ORIGINAL ARTICLE

Prediction of the age at onset in spinocerebellar ataxia type 1, 2, 3 and 6

Sophie Tezenas du Montcel,[1,2] Alexandra Durr,[3,4,5,6] Maria Rakowicz,[7]
Lorenzo Nanetti,[8] Perrine Charles,[3,4,5,6] Anna Sulek,[9] Caterina Mariotti,[8] Rafal Rola,[10]
Ludger Schols,[11,12] Peter Bauer,[13] Isabelle Dufaure-Garé,[1] Heike Jacobi,[14]
Sylvie Forlani,[3,4,5] Tanja Schmitz-Hübsch,[15] Alessandro Filla,[16] Dagmar Timmann,[17]
Bart P van de Warrenburg,[18] Cecila Marelli,[3,4,5,6,19] Jun-Suk Kang,[20] Paola Giunti,[21]
Arron Cook,[21] Laszlo Baliko,[22] Melegh Bela,[22] Sylvia Boesch,[23] Sandra Szymanski,[24]
José Berciano,[25,26] Jon Infante,[25,26] Katrin Buerk,[27] Marcella Masciullo,[28]
Roberto Di Fabio,[29] Chantal Depondt,[30] Susanne Ratka,[31] Giovanni Stevanin,[3,4,5,6,32]
Thomas Klockgether,[33,34] Alexis Brice,[3,4,5,6,35] Jean-Louis Golmard[1,2]

Correspondence to
Dr Sophie Tezenas du Montcel, Unité de Biostatistiques et Information Medicale, Département de Santé Publique, Bâtiment Mazarin 3eme étage, Groupe Hospitalier Pitié-Salpêtrière, 47-83 Bd de l'Hopital, 75651 Paris Cedex 13, France; sophie.tezenas@psl.aphp.fr

Open Access Scan to access more free content

ABSTRACT

Background The most common spinocerebellar ataxias (SCA)—SCA1, SCA2, SCA3, and SCA6—are caused by (CAG)n repeat expansion. While the number of repeats of the coding (CAG)n expansions is correlated with the age at onset, there are no appropriate models that include both affected and preclinical carriers allowing for the prediction of age at onset.

Methods We combined data from two major European cohorts of SCA1, SCA2, SCA3, and SCA6 mutation carriers: 1187 affected individuals from the EUROSCA registry and 123 preclinical individuals from the RISCA cohort. For each SCA genotype, a regression model was fitted using a log-normal distribution for age at onset with the repeat length of the alleles as covariates. From these models, we calculated expected age at onset from birth and conditionally that this age is greater than the current age.

Results For SCA2 and SCA3 genotypes, the expanded allele was a significant predictor of age at onset (-0.105 ± 0.005 and -0.056 ± 0.003) while for SCA1 and SCA6 genotypes both the size of the expanded and normal alleles were significant (expanded: -0.049 ± 0.002 and -0.090 ± 0.009, respectively; normal: $+0.013 \pm 0.005$ and -0.029 ± 0.010, respectively). According to the model, we indicated the median values (90% critical region) and the expectancy (SD) of the predicted age at onset for each SCA genotype according to the CAG repeat size and current age.

Conclusions These estimations can be valuable in clinical and research. However, results need to be confirmed in other independent cohorts and in future longitudinal studies.

ClinicalTrials.gov, number NCT01037777 and NCT00136630 for the French patients.

INTRODUCTION

Autosomal dominant cerebellar ataxias, also known as spinocerebellar ataxias (SCA), are neurodegenerative diseases that are clinically and genetically heterogeneous. Major advances have been made in the understanding of their causes since the 1990s and mutations in more than 20 genes have been identified thus far to be responsible for different forms of the disease. These mutations are comprised of conventional mutations, non-coding nucleotide expansions, and coding (CAG)n expansions.[1] SCA1, SCA2, Machado-Joseph or SCA3, SCA6, SCA7, SCA12, SCA17, and dentatorubral-pallidoluysian atrophy (DRPLA) are caused by (CAG)n repeat expansions in the *ATXN1*, *ATXN2*, *ATXN3*, *CACNA1A*, *ATXN7*, *PPP2R2B*, *TBP*, and *ATN1* genes, respectively, and all lead to the expansion of a polyglutamine tract in the corresponding proteins. Repeat-associated non-ATG translation (RAN) of polyglutamine tracts has also been observed in SCA8 and may contribute to the disease process.[2] All so-called polyglutamine ataxias share many common features, including a negative relationship between age at onset and the number of repeats in the expansion, and a more severe disease with larger expansions. The mean age at onset of symptoms for SCA1, SCA2, SCA3, and SCA7 carriers is generally in the third or fourth decade of life, but an average of 20 years later for SCA6 carriers.[3] The threshold of CAG expansions, or the number of expansions that determines disease carrier status, varies between the different forms of SCA as do the boundaries between what is considered an expanded and normal size (overlapping in SCA1). In most forms this threshold can be found around 40 repeats, except for in SCA6 where it is closer to 20.[1]

Gait ataxia is the first symptom identified in the majority of cases of these diseases. Globas et al[4] have shown that only 12% of SCA1, 13% of SCA2, 15% of SCA3, and 24% of SCA6 patients have other symptoms before the onset of gait ataxia. Nevertheless, the onset and the phenotype may differ considerably between two individuals with the same genotype.[5] Previous studies investigating the relationship between CAG repeat length and age at onset are of limited use in predicting the mean age at onset, as they have relied on simple linear correlations in patients and did not build predictive models that take into account information

479

from clinically unaffected mutation carriers, thus creating a bias favouring pathology. In another polyglutamine disease, Huntington's disease, similar modelling has been performed using statistical models that elucidated the relationship between CAG length and age at onset.[6–8] In SCA, a similar approach was used in the Cuban SCA2 population,[9] although this approach has not been repeated in other forms of SCA.

