This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/ene.13779
This article is protected by copyright. All rights reserved.


Departamento de Bioquímica, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, 90035-000, Brazil.

Departamento de Medicina Interna, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, 90035-003, Brazil.

* Contributed equally.

**Corresponding author:**

Laura B. Jardim, MD, PhD

Hospital de Clínicas de Porto Alegre

Rua Ramiro Barcelos 2350, 90035-003 Porto Alegre, Brazil

Phone: + 55 51 3359 8011 | Fax: + 55 51 3359 8010

e-mail: ljardim@hcpa.edu.br

**Running title:**

Age at onset prediction in SCA3

This article is protected by copyright. All rights reserved.
Keywords:
age at onset; genetic modifier; Machado-Joseph disease; spinocerebellar ataxia type 3; survival models

Disclosure of Conflict of Interest:
EPM, GVF, MLSP and LBJ were supported by the National Council for Research and Development (CNPq), Brazil. MMML received grants from Fundação para a Ciência e Tecnologia, Portugal. JAMS received unrestricted research grant from and is in the advisory board of PTC therapeutics. LBJ received grants from Fundo de Incentivo à Pesquisa do Hospital de Clínicas de Porto Alegre, and Fundo de Apoio à Pesquisa do Rio Grande do Sul, Brazil.

Age at onset prediction in spinocerebellar ataxia type 3 changes according to population of origin

Eduardo P. de Mattos¹*, Vanessa B. Leotti²*, Bing-wen Soong³*, Mafalda Raposo⁴, Maria M. M. Lima⁵, João Vasconcelos⁶, Helena Fussiger⁷, Gabriele N. Souza⁸, Nathalia Kersting⁹, Gabriel V. Furtado³, Jonas A. M. Saute⁸, Suzi A. Camey⁹, Maria Luiza Saraiva-Pereira³⁸, Laura B. Jardim¹¹


² Departamento de Estatística, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, 91509-900, Brazil.

³ Department of Neurology, Shuang Ho Hospital, Taipei Medical University School of Medicine, Taiwan.
Corresponding author:

Laura B. Jardim, MD, PhD

Hospital de Clínicas de Porto Alegre

Rua Ramiro Barcelos 2350, 90035-003 Porto Alegre, Brazil

Phone: + 55 51 3359 8011 | Fax: + 55 51 3359 8010

e-mail: ljjardim@hcpa.edu.br

This article is protected by copyright. All rights reserved.
Abstract

Introduction: In spinocerebellar ataxia type 3/Machado-Joseph disease (SCA3/MJD), length of CAG repeats expansions in ATXN3 shows an inverse correlation with age at onset (AO). Recently, a formula for predicting age at onset based on CAG expansion was developed for European carriers. We tested this formula in SCA3/MJD carriers from distinct origins and developed population-specific models to predict AO.

Methods: Parametric survival modelling.

Results: The European formula was tested in 739 independent SCA3/MJD carriers from South Brazil, Taiwan and Portuguese Azorean islands, and it largely underestimated AO in South Brazilian and Taiwanese test cohorts. This finding challenged the universal use of the European formula, leading us to develop and validate population-specific models for AO prediction. Using validation cohorts, we showed that Brazilian and Taiwanese formulas largely outperformed the European formula in a population-specific manner. Inversely, the European formula was more accurate at predicting AO among Portuguese Azorean carriers.
patients. Hence, specific prediction models were required for each SCA3/MJD ethnic group.

**Discussion:** Our data strongly support the existence of yet unknown factors that modulate AO in SCA3/MJD in a population-dependent manner, independently from CAG expansion length. The generated models are made available to the scientific community, since they can be useful for future studies on SCA3/MJD carriers from distinct geographical origins.

1. **Introduction**

Spinocerebellar ataxias (SCAs) are adult-onset neurodegenerative disorders with autosomal dominant inheritance. Several SCAs are caused by CAG repeat expansions (CAGexp) within coding regions of unrelated genes, being translated into neurotoxic polyglutamine (polyQ)-containing proteins [1].

In SCA3/Machado-Joseph disease (SCA3/MJD), CAGexp length determines ~50% of age at onset (AO) variability [2–5]. While the average (range) CAGexp length in European and North American SCA3/MJD patients was found to be 68.0 (54–77) [3,6,7], these values are skewed towards longer repeats (average: 75.1, range: 65–91) in Brazilian carriers [6,8]. Consistently, younger AO were observed in Brazilians, when compared to European and North American patients [2,3,6,8–11].

