this study is more expensive than the general comprehensive health examination courses. Consequently, it is a possibility that many examinees that had risks or medical history related to arteriosclerosis were included in the study. Because of the possibility that subjects in this study may be a high-risk group, it is not possible to conclude that the prediction of all cerebral white matter lesions can be applied for all subjects. However, it was shown to be possible to discriminate a risk group in subjects that had not received an MRI by using a prediction based on the clinical examination data.

Conclusions

To predict cerebral white matter lesions based on clinical examination data, a logistic regression model was selected from some candidate models, created by various algorithms, based on a comparison of accuracy and interpretability. The explanatory variables of the model were age, gender, PS, LDL, SBP, and the administration of antihypertensive medication. Variable selection was important to the establishment of a high accuracy model, but the model algorithm did not significantly influence the discrimination performance if an appropriate variable selection was conducted while constructing the prediction model. This model will allow clinicians to discriminate a risk group in subjects who have not received a head MRI test.

Supporting information

S1 Doc. Questionnaire in the specific health examination, which was translated from the form in Japanese, in Shin Takeo Hospital.
(DOCX)

Acknowledgments

We are grateful to Drs. T. Kakuma, K. Furukawa, T. Yanagawa, S. Hattori, K. Murotani, T. Ohyama, T. Miyamoto, Mr Y. Koga, Mrs. R. Todoroki and other members in Kurume University Biostatistics Center for support and discussion of data analysis, and for comments on an earlier draft of this study. We thank Mrs. T. Yamaguchi and other staff in Shin Takeo Hospital for preparation of clinical examination data.

Author Contributions

Conceptualization: Yuya Shinkawa, Kazuo Ishii.

Formal analysis: Yuya Shinkawa, Takashi Yoshida, Yohei Onaka, Makoto Ichinose, Kazuo Ishii.

Investigation: Yuya Shinkawa, Kazuo Ishii.

Methodology: Yuya Shinkawa, Kazuo Ishii.

Project administration: Yuya Shinkawa, Kazuo Ishii.

Resources: Yuya Shinkawa, Takashi Yoshida, Yohei Onaka, Makoto Ichinose.

Software: Kazuo Ishii.

Supervision: Kazuo Ishii.

Validation: Yuya Shinkawa, Kazuo Ishii.

Visualization: Yuya Shinkawa.

Writing – original draft: Yuya Shinkawa, Kazuo Ishii.

Writing – review & editing: Yuya Shinkawa, Kazuo Ishii.

References

1. McCurry J. Japan will be model for future super-ageing societies. Lancet. 2015; 386: 1523. https://doi.org/10.1016/S0140-6736(15)00525-5 PMID: 26530607
2. Matsumoto K, Hanaoka S, Wu Y, Hasegawa T. Comprehensive Cost of Illness of Three Major Diseases in Japan. J Stroke Cerebrovasc Dis. 2017; 26: 1934–1940. https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.06.022 PMID: 28712721
3. Himiyama Y. Area Studies—Japan (Regional Sustainable Development Review). EOLSS Publishers, Oxford. 2009; pp84.
4. Ogura S, Tachibanaki T, Wise DA, Issues in the United States and Japan. The University of Chicago Press, Chicago and London. 2007; pp196.
5. Debette S, Beiser A, DeCarli C, Au R, Himali JJ, Kelly-Hayes M et al. Association of MRI markers of vascular brain injury with incident stroke, mild cognitive impairment, dementia, and mortality: the Framingham Offspring Study. Stroke. 2010; 41: 600–606. https://doi.org/10.1161/STROKEAHA.109.570044 PMID: 20167919
6. Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, et al. STandards for ReportIng Vascular changes on nEuroimaging (STRIVE v1). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 2013; 12: 822–838. https://doi.org/10.1016/S1474-4422(13)70124-8 PMID: 23867200

7. Sundaresan V, Griffanti L, Kindalova P, Alfaro-Almagro F, Zamboni G, Rothwell PM, et al. Modelling the distribution of white matter hyperintensities due to ageing on MRI images using Bayesian inference. Neuroimage. 2019; 185: 434–445. https://doi.org/10.1016/j.neuroimage.2018.10.042 PMID: 30359730

8. Chowdhury MH, Nagai A, Bokura H, Nakamura E, Kobayashi S, Yamaguchi S. Age-related changes in white matter lesions, hippocampal atrophy, and cerebral microbleeds in healthy subjects without major cerebrovascular risk factors. J Stroke Cerebrovasc Dis. 2011; 20: 302–309. https://doi.org/10.1016/j.jstrokecerebrovasdis.2009.12.010 PMID: 20634092

9. Nakagawa T, Hashi K, Characteristics and Treatment of Asymptomatic Unruptured Cerebral Aneurysms Detected by "Brain Dock," Jpn J Neurosurg. 1995; 4: 341–350.

10. The Japan Brain Dock Society. The list of the facility certification system for brain dock. 2018 Nov 30 [cited 3 Jan 2019]; http://jbds.jp/nintei_list.html.

11. "The committee of creating new guidelines of brain dock," The Japan Brain Dock Society. Guidelines of brain dock 2014. Sapporo. Kyobunsha; 2014.