It is crucial that studies dealing with prediction of disease onset include both affected individuals and preclinical individuals, which has not been the case in previous models. Ignoring individuals who are free of disease symptoms, are the same age, and have the same number of CAG repeats as affected individuals creates an artificial tendency towards earlier disease onset. For the purposes of this study, we pooled genetic and age at onset data of a large group of SCA1, SCA2, SCA3, and SCA6 patients from the European EUROSCA registry with data of clinically unaffected carriers of SCA1, SCA2, SCA3, and SCA6 mutations from the RISCA study. The EUROSCA registry was established in 2004 to collect core data of European SCA patients. RISCA is a prospective, multicentric, multinational, observational cohort of clinically unaffected at-risk individuals for SCA1, SCA2, SCA3, and SCA6 (ie, first degree relatives of patients with one of these diseases).[10]

PATIENTS AND METHODS
Patients
Two groups of individuals were included: affected patients (EUROSCA registry) and preclinical mutation carriers (RISCA cohort). The EUROSCA registry includes individuals with any form of spinocerebellar ataxia (SCA) from 17 European centres. For the current study, we selected 1187 patients with a positive molecular genetic test for SCA1, SCA2, SCA3 or SCA6, genotyped at a central laboratory, and with information available on age at onset of the disease (317 SCA1, 308 SCA2, 399 SCA3, and 163 SCA6) and, when possible, a SARA (Scale for the Assessment and Rating of Ataxia, with a maximal score of 40 indicating a very severe cerebellar ataxia) score ≥3.[10] Patients were included in the database with age at onset as indicated by self-report during their examination by the neurologist, and as indicated in their medical records. Disease onset was defined by the onset of gait difficulties, as this is the most frequent first symptom. Data were obtained from patients by personal interview. Information obtained by interview was then compared to that from medical records, if available.

The RISCA cohort included individuals at-risk for SCA from 14 European centres.[11] These included adult individuals, children or siblings of an individual with SCA1, SCA2, SCA3 or SCA6. Absence of ataxia was defined as having a score on the SARA scale <3. All individuals were genotyped in the same central laboratory as the EUROSCA registry, and of the 264 individuals included with DNA available, 123 (47%) were carriers of a disease-causing expansion (50 SCA1, 31 SCA2, 26 SCA3, and 16 SCA6). For these preclinical mutation carriers, the age at examination was recorded.

All participants signed informed consent documents approved by institutional review boards and the local ethics committee.

Genotypes
Blood samples to obtain DNA for genetic testing were taken from all study participants including those who had already undergone preclinical genetic testing. All genetic tests were performed at the Institute of Medical Genetics and Applied Genomics (Tübingen, Germany) using established and standardised methods.

For the RISCA cohort, the genetic tests were done anonymously under an arrangement that guaranteed that results were not disclosed to study participants, clinical investigators or anyone else except the statistician's team (STdM and ID-G). However, all study participants were offered genetic counselling with an open preclinical testing procedure according to established clinical standards.

For both cohorts, we defined the pathological thresholds as a CAG repeat expansion of more than 39 repeats in SCA1, more than 31 repeats in SCA2, more than 47 repeats in SCA3, and more than 20 repeats in SCA6.

Prediction of age at onset
Statistical model
The prediction of age at onset was achieved using a statistical model to relate the age at onset of an individual with his genotype. As our final sample included some individuals who had not yet reached an age to be affected by the disease but will inevitably develop symptoms, the methodological framework we used was one of survival analysis. In order to make predictions about age at onset, we used a parametric survival model, namely a log-normal censored model. The age at onset was predicted from the moment of birth for a patient with known genotype using the following formulae:

$$\text{Log}(T) = \mu_G + \gamma_G E \quad (1)$$

where: T is the age at onset from birth, a random variable for a patient with known genotype, μ_G is the expectation of $\text{Log}(T)$, γ_G is the SD of $\text{Log}(T)$, and E is a random variable with a standard Gaussian probability density function.

The mean log age at onset μ_G, for a given genotype, is derived from a regression model as follows using the numbers of repeats of the two alleles n_e and n_{ne}: expanded and not expanded respectively:

$$\mu_G = \alpha_0 + \alpha_e n_e + \alpha_{ne} n_{ne} \quad (2)$$

where α_0, α_e and α_{ne} are the regression parameters that are being estimated.

The random variable T (age at onset) has a log-normal probability density function (pdf) with parameters μ_G and γ_G. Thus its pdf is:

$$f(t) = \frac{1}{t \gamma_G \sqrt{2\pi}} e^{-\frac{[\text{Log}(t) - \mu_G]^2}{2\gamma_G^2}} \quad \text{with } t > 0 \quad (3)$$

As an example $f(t)$ is plotted for SCA2 with a repeat number of the expanded allele of 37 (figure 1).

