Prediction and predictor elucidation of the onset of metabolic syndrome among young workers using machine learning techniques: A nationwide study in Japan

Miyuki Suda, MS^{1*}; Tadao Ooka, MD, PhD^{1*}; Zentaro Yamagata, MD, PhD¹

¹Department of Health Science, University of Yamanashi, Chuo, Yamanashi, Japan

* These authors contributed equally to this manuscript.

Correspondence: Tadao Ooka, Department of Health Science, University of Yamanashi, Chuo, Yamanashi, Japan

Email: tohoka@yamanashi.ac.jp

AUTHOR CONTRIBUTIONS

M.S. and T.O. participated in the design and conception of the study and its coordination, data acquisition, statistical analysis, and manuscript drafting. T.O. and Z.Y. reviewed the analysis and manuscript.

Abstract

Objectives Predictive models for the onset of metabolic syndrome (MS) among people in their 30s are scarce. This study aimed to construct a highly accurate model to predict MS onset by age 40 years and to identify the important predictors of MS onset using health checkup data of Japanese company employees aged 30 and 35.

Methods The study included 6,048 Japanese employees aged 40 years who had undergone periodic health examinations over 10 years. We developed prediction models for MS onset using machine learning methods including the random forest and logistic regression, and evaluated the models using the receiver operating characteristics and precision-recall curves. For the random forest models, the variable importance of each explanatory variable was calculated to identify important predictors of MS onset.

Results The random forest had higher predictive power than logistic regression in all models, although the differences were non-significant. Regarding important predictors, diastolic blood pressure was the most important predictor of MS onset for men aged 30 and 35 years, while body mass index was the most important predictor for women aged 30 and 35 years.

Conclusions We created a machine learning model to predict MS onset at age 40 with high accuracy from health examination data obtained at ages 30 or 35. Sex differences in important predictors of MS onset was shown by the variable importance indices of the random forest.

Applying our model in routine healthcare management should provide early and appropriate health interventions to prevent MS onset in young people.

Keywords: metabolic syndrome, obesity, health check, machine learning, prediction model

1. INTRODUCTION

Metabolic syndrome (MS) is a combination of metabolic disorders, including obesity, hyperglycemia, hypertension, and lipid abnormalities, which predispose to diabetes and cardiovascular disease.¹ There are over 1 billion persons with MS worldwide, indicating that a considerable proportion of the world's population is at risk of MS.² Therefore, preventing MS is an important global issue from a public health perspective.

In 1999, the World Health Organization (WHO) proposed criteria for the diagnosis of MS.³ Since then, two approaches to the diagnostic criteria for MS have been advanced. The first is based on the WHO concept, which includes insulin resistance and visceral fat.^{3,4} The second is based on the overlap of risk factors for cardiovascular disease, including obesity and hypertension.^{5,6} The former concept is mainly used in Japan⁴, while the latter is the mainstream in the U.S. and in European countries.

In Japan, people aged 40 years or older have the option to receive health guidance as a preventive measure against MS in accordance with national law. However, there are no MS prevention measures for people aged less than 40 years. Previous studies have reported that the basis for MS is established before the age of 40⁷ and that the lifestyle in the 30s is associated with the development of MS in the 40s.⁸⁻¹⁰ In addition, a previous study on weight control suggests that health education by age 35 leads to weight reduction at age

40.¹¹ Furthermore, other studies have reported the importance of developing a good lifestyle habits before the age of 40 to prevent weight gain in older age.¹²

Therefore, identifying individuals at high risk of MS in their 30s and intervening to improve their lifestyle habits should be useful in preventing MS. However, most previous studies for identifying MS risk were conducted with the data of people aged over 40 years because of the difficulty of obtaining health information from young people. Moreover, to the best of our knowledge, there is no report on the construction of predictive models for MS onset in young people in their 30s although this seems important for detecting early MS development.

