RESEARCH ARTICLE OPEN ACCESS

Machine Learning Approach for the Prediction of Age-Specific Probability of SCA3 and DRPLA by Survival Curve Analysis

Yuya Hatano, MD, PhD, Tomohiko Ishihara, MD, PhD, Sachiko Hirokawa, MS, and Osamu Onodera, MD, PhD

Neurol Genet 2023;9:e200075. doi:10.1212/NXG.0000000000200075

Correspondence
Dr. Ishihara
ishihara@bri.niigata-u.ac.jp

Abstract

Background and Objectives

As the number of repeats in the expansion increases, polyglutamine diseases tend to show at a younger age. From this relationship, attempts have been made to predict age at onset by parametric survival analysis. However, a method for a more accurate prediction has been desirable. In this study, we examined 2 methods for survival analysis using machine learning and 6 conventional methods for parametric survival analysis of spinocerebellar ataxia (SCA)3 and dentatorubral-pallidoluysian atrophy (DRPLA).

Methods

We compared the performance of 2 machine learning methods of survival analysis (random survival forest [RSF] and DeepSurv) and 6 methods of parametric survival analysis (Weibull, exponential, Gaussian, logistic, loglogistic, and log Gaussian). Training and evaluation were performed using the leave-one-out cross-validation method, and evaluation criteria included root mean squared error (RMSE), mean absolute error (MAE), and the integrated Brier score. The latter was used as the primary end point, and the survival analysis model yielding the best result was used to predict the asymptomatic probability.

Results

Among the models examined, the RSF and DeepSurv machine learning methods had a higher prediction accuracy than the parametric methods of survival analysis. For both SCA3 and DRPLA, RSF had a higher accuracy than DeepSurv for the assessment of RMSE (SCA3: 7.37, DRPLA: 10.78), MAE (SCA3: 5.52, DRPLA: 8.17), and the integrated Brier score (SCA3: 0.05, DRPLA: 0.077). Using RSF, we determined the age-specific probability distribution of age at onset based on CAG repeat size and current age.

Discussion

In this study, we have demonstrated the superiority of machine learning methods for predicting age at onset of SCA3 and DRPLA using survival analysis. Such accurate prediction of onset will be useful for genetic counseling of carriers and for devising methods to verify the effects of interventions for unaffected individuals.

From the Department of Neurology, Brain Research Institute, Niigata University, Niigata-shi, Japan.

Go to Neurology.org/NG for full disclosures. Funding information is provided at the end of the article.

The Article Processing Charge was funded by the authors.

Glossary

DRPLA = dentatorubral-pallidoluysian atrophy; **MAE** = mean absolute error; **RMSE** = root mean squared error; **RSF** = random survival forest; **SCA** = spinocerebellar ataxia.

Neurodegenerative diseases caused by the expansion of CAG triplet repeats encoding polyglutamine chains in specific genes are known as polyglutamine diseases.[1] The penetrance of the pathologic allele is estimated to be 100%, and the CAG repeat size shows a strong inverse correlation with age at onset.[2] Based on this association, the probability of developing the disease at a certain age can be estimated based on the number of CAG repeats in the pathologic allele. This prediction method is helpful for genetic counseling of unaffected carriers in the context of their life plan. Furthermore, an accurate prediction of onset age in nonaffected individuals is highly significant for the design of prophylactic clinical trials.[3] Indeed, in Huntington disease and spinocerebellar ataxia (SCA)3 and SCA6, methods for predicting age at onset using parametric survival analysis have been demonstrated.[2-5] However, development of methods with more predictive accuracy than parametric survival analyses has been desirable.