AO prediction might impact genetic counselling and recruitment of pre-ataxic individuals for future clinical trials. To address this issue, maximum likelihood estimation of AO of first symptoms (AOfs) was previously used in SCA3/MJD [12], while others employed parametric survival models to estimate the probability of onset at a given age in Huntington’s disease [for instance, 13], and in Cuban SCA2 [14], and European SCA1, SCA2, SCA3/MJD, and SCA6 carriers [7]. These models were assumed to be applicable to independent patient cohorts. Recent evidence however, suggests that additional, CAGexp-
independent factors might influence AO [4], arguing against a single AO prediction model and suggesting that the population background should be considered.

Here, we compared known AO of gait ataxia (AOga) in SCA3/MJD to predictions from the published European model [7], using independent cohorts from distinct geographical origins: South Brazil, Taiwan, and Portuguese Azorean Islands. We then developed and validated population-specific models for South Brazilian (BF) and Taiwanese (TF) carriers.

2. Methods

A workflow illustrating the cohorts and analysis rationale used in this study is shown in File S1. SCA3/MJD subjects (n=739) were stratified according to geographical origin. AOga was considered the age at the first walking disturbances, as reported by carriers and/or relatives (see File S2 for details).

2.1. Study populations

2.1.1. South Brazilian cohort one

One hundred symptomatic and 50 asymptomatic carriers were recruited from the Rio Grande do Sul SCA3/MJD population [8]. Their data was used to test the model built with data from European patients, referred here as to “European formula” (EF) for AOga determination [7] (validation cohort for EF) and to generate BF (discovery cohort for BF).

2.1.2. South Brazilian cohort two

Additional 107 patients from the South Brazilian SCA3/MJD population [8] were enrolled in a validation cohort to address predictions of BF, EF, and TF.

This article is protected by copyright. All rights reserved.
2.1.3. Taiwanese cohort three

Forty asymptomatic and 227 symptomatic Taiwanese individuals, randomly assigned from an original cohort of 347 symptomatic subjects, were used to test EF (validation cohort for EF), and to generate TF (discovery cohort for TF).

2.1.4. Taiwanese cohort four

The remaining 120 symptomatic Taiwanese individuals composed a validation cohort to address predictions of BF, EF, and TF.

2.1.5. Portuguese Azorean cohort five

Ninety-five symptomatic individuals from the Azores Islands, Portugal [4], were used to test BF, EF, and TF. SCA3/MJD in South Brazil is virtually entirely traced back to Azoreans who settled this region between 1750 and 1770 [15]. Due to this genetic closeness, cohort five was chosen to validate prediction differences between BF and EF.

2.2. Clinical and molecular diagnosis

Individuals were genotyped at their local institutions where this study was conducted. Length of CAG repeats was determined by polymerase chain reaction using fluorescent primers for the ATXN3 CAG repeat region and capillary electrophoresis.
2.3. Ethical aspects

This study was approved by the Ethics Committees of Hospital de Clínicas de Porto Alegre (14-0204), University of Azores (2/2016) and Shuang Ho Hospital, Taiwan. Confidentiality was guaranteed to all study participants, who gave written informed consent to participate in the study. Results from Brazilian (n=50) and Taiwanese (n=40) asymptomatic individuals were dealt in a pseudonymized manner under an arrangement that ensured that results were not disclosed to anyone but principal investigators in Brazil (LBJ) and Taiwan (BwS), respectively.

2.4. Statistical analysis

Prediction of median AOga for a given CAGexp length was calculated for all individuals from South Brazilian cohort one, Taiwanese cohort three and Portuguese Azorean cohort five, using EF for AOga determination [7]. Critical ranges (CR, 5th and 95th percentiles) for predicted AOga were obtained (File S3). Scatter plots were used to compare differences between observed and predicted AOga in symptomatic individuals from all cohorts, or between the age at the last asymptomatic neurologic evaluation and the predicted AOga for asymptomatic individuals. Differences between observed and predicted values, expressed as mean prediction errors (MPE), were assessed by paired t-tests. Positive MPEs indicated underestimations (observed AOga - predicted AOga > 0 years). Inaccurate AOga predictions for asymptomatic carriers were only detected when the predicted AO was earlier than the actual age of the individual at the time of her/his last clinical evaluation.