In this study, we obtained longitudinal health examination data of men and women from a Japanese company; the data collection began in their 30s and had been recorded for over 10 years. We then built a prediction model for MS onset for young people by applying machine learning methods to these data. In addition, by using a highly interpretable machine learning method¹³ and comparing the detected variables to clinically well-known factors, we confirmed the validity of the models and identified important factors associated with the development of MS in men and women in their 30s.

2. METHODS

2.1. Study design and participants

The study included Japanese employees who were 30 years old in 2008 or 2009 and underwent continuous periodic health examinations conducted from 2008 to 2019 by *Health Insurance Association A*, which is in charge of health management among 525 business sites throughout the nation (**Figure 1**).

For each year, we excluded participants who could not be evaluated for MS because of missing values or who were already identified with MS at age <40. If participants underwent two medical examinations in the same year, only the data point with fewer missing values or the first data of the year (when the number of missing values was the same in the two examinations) were included in the analysis.

Finally, we prepared two datasets for the analyses in this study. The first dataset was created by combining the data from the health examination at age 30 with the MS evaluation data at age 40. The second dataset was created by combining data from health examinations at age 35 with the MS evaluation data at age 40. We used these two datasets to construct prediction models for MS onset at age 40 and to examine important predictors of MS onset.

2.2. Measures

2.2.1. Outcome

The main outcome was the onset of MS at age 40. The diagnostic criteria for MS were based on the 2005 Japanese Journal of Internal Medicine criteria,⁴ which is the most widely used criteria in Japan. For the diagnosis of MS, a notably large abdominal circumference (male ≥ 85 cm, female ≥ 90 cm) was set as a mandatory item, and at least two items among blood pressure (systolic blood pressure ≥ 130 mmHg or diastolic blood pressure ≥ 85 mmHg), lipid levels (triglycerides ≥ 150 mg/dL or high-density lipoprotein cholesterol [HDL-C] ≤ 40 mg/dL), and blood glucose level (fasting blood glucose ≥ 110 mg/dL) needed to exceed the standard values. In addition, patients receiving medication for blood pressure, lipids, and blood glucose were considered to meet the criteria for each item, even if the standard values were not exceeded.

2.2.2. Predictive variables

In constructing the prediction model, we used 16 examination items or 12 interview items from the health examination prescribed by Japanese law. The examination items included (1) body mass index; (2) waist circumference; (3) systolic blood pressure; (4) diastolic blood pressure; (5) high-density lipoprotein cholesterol (HDL-C); (6)

low-density lipoprotein cholesterol (LDL-C); (7) triglycerides; (8) alanine aminotransferase (ALT); (9) aspartate aminotransferase (AST); (10) γ -glutamyl transpeptidase (γ -GTP); (11) blood glucose; (12) hematocrit; (13) hemoglobin; (14) red blood cells; (15) white blood cells; and (16) uric acid.

The interview items included (1) daily alcohol consumption; (2) having breakfast; (3) paying attention to nutritional balance; (4) walking for more than 1 h per day; (5) walking speed; (6) intention to improve health; (7) eating before bed; (8) restful sleep; (9) eating too fast; (10) cigarette smoking; (11) weight gain >10 kg; and (12) exercising more than twice a week

Physical measurements and blood tests were treated as continuous variables, while all questionnaire items were treated as binary variables.

2.2.3. Statistical analysis

The prediction models for MS onset were created using machine learning methods, including random forest (RF) and logistic regression (LR), and were evaluated with the area under the receiver operating characteristics curve (AU-ROC) and the area under the precision recall curve (AU-PRC).

Training data were used to create the machine learning models, and test data were

used to check model accuracy. The training and test datasets were randomly selected from the original dataset in a 4:1 for men and 1:1 ratio for women because of the small number of MS cases for women. In all models, the outcome was the presence or absence of MS at age 40, and the 28 items, including all examination items and interview items, were used as explanatory variables.