Recently, several methods with a high predictive accuracy have been developed using machine learning. Machine learning is a branch of artificial intelligence that extends predictive modeling through traditional statistical analysis. Complex, nonlinear interacting variables can be acquired by machine learning to minimize the error gap between predictions and observations. Several machine learning methods have been used for the diagnosis and prognostication of cancer and neurologic diseases.[6,7] In polyglutamine diseases, the machine learning method XGBoost has also been used for a more accurate prediction of the age at onset of SCA3.[8] However, machine learning was not used in previous studies designed to predict the age at onset of polyglutamine diseases using survival analysis.[2-5] Random survival forest (RSF)[9] and DeepSurv[10] are 2 representative methods of survival analysis that were developed using machine learning. These methods have shown more accurate predictive results than conventional semiparametric survival analyses for patients with oral cancer, those who are critically ill and hospitalized, and those with acute myocardial infarction.[10,11]

In this study, we performed survival analysis of SCA3 and dentatorubral-pallidoluysian atrophy (DRPLA), which are relatively common polyglutamine diseases in Japan, using RSF and DeepSurv, as well as 6 conventional methods of parametric survival analysis, and verified their accuracy. In addition, we used RSF survival analysis to predict the age at disease onset in each age group.

Methods

Patients

Among cases diagnosed by genetic testing at the Department of Neurology, Clinical Neuroscience Branch/Department of Molecular Neuroscience, Resource Branch for Brain Disease Research, Brain Research Institute, Niigata University, between 1992 and 2020, 292 cases of SCA3 and 203 cases of DRPLA with an identifiable age at onset were selected. We defined cases with SCA3 as those with at least 55 CAG repeats on at least 1 of 2 alleles of *ATXN3* and cases with DRPLA as those with at least 49 CAG repeats on at least 1 of 2 alleles of *ATN1*.[12,13]

Genetic Analysis

Genomic DNA was extracted from venous blood using the PAXgene Blood DNA kit (QIAGEN, Hilden, Germany). PCR was performed on the CAG repeat region of the *ATXN3* and *ATN1* genes as previously reported, and fragment size was analyzed by fluorescence capillary electrophoresis.[14,15]

Preanalysis With Boruta

Boruta[16] was used for feature selection, using sex, the number of repeats of expanded alleles, and the number of repeats of nonexpanded alleles as explanatory variables and age at onset as an objective variable. For Boruta analysis, 5 cases with SCA3 for whom the sex or number of repeats was unknown were excluded from the analysis. Features were classified into 3 groups: important, tentative, and unimportant. The unimportant features were excluded from this analysis. Statistical software R version 4.1.0 was used for the analysis, with the Boruta function from the Boruta package, and the parameters were left at their default settings.

Construction of the Prediction Model

Two machine learning methods (RSF and DeepSurv) and 6 methods of parametric survival analysis (Weibull, exponential, Gaussian, logistic, loglogistic, and log Gaussian) that had been previously evaluated[3] were used to estimate the age at onset from CAG repeat length. All models were built using the statistical software R version 4.1.0. We applied the survreg function from the survival package to fit the parametric survival models and the predict function from the survival package to predict the asymptomatic probability. The rfsrc function from the randomForestSRC package was applied to train the RSF model, and the predict function from the randomForestSRC package was applied to predict the asymptomatic probability. The asymptomatic probability obtained with the predict function is a discrete variable. On the contrary, the integrated Brier score that we used to evaluate our model assumes a continuous variable. To estimate the integrated Brier score as accurately as possible, we estimated the asymptomatic probability in units as small as 1/1,000 of a year based on age and the asymptomatic probability obtained by the predict function. We defined the 2 adjacent ages predicted by the predict function as ageA and ageB and defined the asymptomatic probability at ageA and ageB as proA and proB,

Table 1 Fitting Results of the 6 Parametric Survival Models and 2 Machine Learning Models in Patients With SCA3

	RMSE	MAE	Integrated Brier score
Weibull	8.5	6.17	0.163
Exponential	15.3	13.38	0.106
Gaussian	7.51	5.68	0.162
Logistic	7.52	5.66	0.169
Loglogistic	9.32	6.36	0.167
Log Gaussian	9.16	6.33	0.157
Random survival forest	7.37	5.52	0.05
DeepSurv	8.34	6.28	0.058

Abbreviations: DRPLA = dentatorubral-pallidoluysian atrophy; MAE = mean absolute error; RMSE = root mean squared error; SCA = spinocerebellar ataxia. A machine learning approach for the prediction of age-specific probability of SCA3 and DRPLA by survival curve analysis.