New prediction models were developed using data from the South Brazilian cohort one (Brazilian formula, BF) and the Taiwanese cohort three (Taiwanese formula, TF). For each cohort, data from symptomatic and asymptomatic carriers was fitted to four
parametric survival models (Log-Normal, Gaussian, Exponential, and Weibull). The best fitting models were chosen through residual analysis, Akaike Information Criterion (AIC) and visual comparison with Kaplan-Meier curves (File S4). Since data from both cohorts were better explained by a Gaussian parametric survival model, we adapted EF [7] to accommodate a Gaussian distribution (Files S5, S6, and S7).

Both BF and TF were fitted in the statistical software R, version 3.2.2, using the survreg function from the survival package. The print.psm function from the rms package was used to calculate $R^2$ [16]. Results were considered statistically significant when $p<0.05$.

3. Results

3.1. Age at onset and CAGexp differences among distinct SCA3/MJD populations

South Brazilian cohorts had mean AOfs, AOga and CAGexp values very similar to those reported previously for this population [8]. Mean CAGexp was statistically different among Portuguese Azorean (Table 1; smaller mean CAGexp), Taiwanese (intermediate) and South Brazilian carriers (larger mean CAGexp; two-tailed ANOVA with Tukey’s post hoc test, $F_{2,736}=69.51$, $p<0.05$). However, mean AOga was only significantly different between Taiwanese and South Brazilian patients ($F_{2,646}=5.08$, $p<0.05$).

3.2. Accuracy of the European prediction model (EF)

EF underestimated known AOga of South Brazilian (Fig. 1A) and Taiwanese (Fig. 1B) patients. Mean prediction errors ranged from almost 6 years (Taiwanese group), to more than a decade (South Brazilian group; File S8). EF predictions were more accurate for Portuguese Azorean patients (Figure 1C and File S8).
EF also underestimated the predicted AOga of South Brazilian and Taiwanese asymptomatic carriers (File S9). For 33/50 (66.0%) South Brazilian and 21/40 (52.5%) Taiwanese pre-clinical carriers (66%), EF predicted younger AOga than the age of the individuals at the last asymptomatic neurologic evaluation. Underestimations had a median of 6.93 (25th percentile: 5.31, 75th percentile: 13.71) and 6.12 (25th percentile: 3.44, 75th percentile: 13.42) years for South Brazilian and Taiwanese asymptomatic carriers, respectively.

3.3. Development and validation of population-specific models for AOga prediction

Based on parameters estimated from the South Brazilian cohort one and Taiwanese cohort three (see Methods and File S7), new models BF and TF were respectively generated and used to estimate AOga in carriers within a wide range of CAGexp lengths (Fig. 2A, 2B, Files S5 and S6). CAGexp explained, on average, 58.9% and 50.0% of AOga variance in the South Brazilian cohort one and Taiwanese cohort three, respectively. Inclusion of CAG length at the non-expanded ATXN3 allele was not significant (BF: p=0.352; TF: p=0.907).

BF and TF were then validated in the South Brazilian cohort two, Taiwanese cohort four, and Portuguese Azorean cohort five. BF was accurate for South Brazilian individuals (Fig. 3A), but yielded large overestimations in Taiwanese (Fig. 3B) and Portuguese Azorean (Fig. 3C) patients (see also File S8). A similar population specificity was observed for TF, which was more accurate for Taiwanese carriers (Figure 3E), when compared to predictions for South Brazilian (Figure 3D) and Portuguese Azorean (Figure 3F) individuals (Files S8 and S9). Excel files were made available to estimate AOga according to CAGexp for European (EF), Brazilian (BF), and Taiwanese (TF) SCA3/MJD carriers (Files S3, S5, and S6, respectively).

This article is protected by copyright. All rights reserved.
4. Discussion

EF largely underestimated AOga in South Brazilian and Taiwanese SCA3/MJD carriers, even in subjects with CAGexp tracts in the range seen in European carriers. We then developed Brazilian and Taiwanese prediction models, and validated them exclusively in a population-specific manner. These data suggest that (i) different AOga prediction formulas should be used for distinct ethnic groups, and (ii) CAGexp at \( \text{ATXN3} \) has a differential contribution to AOga in distinct populations, possibly due to population-specific modifying factors.