In the construction of the prediction model, RF modelling was performed using the *randomForest* package of the statistical software R. The number of decision trees (*ntree* parameter) was set to 1000, and the number of features used to create the decision trees (*mtry* parameter) was automatically set by the R *caret* package.¹⁴ The analysis was performed with 10-segment cross-validation.

In the LR model, in order to compare the performance under the same conditions as RF, all variables were used as explanatory variables with the forced entry method without selecting variables based on prior knowledge. To avoid the problem of complete separation caused by the small number of female cases, Firth's bias-reduced logistic regression was used to create the LR model for the 30-year-old women.¹⁵

When we created RF models, the variable importance of each explanatory variable was calculated to identify important predictors of MS. In the calculation of variable importance, we used RF with conditional inference trees (using the *cForest* package¹⁶)

because RF has a tendency to underestimate categorical variables with fewer categories. In addition to creating the predictive models, we used multidimensional scaling (MDS) to evaluate the similarity between MS and non-MS patients.¹⁷ All analyses were performed using R version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria), and set the significance level to 0.05.

2.2.4. Ethical Considerations

In this study, we used health checkup data collected by Health Insurance Association A. All datasets used in this study were anonymized and statistically processed in systems that were unconnected to any external networks or the Internet. Personal information was strictly protected and managed in accordance with the ethical guidelines established by the government (Ethical Guidelines for Medical Research Involving Human Subjects).¹⁸ This study was approved by the Ethics Committee of the University of Yamanashi (Ethics Committee receipt number R01688). The study was also approved by the Ethics Committee of Health Insurance Association A (receipt number 2019-002). All participants were provided the opportunity to opt out.

3. RESULTS

Of the 6248 participants who underwent a physical examination at age 30 and 6235 participants who underwent a physical examination at age 35, 2998 (2342 men and 656 women) at age 30 and 4045 (3098 men and 947 women) at age 35 had MS assessment data at age 40 without developing MS (**Figure 1**).

When the blood test data at age 30 were compared between the MS-onset and non-MS-onset groups at age 40, all the blood test data except HDL-C were significantly higher in the MS-onset group, and HDL-C was significantly lower in the MS-onset group (**Table 1**). The blood test data at age 35 were also compared between the MS-onset and non-MS-onset groups at age 40, with the same results as at age 30.

In terms of the health examination questions at ages 30 and 35, we had significantly higher prevalence of “yes” among 6 questions (“Drinking alcohol every day” “Not having breakfast,” “Intention to improve,” “Gained more than 10 kg in weight,” “Eating too fast,” and “Smoke cigarettes”) (**Table 1**). We compared the predictive accuracy of the models using AU-ROC and AU-PRC, and we found that the RF had a higher AU-ROC and AU-PRC than LR in all the models, although the differences in accuracy were not significant (**Table 2 and Figure 2**).

We assessed the importance of the predictors by calculating the variable

importance of the explanatory variables in the RF models, and diastolic blood pressure was shown to be the most important predictor of MS onset in men aged 30 and 35 years. LDL-C, BMI, HDL-C, waist circumference, and walking time were also identified as important factors for predicting MS onset among male participants aged 30 years. HDL-C and skipping breakfast were also revealed as important factors among male participants aged 35 years. For female participants aged 30 and 35 years, BMI, waist circumference, uric acid level, and triglyceride level were the most important predictors of MS onset (**Figure 3**).

The MDS plots of the male participants at ages 30 and 35 were divided into two major clusters, and most of the MS-onset group was included in the right cluster of the plot. On the other hand, in the MDS plots of the female participants at ages 30 and 35, non-MS cases formed a single cluster and MS-onset cases were scattered away from the non-MS-onset cluster (**Figure 4**).

4. DISCUSSION

In this study, we applied the RF and LR machine learning methods to 10-year longitudinal health checkup data from Japanese companies and developed models to predict MS onset at age 40 from health checkup data of individuals aged 30 and 35 years. The accuracy of the prediction models using RF was higher than that of models using LR.

