

Original Paper

Classifying Schizophrenia Cases by Artificial Neural Network Using Japanese Web-Based Survey Data: Case-Control Study

Yupeng He¹, MD, PhD; Masaaki Matsunaga¹, MD, PhD; Yuanying Li², PhD; Taro Kishi³, MD, PhD; Shinichi Tanihara⁴, PhD; Nakao Iwata³, MD, PhD; Takahiro Tabuchi⁵, MD, PhD; Atsuhiko Ota¹, MD, PhD

¹Department of Public Health, Fujita Health University School of Medicine, Toyoake, Japan

²Department of Public Health and Health Systems, Nagoya University Graduate School of Medicine, Nagoya, Japan

³Department of Psychiatry, Fujita Health University School of Medicine, Toyoake, Japan

⁴Department of Public Health, Kurume University School of Medicine, Kurume, Japan

⁵Cancer Control Center, Osaka International Cancer Institute, Osaka, Japan

Corresponding Author:

Yupeng He, MD, PhD

Department of Public Health

Fujita Health University School of Medicine

1-98 Dengakugakubo

Kutsukake-cho

Toyoake, 470-1192

Japan

Phone: 81 562 93 2476

Fax: 81 562 93 3079

Email: yupeng.he@fujita-hu.ac.jp

Abstract

Background: In Japan, challenges were reported in accurately estimating the prevalence of schizophrenia among the general population. Retrieving previous studies, we investigated that patients with schizophrenia were more likely to experience poor subjective well-being and various physical, psychiatric, and social comorbidities. These factors might have great potential for precisely classifying schizophrenia cases in order to estimate the prevalence. Machine learning has shown a positive impact on many fields, including epidemiology, due to its high-precision modeling capability. It has been applied in research on mental disorders. However, few studies have applied machine learning technology to the precise classification of schizophrenia cases by variables of demographic and health-related backgrounds, especially using large-scale web-based surveys.

Objective: The aim of the study is to construct an artificial neural network (ANN) model that can accurately classify schizophrenia cases from large-scale Japanese web-based survey data and to verify the generalizability of the model.

Methods: Data were obtained from a large Japanese internet research pooled panel (Rakuten Insight, Inc) in 2021. A total of 223 individuals, aged 20-75 years, having schizophrenia, and 1776 healthy controls were included. Answers to the questions in a web-based survey were formatted as 1 response variable (self-report diagnosed with schizophrenia) and multiple feature variables (demographic, health-related backgrounds, physical comorbidities, psychiatric comorbidities, and social comorbidities). An ANN was applied to construct a model for classifying schizophrenia cases. Logistic regression (LR) was used as a reference. The performances of the models and algorithms were then compared.

Results: The model trained by the ANN performed better than LR in terms of area under the receiver operating characteristic curve (0.86 vs 0.78), accuracy (0.93 vs 0.91), and specificity (0.96 vs 0.94), while the model trained by LR showed better sensitivity (0.63 vs 0.56). Comparing the performances of the ANN and LR, the ANN was better in terms of area under the receiver operating characteristic curve (bootstrapping: 0.847 vs 0.773 and cross-validation: 0.81 vs 0.72), while LR performed better in terms of accuracy (0.894 vs 0.856). Sleep medication use, age, household income, and employment type were the top 4 variables in terms of importance.

Conclusions: This study constructed an ANN model to classify schizophrenia cases using web-based survey data. Our model showed a high internal validity. The findings are expected to provide evidence for estimating the prevalence of schizophrenia in the Japanese population and informing future epidemiological studies.

KEYWORDS

artificial neural network; schizophrenia; prevalence; Japan; web-based survey; mental health; psychosis; machine learning; epidemiology

Introduction

Schizophrenia is a common mental illness that disrupts a person's thinking processes, perceptions, emotional responsiveness, and social interactions [1]. Estimates of international prevalence range from 0.33% to 0.75% [2,3]. The lifetime prevalence and median 12-month prevalence of schizophrenia were reported to be 0.33% and 0.48%, respectively [4]. In Japan, the point prevalence of schizophrenia, including schizotypal and delusional disorders, is approximately 0.7% according to national data from a patient survey [5]. While the real prevalence was considered quite different owing to the obstacles when operating the investigation. Patients with mild cases might not seek medical attention, and some cases are diagnosed as schizophrenia just for prescriptions to pass the medical insurance review.

We envisioned whether the prevalence of schizophrenia in individuals could be predicted by several factors and estimated the prevalence in the general population. By retrieving data from previous systematic reviews and meta-analyses, we confirmed that individuals with schizophrenia experience poor subjective well-being and various physical, psychiatric, and social comorbidities. For instance, studies conducted in Canada and the United States [6,7] have reported that young adults with schizophrenia tend to experience poorer subjective well-being and lower life satisfaction. Additionally, individuals with schizophrenia are prone to higher risk of noncommunicable diseases and experience poor oral health [8-10]. Patients with schizophrenia frequently exhibit symptoms of depression and experience sleep disorders [11,12]. Furthermore, individuals with schizophrenia typically have lower employment rates [11] and exhibit challenges in social cognition [13]. These factors are strongly associated with the incidence and existence of schizophrenia [14].

Machine learning techniques have recently drawn increasing attention in psychiatric studies. Birnbaum et al [15] built machine learning diagnostic and relapse classifiers for schizophrenia based on internet search activity (timing, frequency, and content), which achieved the area under the curve value of 0.74 and 0.71, respectively. Natural language processing has been applied to detect schizophrenia signs from social media content with extremely high accuracy [16]. Lejeune et al [17] concluded in their review that studies using social media to diagnose mental disorders were promising, while limitations included lack of clinical diagnostic data, small sample size, and heterogeneity in study quality. Previous studies have reported the effectiveness of detecting various types of mental disorders [18,19]. In other epidemiological fields, machine learning techniques also manifested promise, especially excelling at dealing with large-scale data [20,21]. We recently researched developing estimation methods for schizophrenia among the Japanese population [22]. Data were collected using

a large-scale web-based survey. Individuals who participated in this survey were asked to answer questions about demographics, health-related backgrounds, physical comorbidities, psychiatric comorbidities, and social comorbidities. Compared with classical epidemiological surveys, web-based surveys make it easy to reach a large sample size and amount of data. Few studies have referred to the precise prediction of schizophrenia using large-scale web-based surveys.

If schizophrenia cases could be classified by variables of demographic and health-related backgrounds, it would be possible to estimate schizophrenia cases among general population, who are not seeking psychiatric care (namely, whose psychiatric syndromes are unknown). Therefore, we aimed to construct a machine learning model that can accurately classify the schizophrenia case and verify its generalizability.

