Received: 27 November 2023 | Revised: 28 March 2024 | Accepted: 2 May 2024

DOI: 10.1002/psp4.13162

ARTICLE

Item performance of the scale for the assessment and rating of ataxia in rare and ultra-rare genetic ataxias

Alzahra Hamdan[1] | **Andrew C. Hooker**[1] | **Xiaomei Chen**[1] | **Andreas Traschütz**[2,3] | **Rebecca Schüle**[2,4] | **ARCA Study Group*** | **EVIDENCE-RND consortium** | **Matthis Synofzik**[2,3] | **Mats O. Karlsson**[1]

[1]Pharmacometrics Research Group, Department of Pharmacy, Uppsala University, Uppsala, Sweden

[2]Department of Neurodegenerative Diseases, Center for Neurology and Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

[3]German Center for Neurodegenerative Diseases (DZNE) Tübingen, Tübingen, Germany

[4]Division of Neurodegenerative Diseases, Department of Neurology, Heidelberg University Hospital, Heidelberg, Germany

Correspondence
Mats O. Karlsson, Department of Pharmacy, Uppsala University, Box 580, Uppsala 751 23, Sweden.
Email: mats.karlsson@farmaci.uu.se

Abstract

The Scale for the Assessment and Rating of Ataxia (SARA) is widely used for assessing the severity and progression of genetic cerebellar ataxias. SARA is now considered a primary end point in several ataxia treatment trials, but its underlying composite item measurement model has not yet been tested. This work aimed to evaluate the composite properties of SARA and its items using item response theory (IRT) and to demonstrate its applicability across even ultra-rare genetic ataxias. Leveraging SARA subscores data from 1932 visits from 990 patients of the Autosomal Recessive Cerebellar Ataxias (ARCA) registry, we assessed the performance of SARA using IRT methodology. The item characteristics were evaluated over the ataxia severity range of the entire ataxia population as well as the assessment validity across 115 genetic ARCA subpopulations. A unidimensional IRT model was able to describe SARA item data, indicating that SARA captures one single latent variable. All items had high discrimination values (1.5–2.9) indicating the effectiveness of the SARA in differentiating between subjects with different disease statuses. Each item contributed between 7% and 28% of the total assessment informativeness. There was no evidence for differences between the 115 genetic ARCA subpopulations in SARA applicability. These results show the good discrimination ability of SARA with all of its items adding informational value. The IRT framework provides a thorough description of SARA on the item level, and facilitates its utilization as a clinical outcome assessment in upcoming longitudinal natural history or treatment trials, across a large number of ataxias, including ultra-rare ones.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

*For ARCA study group authors, see Appendix.

Matthis Synofzik shared last authorship with Mats O. Karlsson.

The Scale for the Assessment and Rating of Ataxia (SARA) is the most widely used Clinical Outcome Assessment (COA) for assessing the severity and progression of cerebellar ataxias. Regulatory agencies and recent studies have raised concerns about its metric properties, especially at a single-item level. The item measurement model underlying the SARA has not yet been thoroughly tested.

WHAT QUESTION DID THIS STUDY ADDRESS?

Is SARA, and its individual items, a good performing COA, and is it applicable across rare and ultra-rare genetic ataxia subpopulations?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

Item response modeling showed that SARA items have good performance in assessing ataxia severity with high discrimination ability and informativeness. There was no evidence of differences between genetic subpopulations in terms of SARA applicability.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

The presented work provides evidence of SARA adequacy for use as a COA. The presented item response theory framework will facilitate the longitudinal and treatment effects analyses and guide trial designs in upcoming ataxia treatment trials.