When we examined the predictors associated with MS onset, we found the most important predictors that were common to both the RF and LR models. The MDS plots using the RF method showed different characteristics between the participants with MS and those without MS onset.

In this study, the accuracy of the RF prediction model was higher than that of the LR model. When LR is used for prediction, the existence of a linear relationship is assumed between the outcome and explanatory variable. On the other hand, RF does not necessarily require this assumption because of its calculation methods.¹⁹ Furthermore, RF has been reported to have better performance than other methods when there are many explanatory variables and interactions between variables.²⁰ In the present study, 28 explanatory variables were used to create the model, and a group of variables with interactions (e.g., BMI and waist circumference) were included. These may be why the accuracy of predicting the onset of MS was higher in the RF models than in the LR models.

In a previous study, the predictors of MS onset were investigated using regression analysis in Japanese men aged 30, 35, and 40 years, and increase in BMI was reported as the most important predictor.¹⁰ However, in the present study, diastolic blood pressure was presented as the most important predictor. The reason for the difference in the results may be due to multicollinearity that occurred differently in the RF and LR models. A previous

study showed that RF has better performance for nonlinear relationships than regression analysis (Cox-proportional hazards regression).²¹ This characteristic of RF may lead to the presence of different factors for MS prediction.

Another previous study using the neural network method also showed diastolic blood pressure as an important predictor,²² and another study also reported that diastolic blood pressure is associated with the development of MS at a young age.²³ These studies support the results of the present study. The value of diastolic blood pressure may be more useful for judging the future onset of MS than increase in BMI because blood pressure can be ascertained at one point, while increase in BMI cannot be measured at one time.

In the analyses of female participants, the most important predictor of MS was BMI, followed by abdominal circumference, as noted in previous studies.¹⁰ Abdominal circumference is the criterion for determining the onset of MS, and the criteria for abdominal circumference are stricter in women than in men. Therefore, the importance of abdominal circumference and BMI might be higher in female participants than in male participants.

Among the questionnaire items, walking time in men at age 30, skipping breakfast in men at age 35, walking speed in women at age 30, and not feeling rested from sleep in women at age 35 were identified as important predictors. A previous study reported that

daily exercise habits, regular diet, and restful sleep were associated with the development of MS in both men and women.²⁴ This finding is consistent with the results of the current study. To evaluate these items, we used RF with conditional inference trees to calculate the importance of variables, which enabled us to select important variables with only a few categories that were difficult to evaluate with the conventional RF.

In the MDS plot of female participants, MS cases were sporadically located away from the cluster. This sporadic population may represent a future unhealthy population. However, in the plot of male participants, MS cases were concentrated in one of two separate clusters; thus, clusters with a high concentration of MS cases may represent future unhealthy populations. The differences between MS and non-MS participants in the MDS plot were more conspicuous among female participants than among male participants. This may be because the criteria for determining MS were stricter in women than in men, and the women who will experience MS onset at age 40 had more distinctive characteristics than men in the same situation.

The strength of our study is our construction of models with machine learning methods to predict the onset of MS in men and in women using large longitudinal health examination data of young people collected over a 10-year period at a Japanese company. In addition, by using a highly interpretable machine learning method, we were able to

identify important predictors from many health examination items. To our knowledge, this is the first study to develop prediction models with machine learning methods to predict the onset of MS using longitudinal data of men and women in their 30s.

4.1. Limitations

This study has several limitations. First, there is a possibility of selection bias because we developed and evaluated the prediction model using only the data of healthy employees in a large company. Second, it is difficult to apply the same result to populations in other countries because this study used only Japanese health examination data and the MS criteria used in Japan. Third, we could not secure a sufficient number of female subjects, and the prediction models might be less accurate for the female population than for the male population. However, we used Firth's bias reduction method to reduce the problem of complete separation caused by the small number of cases. Fourth, the importance of some variables, especially questionnaire items, might not have been properly evaluated because the variable importance of some items with few categories tends to be underestimated in RF models. However, we used the RF method with conditional inference trees to evaluate the importance of items with few categories.