Table 2 Fitting Results of the 6 Parametric Survival Models and 2 Machine Learning Models in Patients With DRPLA

	RMSE	MAE	Integrated Brier score
Weibull	13.62	9.65	0.158
Exponential	15.74	12.01	0.12
Gaussian	11.4	9.01	0.168
Logistic	11.41	9.02	0.171
Loglogistic	14.8	9.99	0.156
Log Gaussian	16.02	10.51	0.117
Random survival forest	10.78	8.17	0.077
DeepSurv	11.57	8.76	0.086

Abbreviations: DRPLA = dentatorubral-pallidoluysian atrophy; MAE = mean absolute error; RMSE = root mean squared error; SCA = spinocerebellar ataxia. A machine learning approach for the prediction of age-specific probability of SCA3 and DRPLA by survival curve analysis.

respectively. Furthermore, we defined the asymptomatic probability at a certain age, ageC, between ageB and ageA as proC. We estimated proC = ageC × (proB − proA)/(ageB − ageA) + proA − ageA × (proB − proA)/(ageB − ageA) to predict the asymptomatic probability every 1/1,000 of an age. The DeepSurv function from the survivalmodels package was used to train the DeepSurv model, and the predict function from the survivalmodels package was used to predict the asymptomatic probability. As in RSF, the asymptomatic probability was estimated for every 1/1,000 of an age. The parameters of RSF and DeepSurv are listed in eTable 1, links.lww.com/NXG/A607.

Model Evaluation

For the 6 parametric survival analysis methods, RSF, and DeepSurv, training and evaluation were performed using the leave-one-out cross-validation method. Only 1 case was selected from the samples to serve as the test case, and the remaining cases were used as training cases. The validation was then repeated so that all cases became test cases one at a time. The median predicted age at onset was defined as the age at which the asymptomatic probability was 0.5 for parametric survival analysis. In RSF and DeepSurv, the median predicted age at onset was defined as the average of the highest age at which the asymptomatic probability was greater than 0.5 and the lowest age at which the asymptomatic probability was less than 0.5, among the ages at which the asymptomatic probability was estimated for every 1/1,000 of an age. The closeness of this median predicted value of age at onset to the observed value was evaluated using root mean squared error (RMSE) and mean absolute error (MAE).

The Brier score measures the mean squared difference between forecast probability and actual value (1 if it occurs, 0 if it does not), and the original integrated Brier score is the Brier score integrated over time and divided by the maximum time. However, in this analysis, the Brier score cannot be integrated over time. Therefore, in RSF and DeepSurv, we calculated the mean squared difference, N, between the predicted probability and the observed value (1 if the disease has not yet developed and 0 if it has developed), which we estimated every 1/1,000 of a year, and used the average of N in each test case as the variant of the integrated Brier score. For parametric survival analysis, we calculated the observed values at the times when the predicted probability was 0.01, 0.02, 0.03 … 0.99, calculated the mean squared difference, N, between the predicted probability and the observed value, and used the average of N in each test case as the variant of the integrated Brier score.