INTRODUCTION

Cerebellar ataxias are a heterogeneous group of rare and ultra-rare neurodegenerative diseases resulting from progressive damage to the cerebellum and/or its associated tracts.[1,2] With manifold targeted treatment trials upcoming for an increasing number of genetic ataxias, an optimized and thoroughly tested clinical outcome assessment (COA) for ataxia has become a key for academia, industry, and regulatory agencies.[3] The Scale for the Assessment and Rating of Ataxia (SARA) is by far the most widely used COA for assessing the severity and trial sizes in the pleiotropy of cerebellar ataxias,[4,5] and is now also considered a primary end point in several ataxia randomized controlled trials.[6–11] This includes both autosomal-dominant and autosomal-recessive ataxias[12,13] comprising more than 200 genetically stratified ataxia subpopulations, many of them so ultra-rare that it impedes standard COA modeling methodology.[2,14]

While the validity and reliability of the SARA have been shown to be excellent across different ataxia populations,[4,15,16] regulatory agencies and recent studies in both autosomal-dominant and autosomal-recessive ataxias have raised concerns not only related to its functional relevance but also to its metric properties, especially at a single item level.[3,17,18] Thus, modifications to optimize the SARA are now being discussed, for example, omitting two of the appendicular items of SARA, which might improve its sensitivity to change and thus trial sizes in trial settings.[18] However, the underlying evidence and validation of such modifications have remained scarce.[3,18] In particular, the item measurement model underlying the SARA has not yet been thoroughly tested. In fact, the recent guidance of the FDA on fit-for-purpose COAs has highlighted that any COA should be analyzed and validated for its underlying item measurement model in order to be able to combine its multiple items into a single score on a COA. This includes modeling COAs using latent variable modeling approaches such as item response theory models.[19]

Item response theory (IRT) is a class of latent variable models used for the analysis of composite assessment data on the item level.[20,21] It quantifies the relationship between the probability of a particular response to an assessment's item and an unobserved latent variable (ie, disease severity). The IRT framework has been increasingly applied in developing, evaluating, and refining health outcome assessments, especially for neurological diseases such as Parkinson's disease,[22–24] Alzheimer's disease,[25,26] multiple sclerosis,[27] and amyotrophic lateral sclerosis.[28]

Leveraging extensive real-world data from a large-scale international ataxia registry, with prospective SARA assessments across many ataxia genotypes, we here (i) utilize the IRT methodology to evaluate the performance and the quality of SARA as a COA of ataxia and thoroughly assess its dimensionality, and (ii) extend the IRT framework to demonstrate SARA applicability across subpopulations even those ultra-rare ataxias for which a primary IRT modeling would inherently not be possible, but which are now also becoming of treatment interest.[2,29,30]

METHODS

Data

Datasets were taken from the Autosomal Recessive Cerebellar Ataxias (ARCA) registry, a large-scale prospective longitudinal multicenter disease registry aiming to facilitate trial readiness in the ARCA disease area by capturing international real-world (rather than only, eg, consortia- or trial-specific) data.[31] The ARCA registry is compliant with both General Data Protection Regulation (GDPR) and the European medicines agency standards for data quality. Informed consents were obtained before patients' inclusion.[31] Patients had been eligible for inclusion into the ARCA registry if they had (i) a genetically confirmed ARCA; and/or (ii) an early-onset ataxia (EOA) with onset before age 40 years without evidence of an autosomal-dominant family history, a polyglutamine repeat expansion in spinocerebellar ataxia genes, or acquired cause, thus representing a stratum of patients with ataxia known to be enriched—albeit not exclusive—for ARCAs.[14,32,33] The dataset used in this study comprises 990 patients with baseline visits. A total of 420 of the subjects had at least one follow-up visit, and up to 9 predominantly annual visits resulting in a total of 1932 visits. 69% of the patients have established genetic diagnoses with a total of 115 genetic ARCA subpopulations, while 31% have genetically undetermined diagnoses (for further details on the patients and disease characteristics of these diseases included and analyzed here, see[14]).

Ataxia disease severity in this dataset is assessed by the SARA.[4] The SARA is a clinician-reported outcome (ClinRO) consisting of eight task-based items: gait (score 0 to 8), stance (score 0 to 6), sitting (score 0 to 4), speech disturbance (score 0 to 6), finger chase (score 0 to 4), nose-finger test (score 0 to 4), fast alternating hand movements (score 0 to 4), and heel-shin slide (score 0 to 4). It results in a composite score ranging from zero in nonataxic state up to 40 in the most severe ataxia. The last four items are appendicular, that is, assessing (upper or lower) limb function, rated separately for the right and left sides of the body, and the means of right and left scores are calculated (for further details on the SARA details and metrics in the disease populations included and analyzed here, see[18]).