4.2. Clinical indication

By applying the prediction model developed in this study to the health checkup data of men and women in their 30s, it may be possible to prevent MS onset in an effective way. For instance, companies and municipalities with limited medical sources can identify high-risk groups for MS by applying our models to their data. In this study, in addition to BMI and waist circumference, diastolic blood pressure, LDL-C, and HDL-C in men and uric acid and triglyceride in women were noted as important predictors of MS. In addition, walking habits in both men and women, skipping breakfast in men, and restful sleep in women were also presented as important factors.

In this study, we could not determine a causal relationship between these factors and the development of MS. However, based on these results, we can prevent MS onset efficiently by focusing on items by sex when conducting health guidance for people in their 30s. Currently, the Japanese legal health checkup does not include items for determining MS in people in their 30s. However, by applying our prediction model to the health examination of young people, it may be possible to identify high-risk groups for MS at an early stage. Furthermore, by providing health guidance to young people at high risk, focusing on the predictors identified in this study, the onset of MS can be prevented, resulting in a reduction of the nation's healthcare costs.

293

294 **5. CONCLUSION**

295 We developed a high-accuracy prediction model with a machine learning method
296 that predicts the onset of MS at age 40 based on health examination data obtained at ages
297 30 and 35. Some important sex-specific predictors were identified using this highly
298 interpretable machine learning method. Applying our models to routine healthcare
299 management should provide early and appropriate health interventions to young people
300 for preventing the onset of MS in this population.

301

302 **ACKNOWLEDGMENT**

303 We are grateful to all the staff of Health Insurance Association A for preparing the dataset
304 for the current study.

305 **CONFLICT OF INTEREST**

306 There are no conflicts of interest to declare.

307 **DISCLOSURES**

308 The study protocol was examined and approved by the Ethics Committee of the University
309 of Yamanashi (Ethics Committee receipt number R01688). The study was also approved by
310 the Ethics Committee of Health Insurance Association A (receipt number 2019-002). All

participants had the opportunity to opt out.

REFERENCES

1. Neuhauser HK. The metabolic syndrome. *Lancet* 2005;366:1415-1428.
2. Delavari A, Forouzanfar MH, Alikhani S, Sharifian A, Kelishadi R. First nationwide study of the prevalence of the metabolic syndrome and optimal cutoff points of waist circumference in the middle east: The national survey of risk factors for noncommunicable diseases of Iran. *Diabetes Care* 2009;32:1092-1097.
3. World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus. Geneva: World Health Organization; 1999. <http://www.who.int/iris/handle/10665/66040>. Accessed September 28, 2021.
4. Japanese Society of Internal Medicine. Definition and the diagnostic standard for metabolic syndrome--Committee to Evaluate Diagnostic Standards for Metabolic Syndrome. *Nihon Naika Gakkai Zasshi* 2005;94:794-809. Japanese
5. Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JJ, Donato KA, et al. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and

International Association for the Study of Obesity. *Circulation*. 2009;120:1640-1645

6. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, et al.

Diagnosis and management of the metabolic syndrome: an American Heart Association

National Heart, Lung, and Blood Institute Scientific Statement. *Circulation*.

2005;112:2735-2752.

7. Tetutaro S. Metabolic syndrome prevention measures. *Syoujinkai Medical Journal*

2009;48:150-155.

8. Carnethon MR, Loria CM, Hill JO, et al. Risk factors for the metabolic syndrome.

Diabetes Care 2004;27:2707-2715.

9. Hatanaka Y, Tamakoshi A, Tsushita K. Impact of body mass index on men in their 20s

and the effects of subsequent changes in body weight upon the rates of hypertension and

diabetes and medical costs in their 40s. *Sangyo Eiseigaku Zasshi*. 2012;54:141-149.

Japanese.

10. Kadoya M, Nakamura S, Sugio Y, et al. Increased body mass index in young adults is

associated with metabolic syndrome. *Sangyo Eiseigaku Zasshi* 2014;56:121-127.

