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

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.
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 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.6-8 In polyglutamine diseases, the machine learning method XGBoost has also been used for a more accurate prediction of the age at onset of SCA3.6 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 DeepSurv10 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**

Boruta16 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 evaluated17 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 rsfc 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,
respectively. Furthermore, we defined the asymptomatic probability at a certain age, \( \text{age}_C \), between \( \text{age}_B \) and \( \text{age}_A \) as \( \text{pro}_C = \text{age}_C \times (\text{pro}_B - \text{pro}_A)/(\text{age}_B - \text{age}_A) + \text{pro}_A - \text{age}_A \times (\text{pro}_B - \text{pro}_A)/(\text{age}_B - \text{age}_A) \) 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 \((\text{asymptomatic probability at age} \ Y \text{ if asymptomatic at age} \ 0) / (\text{asymptomatic probability at age} \ X \text{ if asymptomatic at age} \ 0)\) 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.
of the RSF analysis (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\(^\text{16}\) 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.
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
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 3

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

<table>
<thead>
<tr>
<th>Current age</th>
<th>CAG repeat</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
<td>51.6</td>
</tr>
<tr>
<td>68</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
</tr>
<tr>
<td>69</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
<td>52.3</td>
</tr>
<tr>
<td>70</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
<td>47.7</td>
</tr>
<tr>
<td>71</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
<td>44.5</td>
</tr>
<tr>
<td>72</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>73</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
<td>38.7</td>
</tr>
<tr>
<td>74</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
<td>32.3</td>
</tr>
<tr>
<td>75</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
</tr>
<tr>
<td>76</td>
<td>24.5</td>
<td>24.5</td>
<td>24.5</td>
<td>24.5</td>
<td>24.5</td>
<td>24.5</td>
<td>24.5</td>
<td>24.5</td>
<td>25</td>
<td>30</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>77</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24.2</td>
<td>24.3</td>
<td>28.8</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>78</td>
<td>22.7</td>
<td>22.7</td>
<td>26.5</td>
<td>26.8</td>
<td>27.2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

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.

### 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 | 62.5 | 62.5 | 62.5 | 62.5 |
| 61          | 54.5       | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 | 54.5 |
| 62          | 43         | 43   | 43   | 43   | 43   | 43   | 43   | 43.5 | 44   | —   | —   | —   | —   | —   |
| 63          | 44.5       | 44.5 | 44.5 | 44.5 | 44.5 | 44.7 | 44.7 | 44.7 | 45   | —   | —   | —   | —   | —   |
| 64          | 36         | 36   | 36   | 37   | 39   | 42   | 42.5 | 42.5 | 43   | 44   | —   | —   | —   | —   |
| 65          | 33.5       | 34   | 35.5 | 35.5 | 36.2 | 36.5 | 41   | 41.8 | 41.8 | 48   | 52.5 | —   | —   | —   |
| 66          | 26         | 26   | 26   | 26   | 30   | 31.5 | 37.5 | 40   | 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 approach 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

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yuya Hatano, MD, PhD</td>
<td>Department of Neurology, Brain Research Institute, Niigata University, Niigata-shi, Japan</td>
<td>Drafting/revision of the article for content, including medical writing for content; study concept or design; and analysis or interpretation of data</td>
</tr>
</tbody>
</table>

References
17. Onodera, Hatano, Hirakawa, Niigata University, Niigata-shi, Brain Research Institute, Department of Neurology, Niigata University, Niigata-shi, Japan, Japan, Japan, Japan

Neurology.org/NG | Neurology: Genetics | Volume 9, Number 3 | June 2023
Machine Learning Approach for the Prediction of Age-Specific Probability of SCA3 and DRPLA by Survival Curve Analysis
Yuya Hatano, Tomohiko Ishihara, Sachiko Hirokawa, et al.
Neurol Genet 2023;9;
DOI 10.1212/NXG.0000000000200075

This information is current as of May 4, 2023

Updated Information & Services
including high resolution figures, can be found at:
http://ng.neurology.org/content/9/3/e200075.full.html

References
This article cites 17 articles, 1 of which you can access for free at:
http://ng.neurology.org/content/9/3/e200075.full.html##ref-list-1

Permissions & Licensing
Information about reproducing this article in parts (figures, tables) or in its entirety can be found online at:
http://ng.neurology.org/misc/about.xhtml#permissions

Reprints
Information about ordering reprints can be found online:
http://ng.neurology.org/misc/addir.xhtml#reprintsus