Item response theory modeling

IRT model

In an IRT model, the response to an item is modeled through item characteristic functions (ICFs), which describe the probability of a particular item response given by individual i with an underlying latent variable (ψ_i). The estimated ICFs can then be visualized using item characteristic curves (ICCs). In this work, graded response model with two item parameters was used to model each item (j):

$$P\left(Y_{ij} \geq k\right) = \frac{e^{a_j\left(\psi_i - b_{j,k}\right)}}{1 + e^{\left(a_j\left(\psi_i - b_{j,k}\right)\right)}} \quad (1)$$

$$P\left(Y_{ij} = k\right) = P\left(Y_{ij} \geq k\right) - P\left(Y_{ij} \geq k+1\right) \quad (2)$$

where Y_{ij} is the observed item score for individual i and item j, $P\left(Y_{ij} \geq k\right)$ is the probability of the ith individual reporting a score at or above k, and $P\left(Y_{ij} = k\right)$ is the probability of reporting score k. The model parameters a_j and $b_{j,k}$ are item-specific fixed-effect parameters; which describe the discrimination of item j and the difficulty parameter for a certain score (k) from item j, respectively. The discrimination parameter represents the slope of the ICC; $P\left(Y_{ij} = k\right)$ versus ψ_i curve, at the inflection point, while the difficulty parameter gives the disease severity level at which the probability of scoring k or higher is 50%. An item having a higher discrimination parameter is able to differentiate better between subjects with underlying latent variables close to the inflection point of the ICC. An item with smaller difficulty parameters is shifted to the left on the latent variable scale and hence is more difficult to complete correctly (individuals with less severe ataxia are probably failing to complete the task). The latent variable ψ_i is a subject-specific random effect and is modeled assuming a normal distribution with a mean of 0 (typical patient) and variance of 1 in the reference population. Note that the latent variable is assumed to be the same across all items for a certain SARA measurement. The ICFs were estimated by treating each SARA assessment as an independent individual to ensure broad coverage of disease severity levels.

The IRT model implementation was performed using the nonlinear mixed-effects modeling software NONMEM version 7.5.[34] Parameter estimation was conducted using a Laplace approximation to the likelihood. The NONMEM model code (Supplementary material 1: Appendix S1) and model diagnostics were generated using the R package piraid.[35]

Model assumptions checking

Upon applying IRT models, a set of assumptions are made that require careful assessment and consideration.[20,21] One important assumption made in this work is unidimensionality, meaning that the response probabilities depend only on one common factor, that is, a single latent

variable ψ_i. Implicit in the unidimensionality assumption is the assumption of no local dependence, which postulates that the response of each item is independent of those of other items.

In this study, a unidimensional IRT model was built to fit the ARCA data, and the adequacy of the unidimensionality assumption was thoroughly assessed in multiple ways. First, the Pearson correlation coefficients between responses from each pair of items (from each visit) were calculated, as well as the correlations between item response residuals for item pairs. The residuals (RES_{ij}) are calculated as shown in Equation 3.

$$RES_{ij} = \frac{Y_{ij} - E\left(\hat{Y}_{ij}\right)}{E\left(SD\left(\hat{Y}_{ij}\right)\right)} \quad (3)$$

where Y_{ij} is the observed score for individual i and item j, $E\left(\hat{Y}_{ij}\right)$ is the corresponding expected score based on the ICFs and individual latent variable estimates, and $E\left(\mathrm{SD}\left(\hat{Y}_{ij}\right)\right)$ is the expected standard deviation of the predicted scores.