11. Nishinoue N, Tatemichi M, Aratake M, Yamazaki A, Fukuda H, Sugita M. Effect of a

group health education program in the workplace at the age of 35: a population

approach. *J UOEH* 2011;33:23-34.

- 347 12. Maeda S. The forefront of disease genome research: A new era of genetic analysis
348 research. Igakuno Ayumi 2018;266:435-440
- 349 13. Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V,
350 Nantasenamat C. Predicting Metabolic Syndrome Using the Random Forest Method.
351 The Scientific World Journal 2015;10
- 352 14. Kuhn M, Wing J, Weston S, et al. Package 'caret': Classification and regression training.
353 2021. <https://cran.r-project.org/web/packages/caret/caret.pdf>. Accessed September 29,
354 2021.
- 355 15. Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80:27-38.
- 356 16. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance
357 for random forests. BMC Bioinformatics 2008;9:307.
- 358 17. Shalab H. Statistical learning from a regression perspective. J R Stat Soc Ser A
359 Statistics Soc 2009;172:935
- 360 18. Ethical Guidelines for Medical Research Involving Human Subjects. Ministry of
361 Education, Culture, Sports, Science and Technology, December 22, 2007 (corrected by
362 the Ministry on February 28, 2007) Ministry of Health, Labour and Welfare. Accessed
363 September 28, 2021.
- 364 19. Murphy KP. Machine learning. Cambridge: The MIT Press; 2012.

20. Yang P, Hwa Yang Y, Zhou BB, Zomaya AYA. A review of ensemble methods in bioinformatics. *Curr Bioinform* 2010;5:296-308.
21. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. *BMJ Nutr Prev Health*. 2021;4:140-148.
22. Hirose H, Takayama T, Hozawa S, Hibi T, Saito I. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. *Comput Biol Med* 2011;41:1051-1056.
23. Lin JD, Chen YL, Wu CZ, et al. Identification of normal blood pressure in different age group. *Medicine (Baltimore)* 2016;95:e3188.
24. Ashizawa E, Katano S, Harada A, et al. Exploring the link between standard lifestyle questionnaires administered during specific medical check-ups and incidence of metabolic syndrome in Chiba Prefecture. *Nihon Kosho Eisei Zasshi*. 2014;61:176-185. Japanese.