When T is censored, we need to express $S(t) = P(T > t)$. We have:

$S(t) = 1 — F(t)$, where

$$F(t) = P(T < t) = P\left(\frac{\text{Log}(T) - \mu_G}{\gamma_G} < \frac{\text{Log}(t) - \mu_G}{\gamma_G}\right)$$
$$= \Phi\left(\frac{\text{Log}(t) - \mu_G}{\gamma_G}\right) \quad (4)$$

And finally:

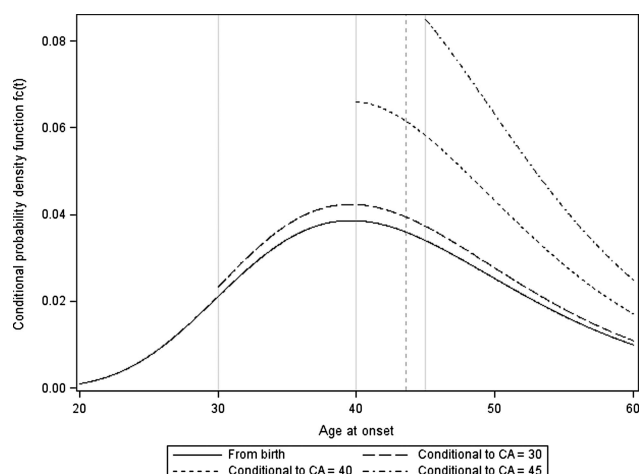$$S(t) = 1 - \Phi\left(\frac{\text{Log}(t) - \mu_G}{\gamma_G}\right) \quad (5)$$

Figure 1 Conditional probability density function of age at onset for different current ages (CA) (t). The figures are drawn for SCA2 data with a repeat number of the expanded allele of 37 for an individual at birth (solid), or at a current age of 30 (medium dash), 40 (short dash) or 45 (short dash dot). Based on a log linear parametric model, the estimation of the mean age at onset from birth was 43 years (γ_G: 3.7428, $\hat{\sigma}_G$: 0.2516) (vertical dashed line). As the distribution is log-normal the mode is neither the mean nor the median. SCA, spinocerebellar ataxias.

Estimation of the parameters

The estimation of the parameters was performed using the affected individuals from the EUROSCA registry whose age at onset is known, and the unaffected individuals carrying an expanded allele from the RISCA study. For the latter patients, the age at last examination was known, and we considered that this age was a censored value of the age at onset of the disease. The parameters were estimated by the maximum likelihood method. Backward selection was used to retain the significant

Table 1 Parameter estimates obtained from the parametric survival model

	Parameter estimate	Parameter standard error	p Value
SCA1			
Intercept	5.4952	0.1924	<0.0001
Expanded allele	−0.0487	0.0019	<0.0001
Shorter allele	0.0133	0.0055	0.0151
Standard deviation (γ_G)	0.1748	0.0069	
SCA2			
Intercept	7.6301	0.1802	<0.0001
Expanded allele	−0.1051	0.0046	<0.0001
Standard deviation (γ_G)	0.2520	0.0101	
SCA3			
Intercept	7.4908	0.1853	<0.0001
Expanded allele	−0.0564	0.0027	<0.0001
Standard deviation (γ_G)	0.2167	0.0076	
SCA6			
Intercept	6.3470	0.2684	<0.0001
Expanded allele	−0.0901	0.0091	<0.0001
Shorter allele	−0.0285	0.0102	0.0053
Standard deviation (γ_G)	0.1738	0.0095	

SCA, spinocerebellar ataxias.

parameter α_e or α_{ne}. The parameter estimation of the model was performed using the SAS V9.3 statistical software.

Computation of the predictive statistics: expectation, SD, and percentiles

In order to take into account that we used parameter estimates, we added to the variance γ_G^2 the variance of $\hat{\mu}_G$, estimated from the estimated parameters $\hat{\alpha}_0$, $\hat{\alpha}_e$ and $\hat{\alpha}_{ne}$, with

$$
\begin{aligned}
\mathrm{var}(\hat{\mu}_G) =& \mathrm{var}(\hat{\alpha}_0 + n_e \times \hat{\alpha}_e + n_{ne} \times \hat{\alpha}_{ne}) \\
=& \mathrm{var}(\hat{\alpha}_0) + \mathrm{var}(\hat{\alpha}_e) \times n_e^2 + \mathrm{var}(\hat{\alpha}_{ne}) \times n_{ne}^2 \\
&+ 2 \times \mathrm{cov}(\hat{\alpha}_0, \hat{\alpha}_e) \times n_e + 2 \times \mathrm{cov}(\hat{\alpha}_0, \hat{\alpha}_{ne}) \\
&\times n_{ne} + 2 \times \mathrm{cov}(\hat{\alpha}_e, \hat{\alpha}_{ne}) \times n_e \times n_{ne}
\end{aligned}
\tag{6}
$$

And finally:

$$
\hat{\sigma}_G^2 = \hat{\gamma}_G^2 + \mathrm{var}(\hat{\mu}_G)
\tag{7}
$$

where, $\mathrm{var}(\hat{\alpha}_0)$, $\mathrm{var}(\hat{\alpha}_e)$ and $\mathrm{var}(\hat{\alpha}_{ne})$ are the variances of the estimated parameters $\hat{\alpha}_0$, $\hat{\alpha}_e$ and $\hat{\alpha}_{ne}$ respectively (table 1), cov ($\hat{\alpha}_0$, $\hat{\alpha}_e$), cov($\hat{\alpha}_0$, $\hat{\alpha}_{ne}$) and cov($\hat{\alpha}_e$, $\hat{\alpha}_{ne}$) are the covariances between the estimated parameters (see online supplementary table S1).