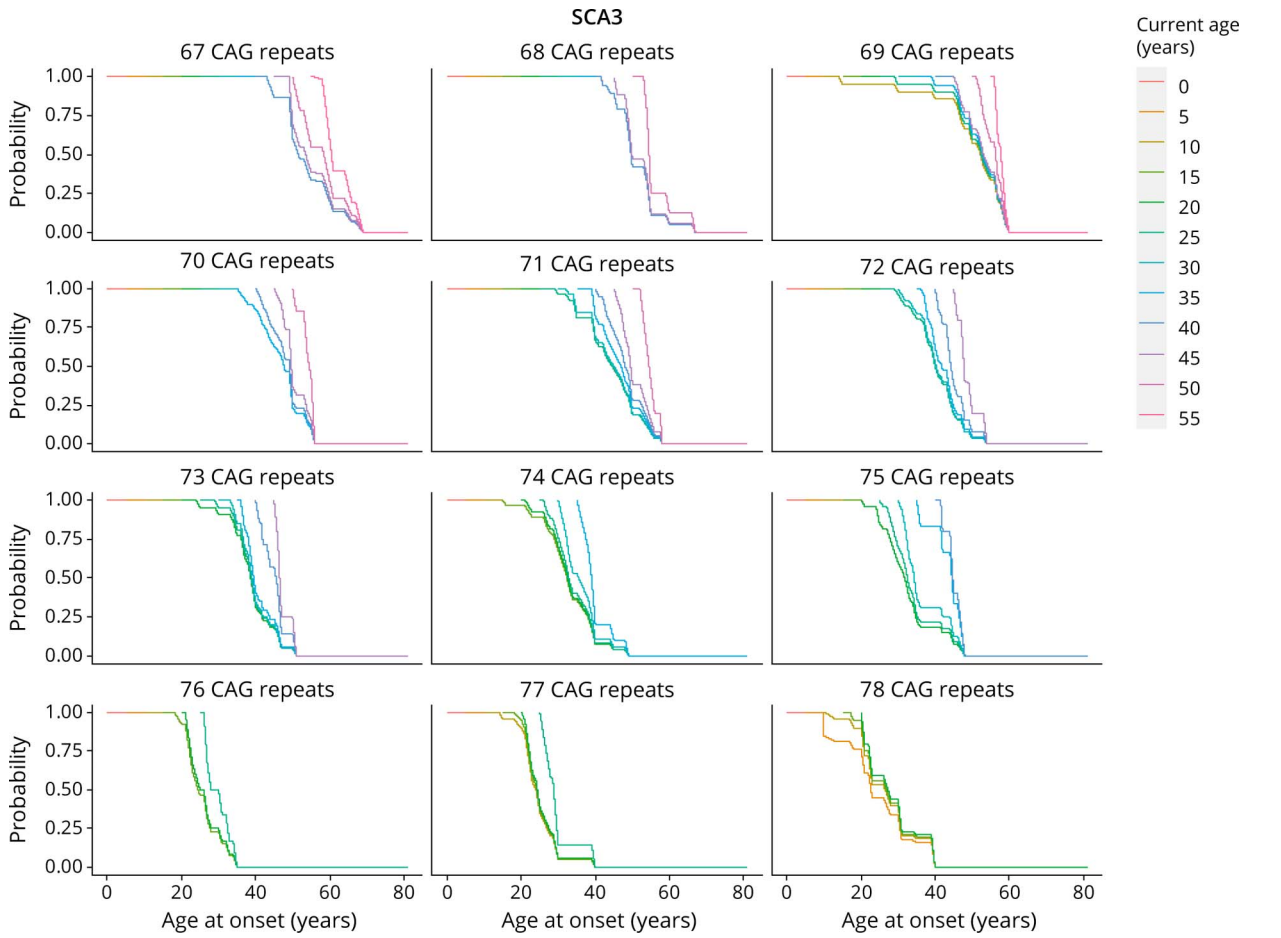
Model Prediction

The integrated Brier score was used as the primary end point, and the survival analysis model with the best result was used to predict the asymptomatic probability. All samples were trained as training cases, and the asymptomatic probability was predicted for each repeat up to 67–78 repeats for SCA3 and 60–70 repeats for DRPLA. Because, in machine learning, the value of asymptomatic probability can change from trial to trial, the asymptomatic probability was predicted 100 times, and the average of the asymptomatic predicted probabilities was calculated. The asymptomatic probability at age Y if asymptomatic at age X was calculated as (asymptomatic probability at age Y if asymptomatic at age 0)/(asymptomatic probability at age X if asymptomatic at age 0) from the definition of conditioning probability.

Model Sharing

We have released an application for Windows 64-bit that can illustrate the asymptomatic probability at a particular age by entering the current age and number of repeats, based on the results

Figure 1 Analysis in Patients With SCA3



The probability unaffected at a given age if currently unaffected is shown in the range 67–78 CAG repeats. Current age is indicated by color coding, with a given age on the x-axis and asymptomatic probability on the y-axis. SCA = spinocerebellar ataxia.

of the RSF analysis (github.com/yuya-hatano/SCA-onset" github.com/yuya-hatano/SCA-onset). The application was developed in Python 3.11.0.

Standard Protocol Approvals, Registrations, and Patient Consents

This study was approved by the Ethics Committee on Genetic Analysis of Niigata University (approval number: G2021-0010). All participants provided written informed consent.

Data Availability

Data set used for analysis in this study is not publicly available. If further information is required, please contact the corresponding author with a reasonable request.