Methods

Study Design: Participants and Survey Items

A prevalence case-control study was conducted using an internet research agency's pooled panel (Rakuten Insight, Inc, incorporated approximately 2.3 million panelists by 2022) [23]. Participants' ages were restricted from 20 to 75 years. Individuals who participated in this study answered a web-based survey.

Among participants who currently have schizophrenia, 5584 individuals who self-reported schizophrenia were sampled in the Rakuten Insight disease panel [24]. A total of 3256 respondents answered the following four questions before the survey: (1) are you currently experiencing schizophrenia only; schizophrenia and migraine; schizophrenia and a sleep disorder; or schizophrenia, migraine, and a sleep disorder? (2) Have you experienced auditory hallucinations lasting more than 1 month? (3) Have you never used stimulants or other illegal drugs and have never been an alcoholic? (4) Have you experienced your first auditory hallucination lasting more than 1 month at less than 60 years of age? Those who answered "yes" to all 4 questions were considered to have schizophrenia. Therefore, 223 participants who currently had schizophrenia were included in the survey.

For participants who do not currently have schizophrenia, all 28,000 participants in the Japan COVID-19 and Society Internet Survey (which was also conducted using the Rakuten Insight Panel) [25] were sampled. A total of 6656 respondents answered the following four questions before the survey: (1) are you currently experiencing mental illness? (2) Have you experienced a mental illness in the past? (3) Have you experienced auditory hallucinations? (4) Have you ever used stimulants or other illegal drugs, been alcoholic, or received psychiatric treatment? Those who answered "no" to all 4 questions were considered to not have schizophrenia. Therefore, 1776 participants who did not currently have schizophrenia were included in the survey.

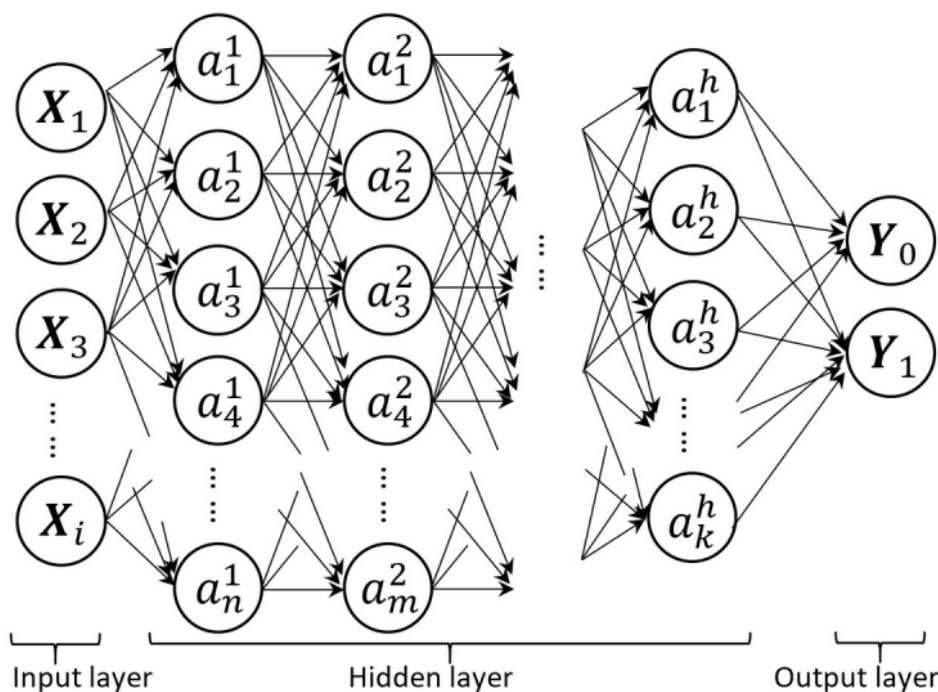
In the survey, 223 participants with schizophrenia and 1776 healthy controls answered a self-administered questionnaire. The question items were designed to assess (1) demographic and health-related backgrounds and physical comorbidities, (2) psychiatric comorbidities, and (3) social comorbidities. Answers to the question items were formatted to 1 response variable (diagnosed with schizophrenia, “yes” or “no”) and 75 feature variables (demographic, health-related backgrounds, physical comorbidities, psychiatric comorbidities, and social comorbidities). Details of the study participants and variable definitions have been published elsewhere [26] and are described in [Multimedia Appendix 1](#).

Artificial Neural Network

An artificial neural network (ANN) is a computing system that imitates the signals transmitted between neurons in biological

brains [27]. Neurons in an ANN are divided into layers: 1 input layer, several hidden layers, and 1 output layer, where the number of neurons and hidden layers is not fixed. “Signals” transmitting is accomplished by weights and activation functions. As long as the initialized weights are updated by the self-learning process, the ANN can generate a “perfect” model. On behalf of the complex structure, an ANN can capture nonlinear associations and reveal potential interactions between variables. In our study, we structured an ANN with 5 hidden layers (neurons of each layer: 128-64-32-16-8), HeNormal weight initializer [28], ReLU activation function in the hidden layers, and sigmoid activation in the output layer [29]. These settings partially referred to previous studies [20,30] ([Figure 1](#)).

Figure 1. Structure of the artificial neural network. X refers to each of the feature variables. a refers to neurons in hidden layers. Y refers to response variables during training process and prediction results during test process.