We thus computed the predictive statistics from the estimated pdf of the age at onset, from birth or conditionally that this age is greater than the current age. The estimated pdf of the age at onset is given by the formulae:

$$
\hat{f}(t) = \frac{1}{t\hat{\sigma}_G\sqrt{2\pi}} e^{-([Log(t)-\hat{\mu}_G]^2/2\hat{\sigma}_G^2)}
\tag{8}
$$

where $\hat{\mu}_G$ is given by formula (2) after replacing the parameters by their estimates and $\hat{\sigma}_G^2$ is given by formula (7).

From these formulae, one can derive the values of the predictive statistics (expectation E(t), variance var(T), and percentiles t_α) of the age at onset distribution. These predictive statistics are calculated first from the moment of birth, without regard to the actual disease progression of the individual. We have:

$$
\hat{E}(T) = e^{\hat{\mu}_G + \frac{\hat{\sigma}_G^2}{2}}
\tag{9}
$$

and

$$
\widehat{\mathrm{var}}(T) = (e^{\hat{\sigma}_G^2} - 1)e^{2\hat{\mu}_G + \hat{\sigma}_G^2}
\tag{10}
$$

The αth percentile t_α is thus obtained from the inverse of F such as:

$$
\hat{t}_\alpha = \hat{F}^{-1}(\alpha) = e^{\hat{\mu}_G + \hat{\sigma}_G \Phi^{-1}(\alpha)}
\tag{11}
$$

In order to account for the fact that any given asymptomatic individual has reached his current age c without yet being affected by disease, we thus estimated the age at onset given a current age (c). As shown in figure 1, this leads to a truncation of the log-normal distribution which increases with c. As the individuals are not observed at birth, but at a current age c, we need to estimate E(T|T>c), the expectation of T given that the individual's age is more than c, the corresponding variance Var (T|T>c) and the corresponding percentiles t_α.

$$\widehat{E}_c(T) = E(T|T > c) = \frac{e^{\hat{\mu}_G + (\hat{\sigma}_G^2/2)}\Phi((\hat{\mu}_G + \hat{\sigma}_G^2 - \text{Log}(c)/\hat{\sigma}_G))}{\Phi((\hat{\mu}_G - \text{Log}(c)/\hat{\sigma}_G))} \quad (12)$$

$$\widehat{\text{var}}_c(T) = \widehat{\text{var}}(T|T > c) = \widehat{E}_c(T^2) - \widehat{E}_c(T)^2 \quad (13)$$

with

$$\widehat{E}_c(T^2) = \frac{e^{2(\hat{\mu}_G + \hat{\sigma}_G^2)}\Phi((\hat{\mu}_G + 2\hat{\sigma}_G^2 - \text{Log}(c)/\hat{\sigma}_G))}{\Phi((\hat{\mu}_G - \text{Log}(c)/\hat{\sigma}_G))} \quad (14)$$

And the ath percentile is given by:

$$\hat{\tau}_\alpha = \exp\left(\hat{\mu}_G + \hat{\sigma}_G \Phi^{-1}\left(\alpha + (1 - \alpha)\Phi\left(\frac{\text{Log}(c) - \hat{\mu}_G}{\hat{\sigma}_G}\right)\right)\right) \quad (15)$$

In this paper, we computed the 5th, 50th (median) and 95th percentiles of the T pdf. We called the (5th; 95th) interval the critical region (90% CR).

Model validation
We conducted a validation study in order to assess the goodness-of-fit of the log-normal model. For each type of SCA disease, the model's validation is based upon the comparison of the observed survival function of the whole sample, as obtained by the Kaplan-Meier method, and the sample estimated survival function. The sample estimated survival function was obtained by the crossover method: for each individual of the sample, the parameters of the model were obtained by removing the individual from the sample, and estimating the survival function of the individual based on its genotype. The sample estimated survival function is the mean of these estimated survival functions for each individual.

Sensitivity analysis
One limitation of our study is that the sample we used was obtained by merging two samples, one with affected patients and one with preclinical mutation carriers. As discussed previously in this paper, while it is crucial to include both affected and unaffected carriers, the two samples do not have the same parameters and thus the accuracy of our results may depend of the respective proportions of the two populations. In order to study the sensitivity of the results to these proportions, we conducted a sensitivity analysis with the following method: we modified the proportions of the two sub-samples by multiplying the unaffected sample size by the factors 0.5 (half of the unaffected) and 2 (twice as many unaffected). This was done by giving these weights to each individual within the preclinical mutation carrier sample, and by making all computations with these weighted samples.

RESULTS
Description of the populations
We included 1310 individuals; of these 1187 were EUROSCA affected individuals (SCA1: 317, SCA2: 308, SCA3: 399, SCA6: 163) from 735 families and the remaining 123 were RISCA unaffected individuals (SCA1: 50, SCA2: 31, SCA3: 26, SCA6: 16) from 102 families. Forty-two families included both EUROSCA affected (120 individuals) and RISCA unaffected

individuals (51 individuals). Half of the individuals were males, and half were females. SCA6 individuals were older than the individuals from the other genotypes. As expected, within each genotype, the mean age at last examination for the unaffected individuals was lower than the mean age at onset of the affected individuals (SCA1: <0.0001, SCA2: 0.0047, SCA3: 0.0051, SCA6: 0.0287). However, there was overlap as the age of some unaffected individuals was higher than the age at onset of some affected individuals (table 2).