12. Kaur J. A comprehensive review on metabolic syndrome. Cardiol Res Pract. 2014; 2014: 943162. https://doi.org/10.1155/2014/943162 PMID: 24711954

13. Zhang H, Liu M, Ren T, Wang X, Liu D, Xu M, et al. Associations between carotid artery plaque score, carotid hemodynamics and coronary heart disease. Int J Environ Res Public Health. 2015; 12: 14275–14284. https://doi.org/10.3390/ijerph121114275 PMID: 26569275

14. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; http://www.R-project.org.

15. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J. Comp. Graph. Stat. 1996; 5: 299–314.

16. Student. The Probable Error of a Mean. Biometrika. 1908; 6: 1–25.

17. Fisher RA. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. J Royal Stat Soc. 1922; 85: 87–94.

18. Cox DR. The regression analysis of binary sequences (with discussion). J Roy Stat Soc B. 1958; 20: 215–242.

19. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995; 20: 273–297.

20. Breiman L. Random Forests. Mach Learn. 2001; 45: 5–32.

21. Pedro D, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn. 1997; 29: 103–137.

22. Unal I. Defining an optimal cut-point value in ROC analysis: an alternative approach. Comput Math Methods Med. 2017; 2017: 3762651. https://doi.org/10.1155/2017/3762651 PMID: 28642804

23. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3: 32–35. PMID: 15405679

24. Chang CC, Lin CJ. LIBSVM: A library for support vector machines, ACM Trans Int Syst Tech. 2011; 2: 1–27S

25. Weihs C, Ligges U, Luebke K, Raabe N. klaR analyzing German business cycles. In Baier D, Decker R, Schmidt-Thieme L, editors. Data analysis and decision support. Springer-Verlag, Berlin; 2005: pp 335–343

26. Honorio J, Samaras D, Rish I, Cecchi G. Variable selection for gaussian graphical models. Proc Mach Learn Res. 2012; 22: 538–546.

27. Farrar DE, Glauber RR. Multicollinearity in Regression Analysis: The problem revisited. Rev Econ Stat. 1967; 49: 92–107.

28. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol. 2005; 4:Article 32.

29. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: network visualizations of relationships in psychometric data. J Stat Soft, 2012; 48: 1–18.

30. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann. San Mateo. 1995;2: 1137–1143.

31. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994; 308: 1552. PMID: 8019315

32. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994; 309: 102. PMID: 8038641

33. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pat Rec. 1997; 30: 1145–1159.

34. Sachs MC. plotROC: A Tool for Plotting ROC Curves. J Stat Sof. 2017; 79: Code Snippet 2.

35. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York: 2009.

36. Naqvi TZ, Lee MS. Carotid intima-media thickness and plaque in cardiovascular risk assessment. JACC Card Imag. 2014; 7: 1025–1038.

37. Zhang H, Xu MG, Xie LJ, Huang M, Shen J, Xiao TT. Meta-analysis of risk factors associated with atherosclerosis in patients with Kawasaki disease. World J Pediatr. 2016; 12: 308–313. https://doi.org/10.1007/s12519-016-0023-0 PMID: 27351565

38. Hernández-Mijares A, Ascaso JF, Blasco M, Brea Á, Díaz Á, Mantilla T, et al. Grupo de trabajo sobre Dislipidemia Aterogénica, Sociedad Española de Arteriosclerosis. Residual cardiovascular risk of lipid origin. Components and pathophysiological aspects. Clin Investig Arterioscler. 2018;pii: S0214-9168: 30085–30088.

39. Hatanaka Y, Shimokata K, Osugi S, Kaneko N. Impact of drinking and smoking habits on cerebrovascular disease risk among male employees. Sangyo Eiseigaku Zasshi. 2016; 58: 155–163. https://doi.org/10.1539/sangyoeisei.B15024 PMID: 27488512

40. Huang Y, Zhang L, Lian G, Zhan R, Xu R, Huang Y,et al. A novel mathematical model to predict prognosis of burnt patients based on logistic regression and support vector machine. Burns. 2016; 42: 291–299. https://doi.org/10.1016/j.burns.2015.08.009 PMID: 26774603

41. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018; 19: 270. https://doi.org/10.1186/s12859-018-2264-5 PMID: 30016950

42. Buntine W. A theory of learning classification rules. Doctoral dissertation, Department of Computer Science. University of Technology, Sydney, Australia; 1990.

43. Harakawa R, Ogawa T, Haseyama M, Akamatsu T. Automatic detection of fish sounds based on multistage classification including logistic regression via adaptive feature weighting. J Acoust Soc Am. 2018; 144: 2709. https://doi.org/10.1121/1.5067373 PMID: 30522274

44. Olesen TK, Denys MA, Goessaert AS, Bruneel E, Decalf V, Helleputte T, et al. Development of a multivariate prediction model for nocturia, based on urinary tract etiologies. Int J Clin Pract. 2018;e13306. https://doi.org/10.1111/ijcp.13306 PMID: 30556626

45. Domingos P. Bayesian averaging of classifiers and the overfitting problem. ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco, CA, USA, 2000; 223–230.

46. Liu H, Yang Y, Xia Y, Zhu W, Leak RK, Wei Z, et al. Aging of cerebral white matter. Ageing Res Rev. 2017; 34: 64–76. https://doi.org/10.1016/j.arr.2016.11.006 PMID: 27865980

47. Gepner AD, Young R, Delaney JA, Budoff MJ, Polak JF, Blaha MJ, et al. Comparison of carotid plaque score and coronary artery calcium score for predicting cardiovascular disease events: The multi-ethnic study of atherosclerosis. J Am Heart Assoc. 2017; 6.