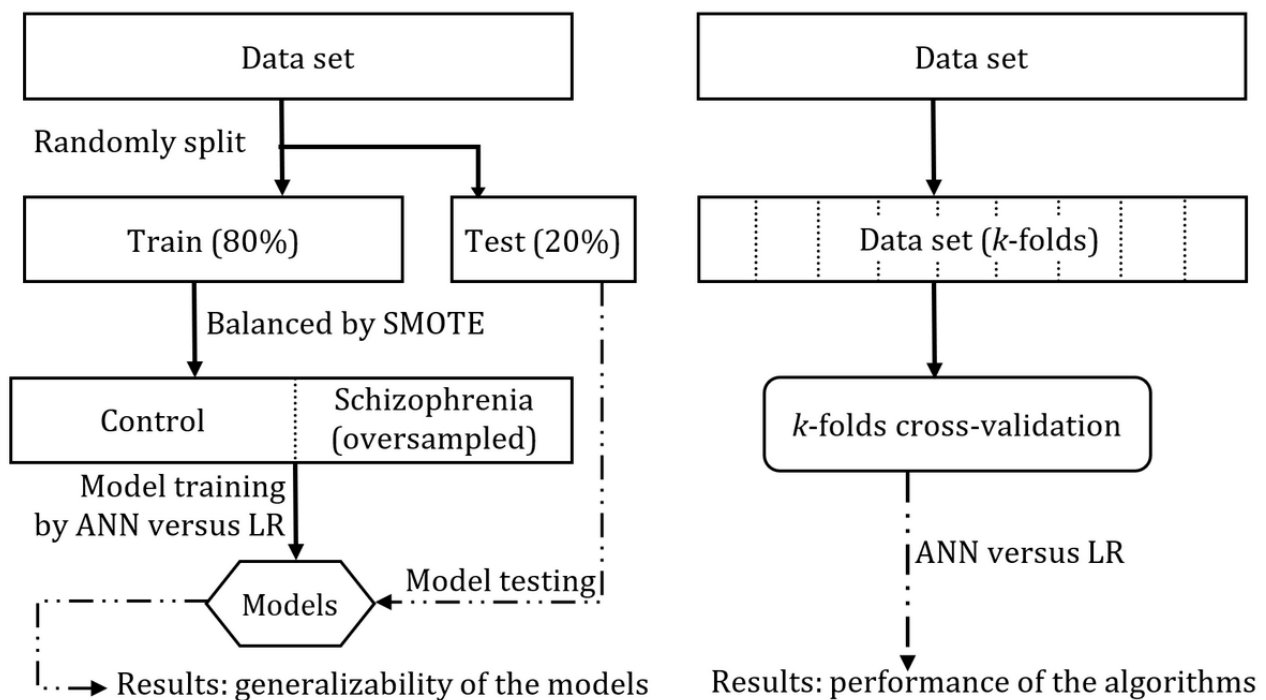
Logistic Regression

Logistic regression (LR) estimates the probability of an event (outcome variable) taking place (success) based on a given data set of independent variables. In the LR equation, the outcome variable is transformed into log odds, the natural logarithm of the probability of success divided by the probability of failure. The independent variables are linearly structured by distributing coefficients to them. These coefficients are commonly estimated via maximum likelihood estimation to optimize the best fit of the log odds. The model is fixed once the optimal coefficients are found [31]. As a typical method that is widely used in epidemiological research, LR is introduced in this study to compare its performance with the novel ANN method.

Data Processing

The data were randomly split into a training data set and a test data set at an 80:20 ratio. Before applying the 2 selected algorithms (ANN and LR), the training data set was balanced based on the synthetic minority oversampling technique [32]. Models were trained using ANN and LR on the training data set separately, and the test data set was used to evaluate the 2 models (ANN and LR model). The area under the receiver operating characteristic curve (AUC) was applied to interpret the results because the outcome was binary. The 95% CI for the AUC was generated from 10,000 bootstraps of the training data set ([Figure 2](#)).

Figure 2. Flowchart of data processing, model training, and evaluation. ANN: artificial neural network; LR: logistic regression; SMOTE: synthetic minority oversampling technique.



Evaluation

The following thresholds were used to evaluate the performance in terms of AUC score: 0.5=no discrimination, 0.5-0.7=poor discrimination, 0.7-0.8=acceptable discrimination, 0.8-0.9=excellent discrimination, and >0.9=outstanding discrimination [33]. Two strategies were designed to evaluate the performance of the models and algorithms. To compare the generalizability of the trained LR and ANN models, their AUCs on the test data sets were compared. To compare the performance of LR and ANN algorithms, the AUCs were compared based on a 10-fold cross-validation. The differences in the AUCs were tested using the Delong method [34].

Model Interpretation (Variable Importance)

To interpret the ANN model, we introduced a shuffle test for each variable to evaluate variable importance. Among all the N variables in the test data set, the n th variable is shuffled at random; this resampled test data set is applied to the ANN model, and the AUC obtained from the resampled test data set is compared with the AUC obtained from the original test data set. The difference between the 2 AUCs explains the importance of the variable. A higher difference indicates that the n th variable has relatively higher importance.

Statistical Analyses

Statistical analyses were performed using Python (version 3.7; Python Software Foundation). The computational environment

used was the Jupyter Notebook (Project Jupyter). Means and SDs are presented for continuous variables. Categorical variables are presented as proportions. Differences in means for continuous variables and categorical variables were tested using analysis of variance and chi-square test, respectively.

Ethical Considerations

This study was approved by the Bioethics Review Committee of Fujita Health University (HM21-408). All procedures performed in this study were in accordance with the Ethical Guidelines for Medical and Health Research Involving Human Subjects enforced by the Ministry of Health, Labour and Welfare, Government of Japan, and the 1964 Helsinki Declaration and its later amendments.

Results

Table 1 shows the characteristics of the normal and schizophrenia groups. Compared with the control group, participants with schizophrenia were more likely to be men. They had a significantly higher proportion of obesity, lower education levels, and lower household income. Participants with schizophrenia were more likely to have poor self-rated health status, depressive symptoms, perceived stress, and lower availability of social support. More details of these characteristics are provided in Table S1 in Multimedia Appendix 1.

Table 1. Characteristics of schizophrenia cases and healthy controls.

	Schizophrenia case (N=223)	Healthy control (N=1776)	<i>P</i> value ^a
Age (years), mean (SD)	46 (9.3)	44 (13.5)	.08
Gender, n (%)			
Women	108 (48)	975 (55)	.07
BMI (≥ 25 kg/m ²), n (%)	103 (46)	313 (18)	<.001
Education, n (%)			
Junior or senior high school or lower	94 (42)	471 (27)	<.001
Household income, n (%)			
<3 million Japanese yen ^b	108 (48)	359 (20)	<.001
Self-rated health status, n (%)			
Bad	108 (48)	386 (19)	<.001
Physical disease, n (%)			
≥ 1 disease	135 (61)	610 (31)	<.001
Depressive symptoms, n (%)			
CES-D ^c ≥ 8	156 (70)	467 (26)	<.001
Perceived stress (PSS-4 ^d), median (IQR)	10 (8-12)	7 (6-8)	<.001 ^e
Social support (ESSI ^f), median (IQR)	21 (15-28)	23 (17-28)	.007 ^e

^aBased on analysis of variance and chi-square test for continuous and categorical variables, respectively, except specified notes.

^bAround US \$20,000.

^cCES-D: Center for Epidemiological Studies Depression.

^dPSS-4: 4-item Perceived Stress Scale.

^e*P* values by the Kruskal-Wallis test.

^fESSI: ENRICH Social Support Instrument.

Figure 3 illustrates the internal validity of the models trained using the ANN and LR. The model trained by ANN performed better than LR in terms of AUC (0.86 vs 0.78), accuracy (0.93 vs 0.91), and specificity (0.96 vs 0.94), whereas the model trained by LR performed better in terms of sensitivity (0.63 vs 0.56). Table 2 shows the algorithm performance comparing

between ANN and LR by using bootstrapping and cross-validation. ANN performed better in terms of AUC (bootstrapping: 0.847 vs 0.773 and cross-validation: 0.81 vs 0.72), whereas LR performed better in terms of accuracy (0.894 vs 0.856).