Parametric model
For SCA2 and SCA3 genotypes only the number of repeats of the expanded allele was significantly associated with the age at onset, while for SCA1 and SCA6 genotypes the number of repeats of both alleles were significantly associated (table 1) with age at onset. The recruiting centre, family, and year at onset separated as quartile did not substantially influence the results. For all genotypes, gender was not significantly associated with age at onset, but, as expected, the expanded allele had a negative effect on the age at onset. For SCA1, the log of the age at onset decreased by 0.049 ± 0.002 (SE) (p<0.001) for each additional repeat, for SCA2 by 0.105 ± 0.005 (p < 0.001), for SCA3 by 0.056 ± 0.003 (p<0.001), and for SCA6 by 0.090 ± 0.009 (p<0.001). In addition, in SCA1, the log of the age at onset increased by 0.013 ± 0.005 (p=0.014) with each additional repeat on the shorter non-expanded allele, and in SCA6, the log age at onset decreased by 0.029 ± 0.010 (p=0.0075).

Prediction of age at onset
Based on a log-normal distribution of the age at onset, we obtained the age at onset for each genotype and the range of observed repeat lengths within each genotype. For example, an individual with 37 repeats in the *SCA2* gene would have a median age at onset of 42 years old (90% CR: 28–64) (figure 1, see online supplementary table S3). Given that this individual is unaffected at the age of 35 years, he would have a 50% risk of developing the disease before the age of 45 years (90% CR: 36–66) and if he remains unaffected at the age of 45 years, he would have a 50% risk of onset before the age of 52 years (90% CR: 46–71) (figure 2B, see online supplementary table S3). Similar results were obtained for SCA1 (figure 2A, see online supplementary table S2), for SCA3 (figure 2C, see online supplementary table S4), and for SCA6 (figure 2D, see online supplementary table S5). For all SCAs, the accuracy of prediction of the age at onset decreased with the size of the allele expansion: for those with large repeat expansions prediction was more accurate compared to those with mildly expanded alleles. In addition, on average, only 4% of the variance of the age at onset (from 1% for SCA3 to 10% for SCA6) was due to the precision of the statistical model estimation, the remaining being due to population dispersion.

The models were fitted to the observed data (see online supplementary figure S1). Furthermore, the models were robust with respect to the proportion of censored data (see online supplementary figure S2).

DISCUSSION
Using two unique cohorts (the EUROSCA and RISCA cohorts) comprised of individuals recruited at the same European centres, examined by the same clinicians and genotyped in the same centralised laboratory, we were able to estimate the relationship between the number of CAG repeats and the age at onset of gait ataxia in the genes corresponding to the four most

Table 2 Characteristics of affected and unaffected individuals according to their spinocerebellar ataxias (SCA) genotype

	SCA1		SCA2		SCA3		SCA6	
	Affected	Unaffected	Affected	Unaffected	Affected	Unaffected	Affected	Unaffected
n	317	50	308	31	399	26	163	16
Gender (M)	169 (53%)	15 (30%)	164 (53%)	20 (65%)	199 (50%)	11 (42%)	81 (50%)	5 (31%)
Age at onset (affected)/last examination (unaffected)	38±11 [11–75]	30±8 [18–51]	36±13 [7–71]	31±9 [18–47]	40±12 [10–78]	35±7 [23–50]	53±11 [24–77]	47±9 [35–68]
Expanded allele	47±5 [39–66]	47±4 [40–63]	39±3 [33–54]	38±2 [33–43]	68±4 [54–77]	68±3 [61–73]	23±2 [21–29]	21±1 [21–23]
Shorter allele	30±2 [26–37]	31±1 [29–36]	22±1 [15–29]	23±2 [22–32]	21±5 [14–35]	20±5 [14–28]	13±1 [7–16]	10±2 [6–13]
SARA	15±8 [3–40] (n=84)	1±1 [0–2.5]	15±8 [3–39] (n=114)	1±1 [0–2.5]	15±8 [3–35.5] (n=98)	1±1 [0–2.5]	15±6 [5.5–33] (n=80)	1±1 [0–2.5]

Data are expressed as n (%) or mean±SD with the range in brackets [min-max].
SARA, Scale for the Assessment and Rating of Ataxia.

frequent polyglutamine ataxia diseases: SCA1, SCA2, SCA3, and SCA6.

Disease onset as defined by the onset of gait difficulties can be variable among patients, and may also be variable depending on the presence of other patients within a given family, as the other members are likely to pay closer attention to early disease symptoms. In contrast to other neurodegenerative diseases such as Huntington's disease, in SCAs there are no psychiatric changes or anosognosia that could interfere with the identification of onset by the patient or their families. Estimations were made according to the genotype of the major gene (number of repeats of the expand allele and additionally, for SCA1 and SCA6, the number of repeats in the short allele) and according to the current age of the carrier of an expansion. There are two events necessary for the disease to develop—the presence of an abnormal CAG repeat, and advanced age. For the mutation carriers with the longest expansions, the oldest age of onset estimations are 1 or 2 years later than the current age. For these carriers, very few individuals will be unaffected in old age, therefore estimations of onset at the oldest ages are more theoretical than real. The ages at onset estimated in this study were similar to those observed and were dependent on CAG repeat length.[12] [13] In addition, as observed, the estimations' variability decreases with the number of repeats, the smaller the repeat the more accurate the onset estimation. Small but important contributions of the normal polymorphic expansion on the unaffected allele were identified in SCA1 and SCA6, but not in SCA2 and SCA3.[3] Similarly to Van de Warrenburg et al, we found a positive effect of the non-expanded CAG repeat for SCA1 and a negative effect for SCA6. This result must be confirmed in an independent cohort, as slightly less than a third of the affected individuals of the current study were also in the Van de Warrenburg et al study.