Results

Data Set Details

In the data set used, the number of repeats of the expanded allele in patients with SCA3 was 71.5 ± 4.5, (mean ± SD, range 56–84) and the age at onset was 41.8 ± 13.5 (mean ± SD, range 10–81) years. There were 142 male cases, 148 female cases and 2 cases of unknown sex. Patients with DRPLA had 65.1 ± 4.2 (mean ± SD, range 55–79) repeats of the expanded allele, and the age at onset was 32.6 ± 20.1 (mean ± SD, range 0–76) years. There were 86 male and 117 female patients.

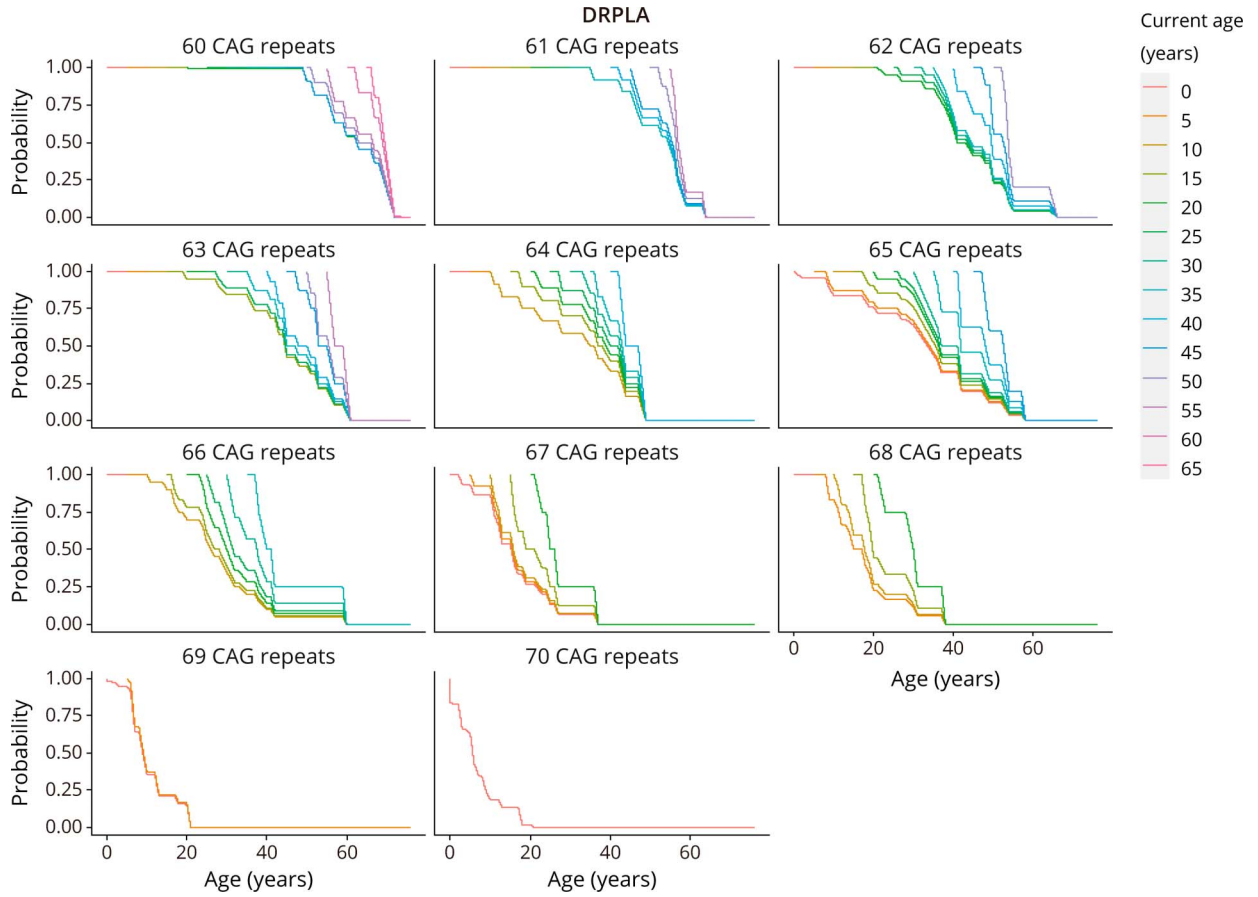
Preanalysis

The feature selection method Boruta[16] was used to select features for this analysis as a preanalysis. Among the features (sex, number of repeats of expanded alleles, and number of repeats of nonexpanded alleles) in cases with SCA3 and DRPLA, only the number of repeats of expanded alleles was considered important, while the other 2 features were considered unimportant. Therefore, for both SCA3 and DRPLA, we performed 6 parametric survival analyses and 2 machine learning analyses using only the number of repeats of the expanded allele as a feature and age at onset as an objective variable.

Model Evaluation

Six parametric survival analyses (Weibull, exponential, Gaussian, logistic, loglogistic, and log Gaussian) and 2 machine learning methods (RSF and DeepSurv) were used

Figure 2 Analysis in Patients With DRPLA



The probability unaffected at a given age if currently unaffected is shown in the range 60–70 CAG repeats. Current age is indicated by color coding, with a given age on the x-axis and asymptomatic probability on the y-axis. DRPLA = dentatorubral-pallidoluysian atrophy.

to predict age at onset. The accuracy of RMSE, MAE, and the integrated Brier score for each analysis is listed in Table 1 (SCA3) and Table 2 (DRPLA). For both diseases, the machine learning method RSF had the highest accuracy for the assessment of RMSE, MAE, and the integrated Brier score.

Prediction of Age at Onset

The probability that an unaffected person with a pathologic allele at a certain age would remain unaffected in subsequent years was predicted using the RSF for each of SCA3 and DRPLA (Figures 1 and 2). The median predicted age at onset is summarized in Table 3 (SCA3) and Table 4 (DRPLA). As expected, the number of CAG repeats of the pathologic allele and age at onset were inversely correlated. Exceptionally, 69 repeats of SCA3 resulted in an older age at onset than 67 and 68 repeats, and even 63 repeats of DRPLA resulted in an older age at onset than 62 repeats.