Differences in AO and CAGexp between European and South Brazilian SCA3/MJD carriers are acknowledged for a long time [8]. Although South Brazilian SCA3/MJD individuals tend to have longer CAGexp tracts than Europeans [2,3,5,6,8–11], we expected that EF should be more accurate, at least for individuals with expansions within the European range. However, this was not the case (Fig. 1A, 1B, and Files S8, S9, and S10). For instance, EF predicted AOga at 39.6 years for individuals with 68 CAG repeats, while BF and TF predicted much later AOga: 57.5 and 47.0 years, respectively (File S3, S5, and S6). Populational specificity for the CAGexp-AO relationship was also suggested by the fact that the Gaussian model yielded the best fit for South Brazilian and Taiwanese cohorts, instead of the LogNormal model used for EF [7]. We speculate that if the \( \text{ATXN3} \) CAGexp range observed in South Brazilians would occur among Europeans, the effect on AOga would be fateful, producing very early-onset cases.

Since BF largely overestimated AOga for Portuguese Azoreans, perhaps the European pattern of AOga dependency on CAGexp might have suffered a bottleneck/founder effect, when Portuguese Azoreans settled in South Brazil. These settlers might have brought longer CAGexp tracts than the median expansion range from the original European population, and simultaneously might have encoded protective factors, maybe in \textit{cis} with CAGexp, that resulted in delayed disease onset and partially
counterbalance the anticipatory effect of longer CAGexp in South Brazil. Conversely, there could be protective factors in the general population background, either genetic or environmental.

Although the accuracy of predictions using EF, BF and TF in cohorts from other geographical origins is difficult to foresee, we anticipate that these models will require further local adjustments of parameters.

In conclusion, EF, BF and TF were validated for Portuguese Azorean, South Brazilian and Taiwanese SCA3/MJD carriers, respectively. We generated open-access files that help predicting AOga according to a given CAGexp length, both at birth and at any given age at evaluation, for European, South Brazilian and Taiwanese carriers (Files S3, S5, and S6, respectively). If regional differences in AOga determination are confirmed in additional SCA3/MJD cohorts, this finding could have a deep impact not only on genetic counselling and on recruitment strategies of asymptomatic individuals for clinical trials, but also on the search for modifier factors that delay disease onset.

Acknowledgements:

The authors would like to thank the individuals who agreed to participate in this study.

Funding:

This work was supported by the following Brazilian agencies: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; 402968/2012-3); Fundo de Incentivo à Pesquisa do Hospital de Clínicas de Porto Alegre (14-0204); and Fundo de Apoio à Pesquisa do Rio Grande do Sul (1209-2551/13-4). Work in the Azores was supported by
the project EXOS3 (PTDC/DTP-PIC/2638/2014) funded by FEDER through the COMPETE Program and by National funding through Fundação para a Ciência e a Tecnologia (FCT).

References


Legends:

Table 1: Clinical and molecular data of SCA3/MJD cohorts enrolled in the present study.

Figure 1: Prediction of age at onset of gait ataxia (AOga) in South Brazilian (A), Taiwanese (B) and the Portuguese Azorean (C) symptomatic carriers using the European prediction model from Tezenas du Montcel et al. (2014b).

Figure 2: Population-specific predictions of age at onset of gait ataxia (AOga) in South Brazilian (A) and Taiwanese (B) SCA3/MJD individuals. Curves represent onset estimates at birth and at 25, 30, 35, 40, and 45 years of age. The x-axes of A and B are not drawn to the same scale due to distinct ranges of CAG repeat length in the two cohorts.

Figure 3: Validation of Brazilian (gray) and Taiwanese (green) models for prediction of age at onset of gait ataxia in South Brazilian (A, D), Taiwanese (B, E) and Azorean (C, F) SCA3/MJD patients. Shaded areas represent a 90% critical range.

Supplementary Files

Supplementary File 1: Recruitment and analysis workflow of SCA3/MJD cohorts included in the present study.
Supplementary File 2: Age at onset of first symptom versus age at onset of gait ataxia in SCA3/MJD patients from the Brazilian cohort one, and a discussion on the differences between these concepts.

Supplementary File 3: Predictions of the age at onset of gait ataxia for SCA3/MJD carriers based on length of the CAG expansion in ATXN3, according to the European formula (Tezenas du Montcel et al. 2014b).