As has been done previously in Huntington's disease,[6–8] we included both affected and unaffected carriers of the mutation. If only affected subjects had been used this could have introduced a bias. Healthy carriers of an abnormally expanded repeat could be different from affected carriers of the same age. In particular, individuals with abnormal expansions in the range just above normal size are expected to start the disease late in life; consequently, they have competing mortality risk and could die of other diseases before the onset of ataxia. This is particularly relevant for SCA6 which has the latest onset of all SCAs. Using only data of affected individuals for estimating the influence of the size of the CAG repeat could lead to a bias providing unduly pessimistic estimates of age at onset.[7] To avoid these biases, a survival analysis allowed us to take into account the unaffected but censored individuals. Almaguer-Mederos et al[9] published the mean and median age at onset from birth for a range of CAG expansion sizes in SCA2 mutation carriers from a Cuban founder population. Compared to their results, our estimates produced a younger age at onset in SCA2. However, the Cuban estimations were not corrected for the current age of the patients. Even when we applied the same methodology — that is, Kaplan-Meier estimates stratified by repeat length—our estimations of age at onset were lower for each repeat length than those in the Cuban population (data not shown). This may be due to specific properties of the founder population of Cuba or to different recruitment strategies. This was also observed in Huntington's disease in the Venezuelan population as reported by the US–Venezuela Collaborative Research Project and Nancy S Wexler.[14] Both the Venezuelan Huntington cohort and the Cuban SCA sample contained affected and preclinical carriers, from large families with a

Figure 2 Median age at onset according to the genotype and the current age of the presymptomatic individual. (A) SCA1 genotype, (B) SCA2 genotype, (C) SCA3 genotype, (D) SCA6 genotype. For all panels, the x axis is the number of repeats for the expanded alleles, and the y axis the estimated age at onset. For SCA1 and SCA6 genotypes, each sub-panel representing different repeats numbers of the shorter allele are depicted. Curves are plotted from birth and for an individual of 25/30/35/40/45 years old. SCA, spinocerebellar ataxias.
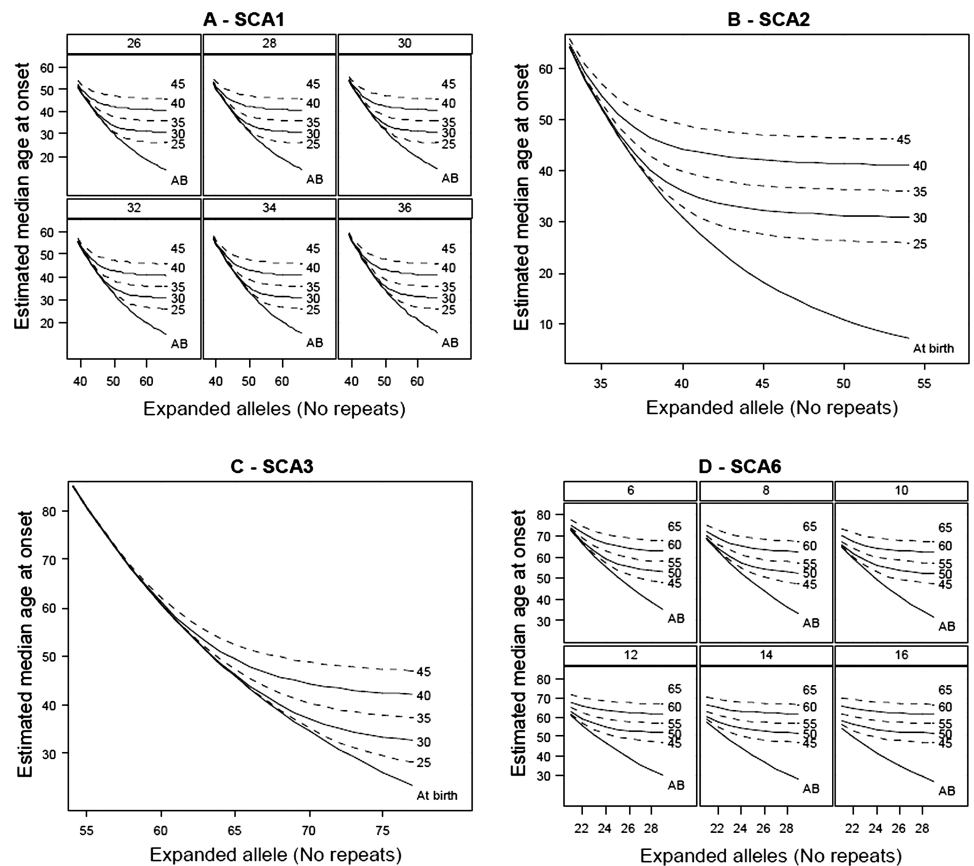
homogeneous genetic background. Their results could be due partly to a specificity of the population, for example, a modifier gene or an environmental effect present in the Cuban population but absent from our study sample population. The samples of our study were recruited in a two-step procedure: first, the affected subjects, and then, their unaffected relatives without systematic screening of the families. Because of this, there may be some carriers within these families with subclinical signs that were not included in either the affected or unaffected cohorts. This could have led to pessimistic estimations of onset age as not all unaffected expansion carriers are necessarily included. Conversely, we have shown that the inclusion of some additional unaffected carriers would have only a small impact on the estimations.

The range of repeat lengths did not cover the entire range that has been previously published. Thus, our results are only valid and usable within this smaller range. An extrapolation outside the range of observed repeats would be misleading. In addition, the present results need to be confirmed either in a replication cohort, or by longitudinal data. These data are not currently available. In addition, the subjects included in the EUROSCA and RISCA cohorts are of primarily European origin. Thus, the extension of the results to other geographical origins must be done cautiously.