380 **Table 1. Characteristics of study participants**

				30 years old without MS			35 years old without MS			40 years old with/without MS		
				40 years MS(-)	40 years MS(+)		40 years MS(-)	40 years MS(+)		40 years MS(-)	40 years MS(+)	
				n	2834	164	n	3865	180	n	5619	375
				male	2186	156	male	2927	171	male	3857	351
				female	648	8	female	938	9	female	1762	24
Examination data				40 years MS(-)	40 years MS(+)	p-value†	40 years MS(-)	40 years MS(+)	p-value†	40 years MS(-)	40 years MS(+)	p-value†
Body mass index, kg/m ²				21.58 ± 2.87	25.62 ± 3.52	<0.001	22.15 ± 3.14	26.61 ± 3.59	<0.001	22.71 ± 3.41	29.17 ± 4.39	<0.001
Waist circumference, cm				76.38 ± 8.02	87.18 ± 8.81	<0.001	78.57 ± 8.70	90.44 ± 8.80	<0.001	80.13 ± 9.33	97.04 ± 9.81	<0.001
Systolic blood pressure, mmHg				112.89 ± 12.05	123.40 ± 12.33	<0.001	114.28 ± 11.98	125.96 ± 12.04	<0.001	115.70 ± 12.46	135.24 ± 13.04	<0.001
Diastolic blood pressure, mmHg				67.81 ± 8.71	75.97 ± 9.34	<0.001	69.78 ± 9.05	78.92 ± 9.82	<0.001	71.81 ± 9.96	87.89 ± 10.40	<0.001
Triglycerides, mg/dL				82.49 ± 59.09	158.97 ± 176.22	<0.001	91.86 ± 70.12	168.82 ± 130.09	<0.001	95.81 ± 73.83	230.13 ± 144.77	<0.001
HDL-C, mg/dL				60.95 ± 13.76	52.30 ± 13.52	<0.001	59.91 ± 14.11	50.07 ± 11.73	<0.001	61.51 ± 14.86	47.05 ± 10.97	<0.001
LDL-C, mg/dL				108.75 ± 28.12	127.63 ± 32.57	<0.001	116.07 ± 30.35	132.56 ± 35.15	<0.001	118.99 ± 30.76	135.47 ± 32.94	<0.001
Blood glucose, mg/dL				88.14 ± 7.75	92.27 ± 12.55	<0.001	89.39 ± 8.32	94.69 ± 9.40	<0.001	89.81 ± 9.52	107.01 ± 33.61	<0.001
Uric acid, mg/dL				5.55 ± 1.30	6.58 ± 1.21	<0.001	5.61 ± 1.32	6.98 ± 1.19	<0.001	5.68 ± 1.47	7.13 ± 1.42	<0.001
Alanine aminotransferase, U/L				21.22 ± 16.70	39.18 ± 29.54	<0.001	23.19 ± 19.37	46.42 ± 35.53	<0.001	24.62 ± 26.00	55.19 ± 37.83	<0.001
Aspartate aminotransferase, U/L				20.03 ± 7.56	26.49 ± 12.69	<0.001	21.09 ± 9.68	28.82 ± 14.57	<0.001	21.99 ± 13.06	34.85 ± 23.01	<0.001
γ-glutamyl transpeptidase, U/L				26.75 ± 23.95	49.27 ± 47.14	<0.001	30.83 ± 31.05	59.06 ± 51.54	<0.001	35.31 ± 42.73	85.09 ± 83.40	<0.001
Hemoglobin, g/dL				14.75 ± 1.36	15.54 ± 0.97	<0.001	14.76 ± 1.39	15.70 ± 0.96	<0.001	14.51 ± 1.51	15.86 ± 1.05	<0.001
Hematocrit, %				45.07 ± 3.77	47.34 ± 2.95	<0.001	44.71 ± 3.83	47.40 ± 2.98	<0.001	44.20 ± 4.10	47.81 ± 3.01	<0.001
Red blood cells, 10 ⁴ /μL				488.39 ± 41.39	514.19 ± 32.48	<0.001	487.49 ± 40.76	515.29 ± 33.64	<0.001	481.58 ± 42.38	520.45 ± 36.19	<0.001
White blood cells, 10 ³ /μL				58.16 ± 14.32	65.87 ± 16.06	<0.001	59.63 ± 17.12	67.13 ± 18.43	<0.001	59.48 ± 15.83	71.63 ± 17.67	<0.001
Questionnaire				40 years MS(-)	40 years MS(+)	p-value*	40 years MS(-)	40 years MS(+)	p-value*	40 years MS(-)	40 years MS(+)	p-value*
Drinking alcohol every day				(%)	Yes	0.033	466 (12.1)	33 (18.4)	0.019	771 (13.7)	57 (15.2)	0.439
					No		3383 (87.9)	146 (81.6)		4844 (86.3)	318 (84.8)	
					(86.6)							
Having breakfast				(%)	Yes	0.032	2771 (72.0)	108 (60.3)	0.001	4200 (74.8)	247 (66.0)	<0.001
					No		1080 (28.0)	71 (39.7)		1414 (25.2)	127 (34.0)	
Paying attention to nutritional balance				(%)	Yes	0.287	3162 (82.2)	149 (83.2)	0.765	4661 (83.0)	300 (80.2)	0.157
					No		687 (17.8)	30 (16.8)		952 (17.0)	74 (19.8)	
Walking more than 1 hour per day				(%)	Yes	0.237	2524 (65.6)	114 (63.7)	0.63	3538 (63.0)	241 (64.3)	0.659
					No		1324 (34.4)	65 (36.3)		2075 (37.0)	134 (35.7)	
Walking speed				(%)	Yes	0.228	2011 (52.2)	100 (55.9)	0.359	2851 (50.8)	236 (62.9)	<0.001
					No		1839 (47.8)	79 (44.1)		2762 (49.2)	139 (37.1)	
Intention to improve health				(%)	Yes	0.007	2687 (69.8)	148 (82.7)	<0.001	4086 (72.8)	322 (85.9)	<0.001
					No		1161 (30.2)	31 (17.3)		1528 (27.2)	53 (14.1)	
Eating before bed				(%)	Yes	0.468	1971 (51.2)	108 (60.7)	0.014	2355 (42.0)	201 (53.7)	<0.001
					No		1875 (48.8)	70 (39.3)		3256 (58.0)	173 (46.3)	
Getting rest from sleep				(%)	Yes	0.627	1640 (42.6)	74 (41.3)	0.758	2350 (41.9)	154 (41.1)	0.787
					No		2206 (57.4)	105 (58.7)		3262 (58.1)	221 (58.9)	
Eating too fast				(%)	Yes	0.003	1180 (30.6)	87 (48.6)	<0.001	1836 (32.7)	177 (47.2)	<0.001
					No		2670 (69.4)	92 (51.4)		3779 (67.3)	198 (52.8)	
Smoke cigarettes				(%)	Yes	0.001	1008 (26.2)	63 (35.2)	0.009	1349 (24.0)	136 (36.3)	<0.001
					No		2842 (73.8)	116 (64.8)		4265 (76.0)	239 (63.7)	
Gained more than 10 kg in weight				(%)	Yes	<0.001	915 (24.9)	117 (66.9)	<0.001	1936 (35.8)	307 (83.2)	<0.001
					No		2760 (75.1)	58 (33.1)		3473 (64.2)	62 (16.8)	
Exercise more than twice a week				(%)	Yes	0.004	3257 (84.6)	147 (82.6)	0.459	4650 (82.8)	312 (83.2)	0.888
					No		594 (15.4)	31 (17.4)		965 (17.2)	63 (16.8)	