Figure 3. Confusion matrixes of artificial neural network and logistic regression models. AUC: area under the receiver operating characteristic curve.

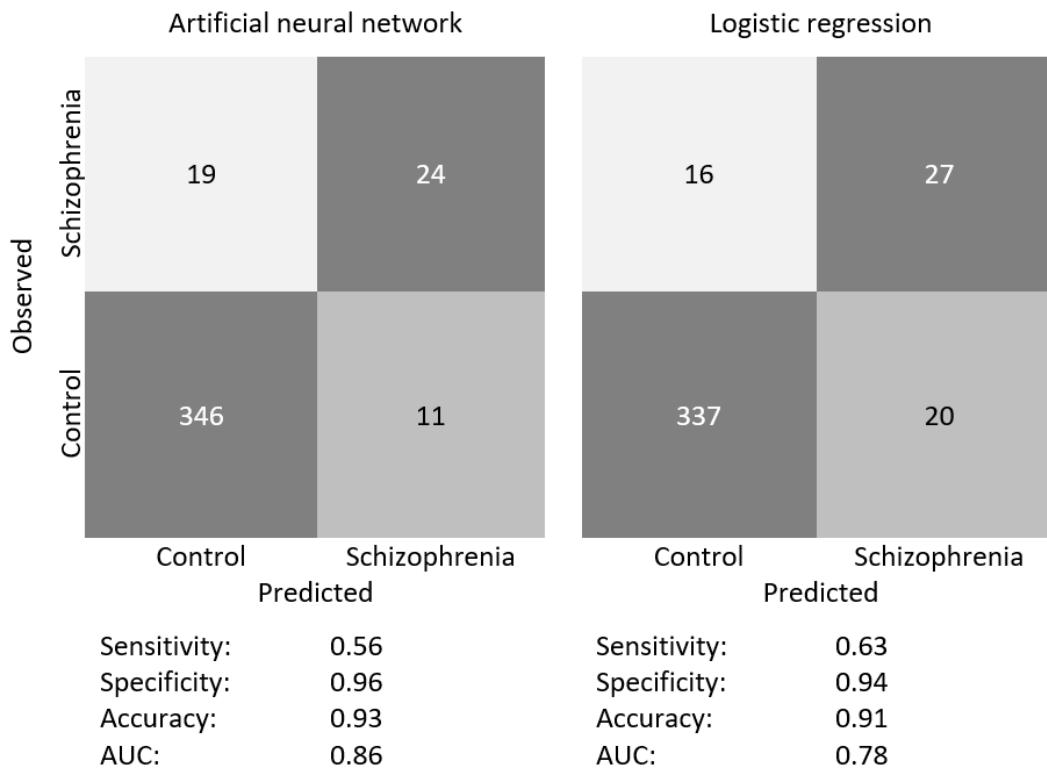


Table 2. Comparison between artificial neural network and logistic regression by bootstrapping and cross-validation.

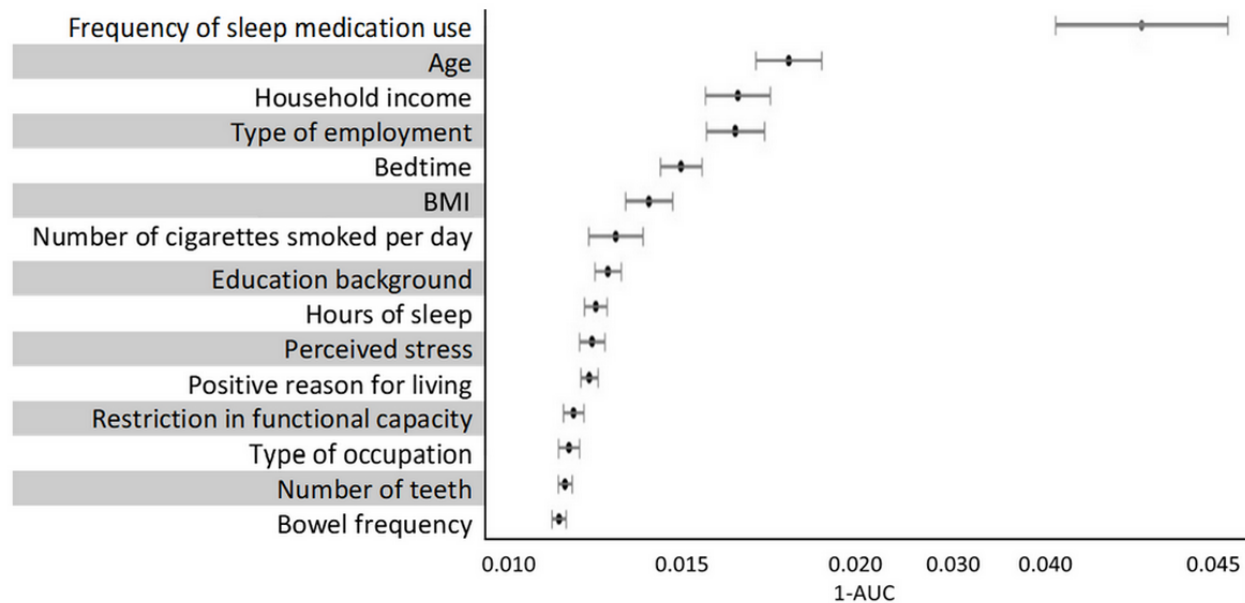
	Artificial neural network, mean (SD)	Logistic regression, mean (SD)	P value
Results from 10,000 bootstrapping			
Accuracy	0.856 (0.18)	0.894 (0.01)	<.001
AUC ^a	0.847 (0.11)	0.773 (0.02)	<.001
Results from 10-fold cross-validation			
Accuracy	0.92 (0.04)	0.82 (0.28)	.27
AUC	0.81 (0.28)	0.72 (0.26)	.49

^aAUC: area under the receiver operating characteristic curve.

Figure 4 shows the feature importance ranking estimated from the ANN model (only the top 15 items are displayed). The frequency of sleep medication use ranked first, illustrating the most important factor associated with schizophrenia. Age took second place. Household income and type of employment ranked third as they showed similar values. Bedtime, BMI, number of cigarettes smoked per day, and educational background

followed, with a marked decrease in importance. Hours of sleep, perceived stress, positive reason for living (aka, *ikigai*, a Japanese term), restriction in functional capacity, type of occupation, number of teeth, and bowel frequency ranked 9th to 15th; however, their importance was not as high as those in the above places.