Discussion

In this study, we demonstrated the superiority of machine learning methods for predicting age at onset for SCA3 and DRPLA using survival analysis. We validated the accuracy of prediction of age at onset in SCA3 and DRPLA using 8 methods of survival analysis, including 2 machine learning methods (RSF and DeepSurv), and parametric survival analysis. The results showed that RSF and DeepSurv had a higher prediction accuracy than parametric survival analyses in the leave-one-out cross-validation method, indicating the superiority of machine learning methods for predicting the age at onset of SCA3 and DRPLA (Tables 1 and 2). These results may be attributed to the fact that parametric survival analysis requires fitting an appropriate probability distribution to the survival function, whereas RSF and DeepSurv do not require such an assumption. Because RSF performed slightly better than DeepSurv in this study (Tables 1 and 2), we used RSF to predict age at onset.

Predicting the probability of developing a genetic disease at each subsequent age is useful for genetic counseling of carriers and for devising methods for verifying the effect of the intervention on unaffected persons. The treatment effect can be measured by comparing the actual with the assumed onset age from the chronological age and number of CAG repeats. In addition, through prospective observation, genetic and

Table 3 Expected Age at Onset Using Random Survival Forest From Different Current Ages According to the CAG Repeat in Patients With SCA3

	Current age											
CAG repeat	0	5	10	15	20	25	30	35	40	45	50	55
67	51.6	51.6	51.6	51.6	51.6	51.6	51.6	51.6	51.6	53.5	58.4	60.5
68	49.6	49.6	49.6	49.6	49.6	49.6	49.6	49.6	49.6	49.9	54.5	—
69	52.3	52.3	52.3	52.5	52.5	52.5	52.8	52.8	53	53	56.3	57.8
70	47.7	47.7	47.7	47.7	47.7	47.7	47.7	47.7	49.1	49.6	54.5	—
71	44.5	44.5	44.5	44.5	44.5	44.5	45	47	48	49.5	54.5	—
72	40	40	40	40	40	40	40.2	41.7	44.2	47.8	—	—
73	38.7	38.7	38.7	38.7	38.7	38.8	39	39.5	45.5	46.5	—	—
74	32.3	32.3	32.3	32.3	32.5	32.8	35.5	39	—	—	—	—
75	31.8	31.8	31.8	31.8	31.8	32.5	34.3	44.5	44.8	—	—	—
76	24.5	24.5	24.5	24.5	25	30	—	—	—	—	—	—
77	24	24	24	24.2	24.3	28.8	—	—	—	—	—	—
78	22.7	22.7	26.5	26.8	27.2	—	—	—	—	—	—	—