381 Values are presented as mean ± SD or as n(%)

382 †Chi-square test

383 MS, metabolic syndrome; HDL-C, High-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol

Table 2. Predictive accuracy of the random forest and logistic regression models

		Random forest	Logistic regression	p-value [†]
Prediction for 30-year-old men	AU-ROC	0.867	0.852	0.64
	AU-PRC	0.392	0.356	
Prediction for 35-year-old men	AU-ROC	0.876	0.852	0.15
	AU-PRC	0.444	0.427	
Prediction for 30-year-old women	AU-ROC	0.985	0.748	0.31
	AU-PRC	0.550	0.377	
Prediction for 35-year-old women	AU-ROC	0.981	0.940	0.16
	AU-PRC	0.613	0.271	

[†]DeLong's test for two correlated ROC curves

AU-ROC, area under the receiver operating characteristic; AU-PRC, area under the precision-recall curve

Figure Legends

Figure 1

Participant selection flow

*Study participants who could not be assessed for MS because of some missing

values.

MS, metabolic syndrome

Figure 2

Receiver operating characteristic (ROC) curve and precision-recall (PR) curve for

predicting the onset of metabolic syndrome via the random forest and logistic

regression models.

Figure 3

Important predictors of metabolic syndrome onset with values of important

variables calculated using the random forest method.

Figure 4

Clustering of young workers with the multidimensional scaling (MDS) plot using the

407 random forest method.

Figure 1

Figure 2

Figure 3

Important predictors among 30-year-old men

Important predictors among 30-year-old women

Important predictors among 35-year-old men

Important predictors among 35-year-old women

Figure 4

MDS plot: 30-year-old men

MDS plot: 30-year-old women

MDS plot: 35-year-old men

MDS plot: 35-year-old women