Both the SDs of age at onset and the critical regions of the predicted ages—the interval where we have a 90% chance to have the observed age—were quite large. Most of the estimated age variance comes from age dispersion within the population, so it cannot be significantly decreased by a larger sample size. The use of these estimates for clinical purposes, particularly in the context of predictive testing, must be done very carefully, taking into account the variability of the estimates. Keeping

these limitations in mind, the estimations can be of help when counselling presymptomatic carriers for the patient that requests it. One risk of this kind of use could be that the knowledge of one's expected age at onset might induce an earlier onset for carriers that are aware of their genetic status. However, data from Huntington's disease do not seem to confirm this kind of effect. In a cohort of presymptomatic Huntington carriers, knowledge of one's genetic status after presymptomatic testing did result in increased auto-observation, but the onset of this disease has always been difficult to define for the carrier and the care taker, as psychiatric symptoms and anosognosia can complicate the determination of disease onset.[15] In the case of SCAs this estimation could be more accurate as anosognosia is not present in this disease. In addition, the estimates can be used for epidemiological purposes—for example, to correlate the time to onset to a particular clinical phenotype such as the score on a disease rating scale or to associated phenotypes such as cerebral imaging results. Knowing the expected age at onset in preclinical individuals, Jacobi et al[10] were able to infer that for SCA1 and SCA2 mutation carriers the extent of functional and brain structural alterations increased as the interval to the predicted age of ataxia onset decreased.

Author affiliations
[1]UPMC Univ Paris 06, ER4, Modelling in Clinical Research, Paris, France
[2]Department of Biostatistics and Medical Informatics, AP-HP, Hopitaux Universitaires Pitié-Salpétrière Charles-Foix, Paris, France
[3]UPMC Univ Paris 06, Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière, UMR-S975, Paris, France
[4]Inserm, U975, Paris, France
[5]Cnrs, UMR 7225, Paris, France
[6]Département de Génétique et Cytogénétique, AP-HP, Hopitaux Universitaires Pitié-Salpétrière Charles-Foix, Paris, France

[7]Department of Clinical Neurophysiology, Institute of Psychiatry and Neurology, Warsaw, Poland
[8]Unit of Genetics of Neurodegenerative and Metabolic Diseases, Fondazione IRCCS Istituto Neurologico "Carlo Besta", Milan, Italy
[9]Department of Genetics, Institute of Psychiatry and Neurology, Warsaw, Poland
[10]First Department of Neurology Institute of Psychiatry and Neurology, Warsaw, Poland
[11]Department of Neurology, Hertie-Institute for Clinical Brain Research, University of Tübingen, Germany
[12]German Center for Neurodgenerative Diseases (DZNE), Tübingen, Germany
[13]Institute of Medical Genetics and Applied Genomics, University Tübingen, Tübingen, Germany
[14]Department of Neurology, University Hospital of Bonn, Bonn, Germany
[15]Department of Neurology, Charité University Medicine, Berlin, Germany
[16]Department of Neurosciences, Federico II University, Napoli
[17]Department of Neurology, University Clinic Essen, University of Duisburg-Essen, Essen, Germany
[18]Department of Neurology, Radboud University Medical Centre, Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, The Netherlands
[19]Departement of Neurology, University Hospital Gui de Chauliac, Montpellier, France
[20]Department of Neurology, Goethe University Frankfurt, Frankfurt, Germany
[21]Department of Molecular Neuroscience, Institute of Neurology, UCL London, UK
[22]Department of Medical Genetics, Szentagothai Research Center, University Pécs, Hungary
[23]Department of Neurology, Medical University Innsbruck, Innsbruck, Austria
[24]Department of Neurology, St. Josef Hospital, University Hospital of Bochum, Bochum, Germany
[25]Service of Neurology, University Hospital "Marqués de Valdecilla (IFIMAV)", University of Cantabria, Spain
[26]Centro de Investigación Biomédica en Red de Enfermedades Neurodegenerativas (CIBERNED), Santander, Spain
[27]Department of Neurology, Philipps University of Marburg, Marburg, Germany
[28]IRCCS San Raffaele Pisana, Rome, Italy
[29]Department of Medical-Surgical Sciences and Biotechnologies, Sapienza University of Rome, Rome, Italy
[30]Department of Neurology, Université Libre de Bruxelles, Brussels, Belgium
[31]Department of Neurodegeneration and Restorative Research, Centers of Molecular Physiology of the Brain and Neurological Medicine, University Hospital of Göttingen, Göttingen
[32]Laboratoire de Neurogenetique, Ecole Pratique des Hautes Etudes (EPHE), Institut du Cerveau et de la Moelle épinière, Hôpital de la Salpêtrière, Paris, France
[33]Department of Neurology, University Hospital of Bonn, Bonn, Germany
[34]German Center for Neurodgenerative Diseases (DZNE), Bonn, Germany
[35]Institut du Cerveau et de la Moelle Epinière, Paris, France