Figure 4. Feature importance (top 15) from artificial neural network model. AUC: area under the receiver operating characteristic curve.



Discussion

Principal Findings

In this study, we developed an ANN model for classifying schizophrenia cases with high internal validity, which achieved an excellent AUC of 0.86. This ANN model also achieved a specificity of 0.96, which implies that it has a good ability to designate an individual without schizophrenia as negative, and a sensitivity of 0.56, which represents the model's limitation in designating an individual with schizophrenia as positive. Our study demonstrated that the ANN has the potential for applying to estimate the prevalence of schizophrenia in large-scale epidemiological studies.

To our knowledge, this is the first study that uses a machine learning technique (ie, ANN) to classify schizophrenia cases from web-based survey data in the Japanese population. A novel machine learning approach has been reported for the detection of schizophrenia from social media content, which achieved an accuracy of 96% [16]. In addition, machine learning techniques might provide an opportunity to improve diagnostic certainty [35] and explain mental disorders in complex states [36]. On the other hand, ANN algorithms have proven their advantages in disease prediction using large-scale survey data. For predicting type 2 diabetes, models developed by ANN have achieved an AUC of 0.86, which is the highest compared to other algorithms such as random forest and support vector machine [21]. Another study reported that models developed by ANN for predicting hypertension achieved an AUC of 0.78; however, there was no significant advantage compared with the classic method [20].

In comparison of the algorithm performance, ANN performed better than LR. Several reasons might be considered: (1) ANN is more competent for approximate relations that do not follow the linearized assumption owing to its structure [37]. In generalized linear models (eg, LR), the relation between the response variable and the feature variables is applied to a linear

equation; therefore, linear models face difficulty in analyzing nonlinear combinations. As the number of feature variables in a linear model increases, multicollinearity [38] and overfitting may occur easily [39]. (2) The ANN is more effective for analyzing interactions among more than 2 variables. In linear models, interaction terms (usually products between 2 variables) are added. However, interaction features are often selected based on rules of thumb. Additionally, multicollinearity should be cautiously considered as the number of interaction terms increases [38]. While in an ANN, the variable relationships are assumed to be extremely high-dimensional and complex [40]. After inputting all the variables, each hidden neuron takes, as input, all nodes from the previous layer and creates a high-order interaction between these nodes [41]. Nevertheless, the complex structure of an ANN prevents the model from being easily visualized and understood.

In terms of the AUC value, ANN outperformed LR both in algorithm and model comparisons. However, in this study, the ANN model exhibited better specificity but lower sensitivity compared to LR as the cutoff threshold was set to the default 0.5. The model's performance metrics indicate a trade-off between sensitivity and specificity. This trade-off could impact the model's ability to correctly classify schizophrenia cases and noncases, which has clinical implications. Further experiments are necessary, involving the adjustment of the cutoff threshold [42] to determine a suitable balance point in practical scenarios.

This study suggested the possibility of classifying schizophrenia cases among general population. We expected to estimate schizophrenia cases among those who are not seeking psychiatric care, typically, whose psychiatric syndromes are unknown. Hence, no clinical assessment or diagnostic criteria were involved in the feature variables. As we ranked the feature importance, sleep problems were the most important factor associated with schizophrenia in terms of sleep medication use, bedtime, and sleep duration. The importance of factors such as age, socioeconomic background, BMI, physical activity, smoking, depression, and oral health followed. These factors

have also been reported to be strongly associated with schizophrenia [8,43-50]. The advantage of our study was that we ranked those previously reported factors according to their “priority” for classifying schizophrenia, which may provide potential evidence for screening and early detection of schizophrenia when using massive data.

In this study, no variable selection technique was preselected because we hope to mine as much potentially useful information as possible when training the models. Previous studies have reported that fully input variables might lead to unstable estimates in linear models such as LR [51]. Some methodologists have suggested that statistical significance-based variable selection techniques are mechanical and, as such, have limitations [51,52]. We conducted an additional experiment for the sake of “fairness”; the LR model was trained using only the top 15 most important features reported by the ANN model. The results obtained from this partial-featured LR model did not improve compared with the all-feature LR model (sensitivity: 0.69, specificity: 0.87, accuracy: 0.91, and AUC: 0.78). The machine learning approach might be used for feature selection to compensate for the limitations of classical epidemiology studies.

Limitations

This study has several limitations. First, participants with mental disorders other than schizophrenia were not included. Hence, our model is not suitable for distinguishing schizophrenia from other types of mental disorders. Second, although the data in this study were obtained from a large Japanese internet research pooled panel, we should cautiously explain the representativeness for the entire Japanese population. This could introduce sampling bias, potentially excluding individuals who are less likely to participate in web-based surveys, such as those with limited internet access or severe mental health conditions. In addition, the self-reported diagnoses of schizophrenia might not be as reliable as clinically confirmed diagnoses, as individuals might misinterpret symptoms or misunderstand their condition. There might also be issues related to stigma or disclosure bias, where individuals might be hesitant to disclose mental health diagnoses. Third, the important variables

determined by our model cannot be arbitrarily used as a criterion for identifying schizophrenia. For example, sleep medication use is often observed in patients with depression, although our model determined that it was most associated with schizophrenia. In future studies, we plan to include samples with various types of mental disorders and construct a model that can classify multiple mental disorders. Additionally, Clinical assessments and diagnostic criteria used by health care professionals were not included in the analysis. In future studies, we can introduce essential information to enhance the model construction. Fourth, all variables (feature variables and response variables) were self-reported at the same time. The answers were possibly biased because the participants might have been reluctant to answer some sensitive questions truthfully. The existence of schizophrenia might be different from the actual situation because the history of schizophrenia was reported by the participants themselves. Patients who did not use the internet and those who had difficulty completing the web-based survey due to their illness were not included in this study. These issues may have affected the accuracy and generalizability of the model. Fifth, the findings of this study were derived from a cross-sectional design; therefore, it is difficult to explain any causal or temporal associations. Sixth, because of the “black-box” design of the ANN model, it is difficult to interpret how variables and variable interactions contribute to the classification of schizophrenia. Further research is necessary to focus on model visualization and interpretation. Finally, the ideal model should be dynamic (ie, can be updated to adopt the latest data structure) [53]; hence, we need to input more large-scale data to improve the current model and to assess the model performance by external validation.

Conclusions

In this study, an ANN model was constructed to classify schizophrenia cases using web-based survey data. The model achieved a high internal validity. ANN performed better compared to the classical statistic method. These findings are expected to provide evidence for estimating the prevalence of schizophrenia in the Japanese population and informing future epidemiological studies.