Abbreviations: DRPLA = dentatorubral-pallidoluysian atrophy; SCA = spinocerebellar ataxia.
A machine learning approach for the prediction of age-specific probability of SCA3 and DRPLA by survival curve analysis.

acquired factors that influence the age at onset can be examined by scrutinizing cases that had developed at an age significantly different from that expected. This method can predict the probability of onset at a given age for each CAG repeat based on the current age. From the present results (e.g., assuming an SCA3 carrier with 69 repeat expansions), if the carrier is unaffected immediately after birth, the probability of developing the disease by the age of 55 years is 67%. However, if the patient has not developed the disease at age 50 years, the probability of developing the disease by the age of 55 years is 42%. We believe that these results are more clinically relevant than results from analyses other than survival analysis.

Table 4 Expected Age at Onset Using Random Survival Forest From Different Current Ages According to the CAG Repeat in Patients with DRPLA

	Current age													
CAG repeat	0	5	10	15	20	25	30	35	40	45	50	55	60	65
60	62.5	62.5	62.5	62.5	62.5	62.5	62.5	62.5	62.5	62.5	63	66.5	69	69.5
61	54.5	54.5	54.5	54.5	54.5	54.5	54.5	54.5	55	55.5	56.5	57	—	—
62	43	43	43	43	43	43.5	44	44.5	49.5	52.5	53.7	—	—	—
63	44.5	44.5	44.5	44.5	44.7	44.7	45	45	48	55	55.5	59	—	—
64	36	36	36	37	39	42	42.5	43	44	—	—	—	—	—
65	33.5	34	35.5	35.5	36.2	36.5	41	41.8	48	52.5	—	—	—	—
66	26	26	26	26	30	31.5	37.5	40	—	—	—	—	—	—
67	15.3	15.5	15.8	19	26	—	—	—	—	—	—	—	—	—
68	15	15	17.7	19.8	30	—	—	—	—	—	—	—	—	—
69	8.9	9.1	—	—	—	—	—	—	—	—	—	—	—	—
70	5.6	—	—	—	—	—	—	—	—	—	—	—	—	—

Abbreviations: DRPLA = dentatorubral-pallidoluysian atrophy; SCA = spinocerebellar ataxia.
A machine learning approach for the prediction of age-specific probability of SCA3 and DRPLA by survival curve analysis.

One limitation of this study was the small number of cases examined. The fact that some inversions were observed at the predicted onset age was assumed to be due to bias in the basic data resulting from the small number of cases. To remedy this problem, the number of cases will need to be further increased. Another issue was that the number of CAG repeats in *HTT*, *ATN1*, and *ATXN2* and DNA methylation also affect the age at onset in SCA3,[17] but these factors were not considered in this study. Future development of analytical tools that include these factors in a larger number of cases is expected.

A previous study mentioned the importance of analysis in a multiethnic cohort.[3] They acknowledge the need for a unified model across multiethnic cohorts to identify regional differences and important modifiers in decisions of the age at onset. Other groups have shown that different ethnic groups have different models that fit better within parametric analysis methods.[4] Our study was conducted in a Japanese cohort, and future validation in other ethnic groups would be required.

We have shown that machine learning methods, including RSF, can contribute to the prediction of the age at onset of polyglutamine diseases. Future validation for other diseases is expected. Furthermore, RSF can be applied to survival analysis in various fields and would be expected to improve its accuracy.

Study Funding

This study was supported by a Grant-in-Aid from the Tsubaki Memorial Foundation, Grants-in-Aid from the Research Committee on Ataxia, and a Health Labour Sciences Research Grant from The Ministry of Health, Labour and Welfare, Japan (grant number JPMH20FC1041).