REFERENCES

1 Durr A. Autosomal dominant cerebellar ataxias: polyglutamine expansions and beyond. *Lancet Neurol* 2010;9:885–94.
2 Ikeda Y, Daughters RS, Ranum LP. Bidirectional expression of the SCA8 expansion mutation: one mutation, two genes. *Cerebellum* 2008;7:150–8.
3 Van de Warrenburg BP, Hendriks H, Durr A, van Zuijlen MC, Stevanin G, Camuzat A, Sinke RJ, Brice A, Kremer BP. Age at onset variance analysis in spinocerebellar ataxias: a study in a Dutch-French cohort. *Ann Neurol* 2005;57:505–12.
4 Globas C, du Montcel ST, Baliko L, Boesch S, Depondt C, DiDonato S, Durr A, Filla A, Klockgether T, Mariotti C, Melegh B, Rakowicz M, Ribai P, Rola R, Schmitz-Hubsch T, Szymanski S, Timmann D, Van de Warrenburg BP, Bauer P, Schols L. Early symptoms in spinocerebellar ataxia type 1, 2, 3, and 6. *Mov Disord* 2008;23:2232–8.
5 Stevanin G, Durr A, Brice A. Clinical and molecular advances in autosomal dominant cerebellar ataxias: from genotype to phenotype and physiopathology. *Eur J Hum Genet* 2000;8:4–18.
6 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. *Clin Genet* 2004;65:267–77.
7 Langbehn DR, Hayden MR, Paulsen JS. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. *Am J Med Genet B Neuropsychiatr Genet* 2010;153B:397–408.
8 Paulsen JS, Langbehn DR, Stout JC, Aylward E, Ross CA, Nance M, Guttman M, Johnson S, MacDonald M, Beglinger LJ, Duff K, Kayson E, Biglan K, Shoulson I, Oakes D, Hayden M. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *J Neurol Neurosurg Psychiatry* 2008;79:874–80.
9 Almaguer-Mederos LE, Falcon NS, Almira YR, Zaldivar YG, Almarales DC, Gongora EM, Herrera MP, Batallan KE, Arminan RR, Manresa MV, Cruz GS, Laffita-Mesa J, Cyuz TM, Chang V, Auburger G, Gispert S, Perez LV. Estimation of the age at onset in spinocerebellar ataxia type 2 Cuban patients by survival analysis. *Clin Genet* 2010;78:169–74.
10 Jacobi H, Reetz K, du Montcel ST, Bauer P, Mariotti C, Nanetti L, Rakowicz M, Sulek A, Durr A, Charles P, Filla A, Antenora A, Schols L, Schicks J, Infante J,

Kang JS, Timmann D, Fabio RD, Masciullo M, Baliko L, Melegh B, Boesch S, Burk K, Peltz A, Schulz JB, Dufaure-Gare I, Klockgether T. Biological and clinical characteristics of individuals at risk for spinocerebellar ataxia types 1, 2, 3, and 6 in the longitudinal RISCA study: analysis of baseline data. *Lancet Neurol* 2013;12:650–8.

11 Schmitz-Hubsch T, du Montcel ST, Baliko L, Berciano J, Boesch S, Depondt C, Giunti P, Globas C, Infante J, Kang JS, Kremer B, Mariotti C, Melegh B, Pandolfo M, Rakowicz M, Ribai P, Rola R, Schols L, Szymanski S, van de Warrenburg BP, Durr A, Klockgether T, Fancellu R. Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology* 2006;66:1717–20.

12 Stevanin G, Dürr A, Brice A. Spinocerebellar ataxias caused by polyglutamine expansions. *Adv Exp Med Biol* 2002;516:47–77.

13 Durr A, Stevanin G, Cancel G, Duyckaerts C, Abbas N, Didierjean O, Chneiweiss H, Benomar A, Lyon-Caen O, Julien J, Serdaru M, Penet C, Agid Y, Brice A. Spinocerebellar ataxia 3 and Machado-Joseph disease: clinical, molecular, and neuropathological features. *Ann Neurol* 1996;39:490–9.

14 Wexler NS, Lorimer J, Porter J, Gomez F, Moskowitz C, Shackell E, Marder K, Penchaszadeh G, Roberts SA, Gayán J, Brocklebank D, Cherny SS, Cardon LR, Gray J, Dlouhy SR, Wiktorski S, Hodes ME, Conneally PM, Penney JB, Gusella J, Cha J-H, Irizarry M, Rosas D, Hersch S, Hollingsworth Z, MacDonald M, Young AB, Andresen JM, Housman DE, De Young MM, Bonilla E, Stillings T, Negrette A, Snodgrass SR, Martinez-Jaurrieta MD, Ramos-Arroyo MA, Bickham J, Ramos JS, Marshall F, Shoulson I, Rey GJ, Feigin A, Arnheim N, Acevedo-Cruz A, Acosta L, Alvir J, Fischbeck K, Thompson LM, Young A, Dure L, O'Brien CJ, Paulsen J, Brickman A, Krch D, Peery S, Hogarth P, Higgins DS Jr, Landwehrmeyer B, U.S.-Venezuela Collaborative Research Project. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc Natl Acad Sci USA* 2004;101:3498–503.

15 Gargiulo M, Jutras M, Tezenas du Montcel S, Benaich S, Herson A, Feingold J, Durr A. Does Presymptomatic testing influences Age at Onset in a Negative manner (PAON): follow up of at risk persons for Huntington disease. Presented at the European Society Human genetics, Paris, France, 2013.

Correction

Sophie Tezenas du Montcel, Alexandra Durr, Maria Rakowicz, *et al.* Prediction of the age at onset in spinocerebellar ataxia type 1,2,3 and 6. *J Med Genet* 2014;51:479–86. doi:10.1136/jmedgenet-2013-102200. The author's name Béla Melegh has been corrected.

CrossMark

J Med Genet 2014;**51**:613. doi:10.1136/jmedgenet-2013-102200corr1