Acknowledgments

This study was funded by a Health and Labour Sciences Research grant from the Ministry of Health, Labour and Welfare, Japan (JPMH21GC1018). The founder made no interference with the authors' study design. No generative artificial intelligence or similar tools were used when preparing the paper.

Data Availability

The data sets analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YH made the conceptualization, conducted the data curation, methodology, data analysis, and interpretation, and wrote the original paper. MM made the conceptualization, conducted the investigation and data curation, and revised the paper. YL, NI, and TT revised the paper. TK and ST made the conceptualization and revised the paper. AO made the conceptualization, supervised the investigation, conducted the methodology and data curation, wrote part of the paper, and revised the paper. All authors reviewed the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Data source and variable definition.

[\[DOCX File , 47 KB-Multimedia Appendix 1\]](#)

References

1. Schizophrenia. National Institute of Mental Health. URL: <https://www.nimh.nih.gov/health/statistics/schizophrenia> [accessed 2023-10-18]
2. Saha S, Chant D, Welham J, McGrath J. A systematic review of the prevalence of schizophrenia. *PLoS Med* 2005 May;2(5):e141 [FREE Full text] [doi: [10.1371/journal.pmed.0020141](https://doi.org/10.1371/journal.pmed.0020141)] [Medline: [15916472](https://pubmed.ncbi.nlm.nih.gov/15916472/)]
3. Moreno-Küstner B, Martín C, Pastor L. Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PLoS One* 2018;13(4):e0195687 [FREE Full text] [doi: [10.1371/journal.pone.0195687](https://doi.org/10.1371/journal.pone.0195687)] [Medline: [29649252](https://pubmed.ncbi.nlm.nih.gov/29649252/)]
4. Simeone JC, Ward AJ, Rotella P, Collins J, Windisch R. An evaluation of variation in published estimates of schizophrenia prevalence from 1990-2013: a systematic literature review. *BMC Psychiatry* 2015;15:193 [FREE Full text] [doi: [10.1186/s12888-015-0578-7](https://doi.org/10.1186/s12888-015-0578-7)] [Medline: [26263900](https://pubmed.ncbi.nlm.nih.gov/26263900/)]
5. Okui T. An age-period-cohort analysis for prevalence of common psychiatric disorders in Japan, 1999-2017. *Soc Psychiatry Psychiatr Epidemiol* 2021;56(4):639-648 [FREE Full text] [doi: [10.1007/s00127-020-01988-5](https://doi.org/10.1007/s00127-020-01988-5)] [Medline: [33221927](https://pubmed.ncbi.nlm.nih.gov/33221927/)]
6. Fervaha G, Agid O, Takeuchi H, Foussias G, Remington G. Life satisfaction and happiness among young adults with schizophrenia. *Psychiatry Res* 2016;242:174-179 [FREE Full text] [doi: [10.1016/j.psychres.2016.05.046](https://doi.org/10.1016/j.psychres.2016.05.046)] [Medline: [27288735](https://pubmed.ncbi.nlm.nih.gov/27288735/)]
7. Palmer BW, Martin AS, Depp CA, Glorioso DK, Jeste DV. Wellness within illness: happiness in schizophrenia. *Schizophr Res* 2014;159(1):151-156 [FREE Full text] [doi: [10.1016/j.schres.2014.07.027](https://doi.org/10.1016/j.schres.2014.07.027)] [Medline: [25153363](https://pubmed.ncbi.nlm.nih.gov/25153363/)]
8. Sun XN, Zhou JB, Li N. Poor oral health in patients with schizophrenia: a meta-analysis of case-control studies. *Psychiatr Q* 2021;92(1):135-145 [doi: [10.1007/s11126-020-09752-3](https://doi.org/10.1007/s11126-020-09752-3)] [Medline: [32483766](https://pubmed.ncbi.nlm.nih.gov/32483766/)]
9. Zareifopoulos N, Bellou A, Spiropoulou A, Spiropoulos K. Prevalence of comorbid chronic obstructive pulmonary disease in individuals suffering from schizophrenia and bipolar disorder: a systematic review. *COPD* 2018;15(6):612-620 [FREE Full text] [doi: [10.1080/15412555.2019.1572730](https://doi.org/10.1080/15412555.2019.1572730)] [Medline: [30714418](https://pubmed.ncbi.nlm.nih.gov/30714418/)]
10. Mamakou V, Thanopoulou A, Gonidakis F, Tentolouris N, Kontaxakis V. Schizophrenia and type 2 diabetes mellitus. *Psychiatriki* 2018;29(1):64-73 [FREE Full text] [doi: [10.22365/jpsych.2018.291.64](https://doi.org/10.22365/jpsych.2018.291.64)] [Medline: [29754122](https://pubmed.ncbi.nlm.nih.gov/29754122/)]
11. Crespo-Facorro B, Such P, Nylander A, Madera J, Resemann HK, Worthington E, et al. The burden of disease in early schizophrenia—a systematic literature review. *Curr Med Res Opin* 2021;37(1):109-121 [FREE Full text] [doi: [10.1080/03007995.2020.1841618](https://doi.org/10.1080/03007995.2020.1841618)] [Medline: [33095689](https://pubmed.ncbi.nlm.nih.gov/33095689/)]
12. Waite F, Sheaves B, Isham L, Reeve S, Freeman D. Sleep and schizophrenia: from epiphenomenon to treatable causal target. *Schizophr Res* 2020;221:44-56 [FREE Full text] [doi: [10.1016/j.schres.2019.11.014](https://doi.org/10.1016/j.schres.2019.11.014)] [Medline: [31831262](https://pubmed.ncbi.nlm.nih.gov/31831262/)]
13. Harvey PD, Isner EC. Cognition, social cognition, and functional capacity in early-onset schizophrenia. *Child Adolesc Psychiatr Clin N Am* 2020;29(1):171-182 [doi: [10.1016/j.chc.2019.08.008](https://doi.org/10.1016/j.chc.2019.08.008)] [Medline: [31708046](https://pubmed.ncbi.nlm.nih.gov/31708046/)]
14. He Y, Tanaka A, Kishi T, Li Y, Matsunaga M, Tanihara S, et al. Recent findings on subjective well-being and physical, psychiatric, and social comorbidities in individuals with schizophrenia: a literature review. *Neuropsychopharmacol Rep* 2022;42(4):430-436 [FREE Full text] [doi: [10.1002/npr2.12286](https://doi.org/10.1002/npr2.12286)] [Medline: [35916310](https://pubmed.ncbi.nlm.nih.gov/35916310/)]
15. Birnbaum ML, Kulkarni PP, Van Meter A, Chen V, Rizvi AF, Arenare E, et al. Utilizing machine learning on internet search activity to support the diagnostic process and relapse detection in young individuals with early psychosis: feasibility study. *JMIR Ment Health* 2020;7(9):e19348 [FREE Full text] [doi: [10.2196/19348](https://doi.org/10.2196/19348)] [Medline: [32870161](https://pubmed.ncbi.nlm.nih.gov/32870161/)]
16. Bae YJ, Shim M, Lee WH. Schizophrenia detection using machine learning approach from social media content. *Sensors (Basel)* 2021;21(17):5924 [FREE Full text] [doi: [10.3390/s21175924](https://doi.org/10.3390/s21175924)] [Medline: [34502815](https://pubmed.ncbi.nlm.nih.gov/34502815/)]
17. Lejeune A, Robaglia BM, Walter M, Berrouguet S, Lemey C. Use of social media data to diagnose and monitor psychotic disorders: systematic review. *J Med Internet Res* 2022;24(9):e36986 [FREE Full text] [doi: [10.2196/36986](https://doi.org/10.2196/36986)] [Medline: [36066938](https://pubmed.ncbi.nlm.nih.gov/36066938/)]
18. Thorstad R, Wolff P. Predicting future mental illness from social media: a big-data approach. *Behav Res Methods* 2019;51(4):1586-1600 [FREE Full text] [doi: [10.3758/s13428-019-01235-z](https://doi.org/10.3758/s13428-019-01235-z)] [Medline: [31037606](https://pubmed.ncbi.nlm.nih.gov/31037606/)]
19. Gkotsis G, Oellrich A, Velupillai S, Liakata M, Hubbard TJP, Dobson RJB, et al. Characterisation of mental health conditions in social media using informed deep learning. *Sci Rep* 2017;7:45141 [FREE Full text] [doi: [10.1038/srep45141](https://doi.org/10.1038/srep45141)] [Medline: [28327593](https://pubmed.ncbi.nlm.nih.gov/28327593/)]
20. López-Martínez F, Núñez-Valdez ER, Crespo RG, García-Díaz V. An artificial neural network approach for predicting hypertension using NHANES data. *Sci Rep* 2020;10(1):10620 [FREE Full text] [doi: [10.1038/s41598-020-67640-z](https://doi.org/10.1038/s41598-020-67640-z)] [Medline: [32606434](https://pubmed.ncbi.nlm.nih.gov/32606434/)]