Second, the adequacy of the unidimensional IRT model was confirmed by comparing the observed and model-predicted data in terms of data correlations and residual correlations for item pairs. Using the visual predictive check (VPC) approach,[36] 100 datasets were simulated from the developed IRT model with item-level data, and the quantiles of the simulated data were graphically compared against the corresponding quantiles of observed data. This comparison was performed for both item scores and response residuals.

Fisher Item Information

To evaluate the amount of information provided by each item of the SARA assessment with respect to the underlying latent variable, the Fisher information for each item j was calculated (and visualized) as a function of ψ_i; $I_j(\psi_i)$, as shown in Equation 4.

$$I_j(\psi_i) = -E\left(\frac{d^2 \log L}{d{\psi_i}^2}\right) \quad (4)$$

where E is the expectation of the second derivative of log-likelihood (log$L$); that is, the likelihood of specific parameter values given the observed data. The total item information in the population (population item information; I_j) was calculated as the integration of item information over the latent variable distribution.

ARCA genetic subpopulation analysis

Statistical analysis was conducted on the level of the genetic ARCA subpopulations[2,14] to (i) evaluate the applicability of the developed IRT model to all genetic subpopulations present in the dataset and (ii) assess whether the known population heterogeneity[2,14] manifests as different patterns at the item level of SARA.

In this analysis, the likelihood values (L), resulting from the maximum-likelihood estimation for the IRT model fitted to the whole ARCA dataset, were compared between each of the genetic subpopulations and the entire ARCA population. In practice, an objective function value (OFV) is used as a way to measure the goodness of fit for a model and is calculated as $-2\log(L)$, and the OFV contribution for each individual is reported; iOFV. The difference in the mean of iOFVs for each subpopulation and the mean of iOFVs for the entire population; Δ mean(iOFV) was calculated along with the 95% confidence interval CI (Equations 5–7). The CI was calculated based on pooled two-sampled t-test assuming equal variances between the studied subpopulation and the entire population. A correction factor was used to adjust the test statistic for the overlap between the subpopulation and the entire population.[37] Bonferroni correction was made to the 0.05 confidence level to adjust for the multiple comparisons performed on the same dataset.

$$\Delta\ \mathrm{mean(iOFV)} = \mathrm{mean}\left(\mathrm{iOFV_{subpop}}\right) - \mathrm{mean}\left(\mathrm{iOFV_{all\ pop}}\right) \quad (5)$$

$$\mathrm{SD}_p^2 = \frac{\left(n_{\mathrm{sub}} - 1\right) * \mathrm{SD}_{\mathrm{sub}}^2 + \left(n_{\mathrm{all}} - 1\right) * \mathrm{SD}_{\mathrm{all}}^2}{n_{\mathrm{sub}} + n_{\mathrm{all}} - 2} \quad (6)$$

$$\mathrm{CI} = \Delta\ \mathrm{mean(iOFV)} \pm t^* \sqrt{\mathrm{SD}_p^2 * \left(\frac{1}{n_{\mathrm{sub}}} + \frac{1}{n_{\mathrm{all}}}\right)} * \sqrt{\frac{\left(1 - \frac{n_{\mathrm{sub}}}{n_{\mathrm{all}}}\right)}{\left(1 + \frac{n_{\mathrm{sub}}}{n_{\mathrm{all}}}\right)}} \quad (7)$$

where SD_p^2 is the weighted average of the two sample variations, $\mathrm{SD}_{\mathrm{sub}}^2$ and $\mathrm{SD}_{\mathrm{all}}^2$ are the variances for the subpopulation and the entire population, respectively, n_{sub} and n_{all} are the numbers of observations for the corresponding groups, t^* is the t-critical value based on the confidence level, and $(n_{\mathrm{sub}} + n_{\mathrm{all}} - 2)$ degrees of freedom. The term $\sqrt{\left(1 - \frac{n_{\mathrm{sub}}}{n_{\mathrm{all}}}\right) / \left(1 + \frac{n_{\mathrm{sub}}}{n_{\mathrm{all}}}\right)}$ is the correction factor calculated based on the size proportion of the subpopulation relative to the entire ARCA population.