Supplementary File 4: Generation of population-specific models for prediction of the age at onset of gait ataxia (AOga) in SCA3/MJD individuals.

Supplementary File 5: Predictions of the age at onset of gait ataxia for SCA3/MJD carriers based on length of the CAG expansion in ATXN3, according to the Brazilian formula (BF) described here.

Supplementary File 6: Predictions of the age at onset of gait ataxia for SCA3/MJD carriers based on length of the CAG expansion in ATXN3, according to the Taiwanese formula (TF) described here.

Supplementary File 7: Extended statistical methods.
**Supplementary File 8:** Comparison of performances of the Brazilian, European and Taiwanese prediction models of age at onset of gait ataxia in geographically distinct SCA3/MJD cohorts. Mean prediction error refers to the difference between observed and predicted ages at onset of gait ataxia.

**Supplementary File 9:** Prediction of the age at onset of gait ataxia (AOga) in South Brazilian and Taiwanese asymptomatic SCA3/MJD carriers using the European prediction model (EF).

**Supplementary File 10:** Additional comparisons of three population-specific models for prediction of the age at onset of gait ataxia (AOga) in SCA3/MJD carriers.
Table 1: Clinical and molecular data of SCA3/MJD cohorts enrolled in the present study.

<table>
<thead>
<tr>
<th></th>
<th>South Brazilian cohort one</th>
<th>South Brazilian cohort two</th>
<th>Taiwanese cohort three</th>
<th>Taiwanese cohort four</th>
<th>Portuguese Azorean cohort five</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients</td>
<td>Carriers</td>
<td>Patients</td>
<td>Patients</td>
<td>Carriers</td>
</tr>
<tr>
<td>N</td>
<td>100</td>
<td>50</td>
<td>107</td>
<td>227</td>
<td>40</td>
</tr>
<tr>
<td>Females, N (%)</td>
<td>58 (58.0)</td>
<td>27 (54.0)</td>
<td>57 (53.3)</td>
<td>112 (49.3)</td>
<td>19 (47.5)</td>
</tr>
<tr>
<td>ATXN3 CAGnorm*</td>
<td>22.5</td>
<td>21.3</td>
<td>20.9</td>
<td>20.5</td>
<td>18.5</td>
</tr>
<tr>
<td>(4.6; 14-31)</td>
<td>(5.7; 14-32)</td>
<td>(5.6; 13-36)</td>
<td>(6.7; 14-35)</td>
<td>(6.1; 14-32)</td>
<td>(6.7; 14-36)</td>
</tr>
<tr>
<td>ATXN3 CAGexp*</td>
<td>75.5</td>
<td>73.7</td>
<td>75.6</td>
<td>72.0</td>
<td>70.0</td>
</tr>
<tr>
<td>(3.2; 67-83)</td>
<td>(3.0; 68-84)</td>
<td>(3.4; 67-86)</td>
<td>(5.2; 55-87)</td>
<td>(4.2; 56-78)</td>
<td>(4.8; 60-87)</td>
</tr>
<tr>
<td>AO gait ataxia*</td>
<td>35.0</td>
<td>NA</td>
<td>35.1</td>
<td>38.6</td>
<td>NA</td>
</tr>
<tr>
<td>(10.1; 15-58)</td>
<td>(10.6; 12-55)</td>
<td>(13.5; 7-80)</td>
<td>(12.0; 8-69)</td>
<td>(12.1; 13-71)</td>
<td></td>
</tr>
<tr>
<td>Age at last asymptomatic evaluation*</td>
<td>NA</td>
<td>31.8</td>
<td>NA</td>
<td>36.3</td>
<td>NA</td>
</tr>
<tr>
<td>SARA score*</td>
<td>11.9</td>
<td>NA</td>
<td>13.0</td>
<td>ND</td>
<td>NA</td>
</tr>
<tr>
<td>(4.7; 3.0-21.0)</td>
<td>(4.7; 4.0-29.0)</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

*Mean (standard deviation; range); a Data available for 47 patients; b data available for 57 patients; AO: age at onset; CAGexp: CAG length of the expanded allele; CAGnorm: CAG length of the non-expanded allele; NA: not applicable; ND: not described; SARA: Scale for the Assessment and Rating of Ataxia.