21. Zhang L, Wang Y, Niu M, Wang C, Wang Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. *Sci Rep* 2020;10(1):4406 [FREE Full text] [doi: [10.1038/s41598-020-61123-x](https://doi.org/10.1038/s41598-020-61123-x)] [Medline: [32157171](https://pubmed.ncbi.nlm.nih.gov/32157171/)]
22. Development of a prevalence estimation method for out-of-hospital schizophrenia and schizophrenia-related disorders in the general population using large-scale epidemiological study data and medical reimbursement statement data [in Japanese]. Ministry of Health, Labour and Welfare Grants System. URL: <https://mhlw-grants.niph.go.jp/project/158063> [accessed 2023-10-18]
23. Rakuten Insight, Inc. URL: <https://insight.rakuten.co.jp/en/aboutus.html> [accessed 2023-10-18]
24. Rakuten disease panel. URL: <https://rd.insight.rakuten.net/1/310391/2020-09-07/w7bk3r> [accessed 2023-10-18]
25. The Japan COVID-19 and Society Internet Survey. URL: <https://takahiro-tabuchi.net/jacsis/> [accessed 2023-03-08]
26. Matsunaga M, Li Y, He Y, Kishi T, Tanihara S, Iwata N, et al. Physical, psychiatric, and social comorbidities of individuals with schizophrenia living in the community in Japan. *Int J Environ Res Public Health* 2023;20(5):4336 [FREE Full text] [doi: [10.3390/ijerph20054336](https://doi.org/10.3390/ijerph20054336)] [Medline: [36901345](https://pubmed.ncbi.nlm.nih.gov/36901345/)]
27. Hardesty L. Explained: neural networks. Massachusetts Institute of Technology. URL: <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414> [accessed 2023-10-18]
28. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. 2015 Presented at: 2015 IEEE International Conference on Computer Vision (ICCV); December 7-13, 2015; Santiago, Chile [doi: [10.1109/iccv.2015.123](https://doi.org/10.1109/iccv.2015.123)]
29. Nwankpa C, Ijomah W, Gachagan A, Marshall S. Activation functions: comparison of trends in practice and research for deep learning. *arXiv* 2018:1-20 [FREE Full text] [doi: [10.48550/arXiv.1811.03378](https://doi.org/10.48550/arXiv.1811.03378)]
30. He Y, Chiang C, Hirakawa Y, Yatsuya H. Comparison of artificial neural network and logistic regression for predicting common metabolic outcomes. 2021 Presented at: The 31rd Annual Scientific Meeting of the Japanese Epidemiological Association; January 27-29, 2021; Japan
31. What is logistic regression? IBM. URL: <https://www.ibm.com/topics/logistic-regression> [accessed 2023-10-18]
32. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321-357 [doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953)]
33. Hosmer DW, Lemeshow S, Sturdivant RX. Assessing the fit of the model. In: *Applied Logistic Regression*. 3rd Edition. Hoboken, NJ: John Wiley & Sons, Inc; 2013:153-225
34. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837-845 [Medline: [3203132](https://pubmed.ncbi.nlm.nih.gov/3203132/)]
35. Starke G, De Clercq E, Borgwardt S, Elger BS. Computing schizophrenia: ethical challenges for machine learning in psychiatry. *Psychol Med* 2021;51(15):2515-2521 [FREE Full text] [doi: [10.1017/S0033291720001683](https://doi.org/10.1017/S0033291720001683)] [Medline: [32536358](https://pubmed.ncbi.nlm.nih.gov/32536358/)]
36. Tai AMY, Albuquerque A, Carmona NE, Subramaniepillai M, Cha DS, Sheko M, et al. Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artif Intell Med* 2019 Aug;99:101704 [doi: [10.1016/j.artmed.2019.101704](https://doi.org/10.1016/j.artmed.2019.101704)] [Medline: [31606109](https://pubmed.ncbi.nlm.nih.gov/31606109/)]
37. Cai S, Bileschi S, Nielsen E, Chollet F. Chapter 3. Adding nonlinearity: beyond weighted sums. In: *Deep Learning with JavaScript*. New York: Manning; 2020:79-116
38. Jaccard J. *Interaction Effects in Logistic Regression*. Thousand Oaks: SAGE Publications, Inc; 2001.
39. Sainani KL. Understanding linear regression. *PM R* 2013;5(12):1063-1068 [FREE Full text] [doi: [10.1016/j.pmrj.2013.10.002](https://doi.org/10.1016/j.pmrj.2013.10.002)] [Medline: [24140739](https://pubmed.ncbi.nlm.nih.gov/24140739/)]
40. Olden JD, Jackson DA. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecol Model* 2002;154(1-2):135-150 [FREE Full text] [doi: [10.1016/s0304-3800\(02\)00064-9](https://doi.org/10.1016/s0304-3800(02)00064-9)]
41. Tsang M, Liu H, Purushotham S, Murali P, Liu Y. Neural Interaction Transparency (NIT): Disentangling Learned Interactions for Improved Interpretability. 2018. URL: <https://proceedings.neurips.cc/paper/2018/file/74378afe5e8b20910cf1f939e57f0480-Paper.pdf> [accessed 2023-10-18]
42. Kuhn M, Johnson K. Remedies for severe class imbalance. In: *Applied Predictive Modeling*. New York: Springer; 2013:419-443
43. Meyer N, Faulkner SM, McCutcheon RA, Pillinger T, Dijk DJ, MacCabe JH. Sleep and circadian rhythm disturbance in remitted schizophrenia and bipolar disorder: a systematic review and meta-analysis. *Schizophr Bull* 2020;46(5):1126-1143 [FREE Full text] [doi: [10.1093/schbul/sbaa024](https://doi.org/10.1093/schbul/sbaa024)] [Medline: [32154882](https://pubmed.ncbi.nlm.nih.gov/32154882/)]
44. Kiliçaslan EE, Erol A, Zengin B, Aydin P, Mete L. Association between age at onset of schizophrenia and age at menarche. *Noro Psikiyat Ars* 2014;51(3):211-215 [FREE Full text] [doi: [10.4274/npa.y6675](https://doi.org/10.4274/npa.y6675)] [Medline: [28360628](https://pubmed.ncbi.nlm.nih.gov/28360628/)]
45. Hakulinen C, McGrath JJ, Timmerman A, Skipper N, Mortensen PB, Pedersen CB, et al. The association between early-onset schizophrenia with employment, income, education, and cohabitation status: nationwide study with 35 years of follow-up. *Soc Psychiatry Psychiatr Epidemiol* 2019;54(11):1343-1351 [FREE Full text] [doi: [10.1007/s00127-019-01756-0](https://doi.org/10.1007/s00127-019-01756-0)] [Medline: [31456027](https://pubmed.ncbi.nlm.nih.gov/31456027/)]
46. Mitchell AJ, Vancampfort D, Sweers K, van Winkel R, Yu W, De Hert M. Prevalence of metabolic syndrome and metabolic abnormalities in schizophrenia and related disorders—a systematic review and meta-analysis. *Schizophr Bull* 2013;39(2):306-318 [FREE Full text] [doi: [10.1093/schbul/sbr148](https://doi.org/10.1093/schbul/sbr148)] [Medline: [22207632](https://pubmed.ncbi.nlm.nih.gov/22207632/)]