Disclosure

The authors report no relevant disclosures. Go to Neurology.org/NG for full disclosure.

Publication History

Received by *Neurology: Genetics* October 6, 2022. Accepted in final form March 23, 2023. Submitted and externally peer reviewed. The handling editor was Editor Stefan M. Pulst, MD, Dr med, FAAN.

Appendix Authors

Name	Location	Contribution
Yuya Hatano, MD, PhD	Department of Neurology, Brain Research Institute, Niigata University, Niigata-shi, Japan	Drafting/revision of the article for content, including medical writing for content; study concept or design; and analysis or interpretation of data
Tomohiko Ishihara, MD, PhD	Department of Neurology, Brain Research Institute, Niigata University, Niigata-shi, Japan	Drafting/revision of the article for content, including medical writing for content; study concept or design
Sachiko Hirokawa	Department of Neurology, Brain Research Institute, Niigata University, Niigata-shi, Japan	Major role in the acquisition of data
Osamu Onodera, MD, PhD	Department of Neurology, Brain Research Institute, Niigata University, Niigata-shi, Japan	Drafting/revision of the article for content, including medical writing for content; study concept or design

References

1. Shao J, Diamond MI. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. *Hum Mol Genet.* 2007;16(R2):R115-R123. doi. 10.1093/hmg/ddm213
2. Tezenas du Montcel S, Durr A, Rakowicz M, et al. Prediction of the age at onset in spinocerebellar ataxia type 1, 2, 3 and 6. *J Med Genet.* 2014;51(7):479-486. doi. 10.1136/jmedgenet-2013-102200
3. Peng L, Chen Z, Long Z, et al. New model for estimation of the age at onset in spinocerebellar ataxia type 3. *Neurology.* 2021;96(23):e2885-e2895. doi. 10.1212/WNL.0000000000012068
4. de Mattos EP, Leotti VB, Soong BW, et al. Age at onset prediction in spinocerebellar ataxia type 3 changes according to population of origin. *Eur J Neurol.* 2019;26(1):113-120. doi. 10.1111/ene.13779
5. Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length: prediction of the age of onset and penetrance for HD. *Clin Genet.* 2004;65(4):267-277. doi. 10.1111/j.1399-0004.2004.00241.x
6. Myszczynska MA, Ojamies PN, Lacoste AMB, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol.* 2020;16(8):440-456. doi. 10.1038/s41582-020-0377-8
7. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J.* 2015;13:8-17. doi. 10.1016/j.csbj.2014.11.005
8. Peng L, Chen Z, Chen T, et al. Prediction of the age at onset of spinocerebellar ataxia type 3 with machine learning. *Mov Disord.* 2021;36(1):216-224. doi. 10.1002/mds.28311
9. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat.* 2008;2(3):841-860. doi. 10.1214/08-AOAS169
10. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol.* 2018;18(1):24. doi. 10.1186/s12874-018-0482-1
11. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. *Sci Rep.* 2019;9(1):6994. doi. 10.1038/s41598-019-43372-7
12. Dorschner MO, Barden D, Stephens K. Diagnosis of five spinocerebellar ataxia disorders by multiplex amplification and capillary electrophoresis. *J Mol Diagn.* 2002;4(2):108-113. doi. 10.1016/S1525-1578(10)60689-7
13. Maruyama S, Saito Y, Nakagawa E, et al. Importance of CAG repeat length in childhood-onset dentatorubral-pallidoluysian atrophy. *J Neurol.* 2012;259(11):2329-2334. doi. 10.1007/s00415-012-6493-7
14. Kawaguchi Y, Okamoto T, Taniwaki M, et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat Genet.* 1994;8(3):221-228. doi. 10.1038/ng1194-221
15. Koide R, Ikeuchi T, Onodera O, et al. Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nat Genet.* 1994;6(1):9-13. doi. 10.1038/ng0194-9
16. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw.* 2010;36(11):1-13. doi. 10.18637/jss.v036.i11
17. Li J, Shu A, Sun Y, et al. DNA methylation age acceleration is associated with age of onset in Chinese spinocerebellar ataxia type 3 patients. *Neurobiol Aging.* 2022;113:1-6. doi. 10.1016/j.neurobiolaging.2022.02.002