47. Shah P, Iwata Y, Caravaggio F, Plitman E, Brown EE, Kim J, et al. Alterations in body mass index and waist-to-hip ratio in never and minimally treated patients with psychosis: a systematic review and meta-analysis. *Schizophr Res* 2019;208:420-429 [FREE Full text] [doi: [10.1016/j.schres.2019.01.005](https://doi.org/10.1016/j.schres.2019.01.005)] [Medline: [30685395](https://pubmed.ncbi.nlm.nih.gov/30685395/)]
48. Papiol S, Schmitt A, Maurus I, Rossner MJ, Schulze TG, Falkai P. Association between physical activity and schizophrenia: results of a 2-sample mendelian randomization analysis. *JAMA Psychiatry* 2021;78(4):441-444 [FREE Full text] [doi: [10.1001/jamapsychiatry.2020.3946](https://doi.org/10.1001/jamapsychiatry.2020.3946)] [Medline: [33295946](https://pubmed.ncbi.nlm.nih.gov/33295946/)]
49. Hunter A, Murray R, Asher L, Leonardi-Bee J. The effects of tobacco smoking, and prenatal tobacco smoke exposure, on risk of schizophrenia: a systematic review and meta-analysis. *Nicotine Tob Res* 2020;22(1):3-10 [FREE Full text] [doi: [10.1093/ntr/nty160](https://doi.org/10.1093/ntr/nty160)] [Medline: [30102383](https://pubmed.ncbi.nlm.nih.gov/30102383/)]
50. Etchecopar-Etchart D, Korchia T, Loundou A, Llorca PM, Auquier P, Lançon C, et al. Comorbid major depressive disorder in schizophrenia: a systematic review and meta-analysis. *Schizophr Bull* 2021;47(2):298-308 [FREE Full text] [doi: [10.1093/schbul/sbaa153](https://doi.org/10.1093/schbul/sbaa153)] [Medline: [33252130](https://pubmed.ncbi.nlm.nih.gov/33252130/)]
51. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. *Source Code Biol Med* 2008;3:17 [FREE Full text] [doi: [10.1186/1751-0473-3-17](https://doi.org/10.1186/1751-0473-3-17)] [Medline: [19087314](https://pubmed.ncbi.nlm.nih.gov/19087314/)]
52. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;3:1157-1182 [FREE Full text]
53. Ellner SP, Guckenheimer J. What are dynamic models? In: *Dynamic Models in Biology*. New Jersey: Princeton University Press; 2006:1-30

Abbreviations

ANN: artificial neural network

AUC: area under the receiver operating characteristic curve

LR: logistic regression

Edited by A Mavragani; submitted 26.06.23; peer-reviewed by Y Bafna, E Vashishtha; comments to author 08.09.23; revised version received 18.09.23; accepted 08.10.23; published 15.11.23

Please cite as:

He Y, Matsunaga M, Li Y, Kishi T, Tanihara S, Iwata N, Tabuchi T, Ota A

Classifying Schizophrenia Cases by Artificial Neural Network Using Japanese Web-Based Survey Data: Case-Control Study

JMIR Form Res 2023;7:e50193

URL: <https://formative.jmir.org/2023/1/e50193>

doi: [10.2196/50193](https://doi.org/10.2196/50193)

PMID: [37966882](https://pubmed.ncbi.nlm.nih.gov/37966882/)

©Yupeng He, Masaaki Matsunaga, Yuanying Li, Taro Kishi, Shinichi Tanihara, Nakao Iwata, Takahiro Tabuchi, Atsuhiko Ota. Originally published in *JMIR Formative Research* (<https://formative.jmir.org>), 15.11.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Formative Research*, is properly cited. The complete bibliographic information, a link to the original publication on <https://formative.jmir.org>, as well as this copyright and license information must be included.