To avoid bias in the calculations due to the variation in the number of study visits between subjects, only one observation per subject was selected randomly, with equal

probability from the set of available measurements of a subject, and considered in the analysis. In order to illustrate how a group of individuals with different SARA item performances would perform in this analysis, a hypothetical subpopulation was created by permuting the individual responses for each item across the individuals in the original dataset. This hypothetical population serves as a "positive control group" in this analysis.

RESULTS

Item response theory modeling

IRT model

An initial assessment we performed on SARA showed a distinctively high correlation between gait and stance items ($r=0.9$) compared to the other item–pair correlations ($r=0.36$–0.69), which indicated a local dependency between gait and stance items. Local dependency means that one item's response might affect the probability of a response of another item and failure to account for this characteristic may lead to inaccurate estimation of item parameters and the individual disease severity. To handle this local dependency during our unidimensional IRT analysis, the two interdependent items were combined into a single item (gait-stance) by taking the sum of their subscores, resulting in a 15-category (0–14) item. As a consequence, the single-item characteristics for both gait and stance will not be differentiated from each other in all downstream analyses. This adjustment was performed on the data upon modeling only and not on the SARA itself. Similarly, the correlations between the right and left responses for each of the appendicular items were particularly high (0.85–0.92) so the subscores of both sides were combined into a single value (mean).

The final IRT model consisted of seven ordered categorical submodels with a total of 63 item-specific parameters including item discrimination and difficulty parameters for each item response category. All item parameters were successfully estimated (ie, computed to give the best fit of the model to the observed data) with good precision (Table S1).

Item characteristic curves

The ICCs for each item of the SARA are shown in Figure 1 as the probability of a certain item score as a function of the latent variable (ataxia severity). The goodness of fit of the ICCs was evaluated using a generalized additive model smooth diagnostic,[38] showing that the ICCs describe the data adequately (Figure S1).

The ICCs show an increase in the probability of obtaining higher subscores upon increasing the individual's latent variable (ψ_i). The location of the ICC on the ψ_i scale (x-axis) indicates the difficulty of the corresponding item category. Only little overlap between the ICCs of each SARA item is observed as illustrated by the range of ψ_i where a particular item response has higher probability than the other response categories. For the appendicular items, the curves corresponding to the half-point scores—resulting from calculating the mean—show small peaks reflecting low probabilities of reporting different subscores for the right and left body sides. This implies that patients tend to perform similarly regardless of the measured body side which is in line with the observed data in the ARCA registry dataset where 76%–83% of the study visits have the same reported subscores for both body sides in the four appendicular items.

Regarding the shape of the ICCs, items having nearly flat ICCs over the full ψ_i range are not differentiating well between individuals with different ataxia severities. These items are associated with small discrimination parameters. On the contrary, the ICCs of SARA from our analysis are rather (i) sigmoidal in case of the lowest and highest response categories in each item or (ii) bell shaped for the middle response categories. These ICCs indicate the ability of SARA, for all of its items and across its full range, to detect a change in the item response upon changing the individual latent variable at the non-flat regions of the curves. The estimated discrimination parameters vary between 1.5 and 2.9 with a mean of 2.11 (Table S1) which are considered high when, for example, compared with the discrimination values of 0.02–0.3 resulting from fitting the model using a dataset with permuted subscores.

To allow for direct mapping between the latent variable scale and SARA total score, IRT-informed link functions were calculated based on the ICFs of the developed IRT model[39] and visualized in Figure S3.

Model assumptions checking

After handling the observed local dependencies as described earlier, all item pairs have shown similarly high data-correlation levels, implying that all SARA items share a common latent variable (Figure 2). The correlations between model residuals of item pairs are slightly negative and did not show any distinct patterns across items (Figure 3). This indicates that model misspecifications are absent and that the unidimensional IRT model has accounted for the data correlations between SARA items. Model simulations were able to approximately mimic the real-data patterns, as illustrated by both the average correlations of the 100 simulated datasets, and

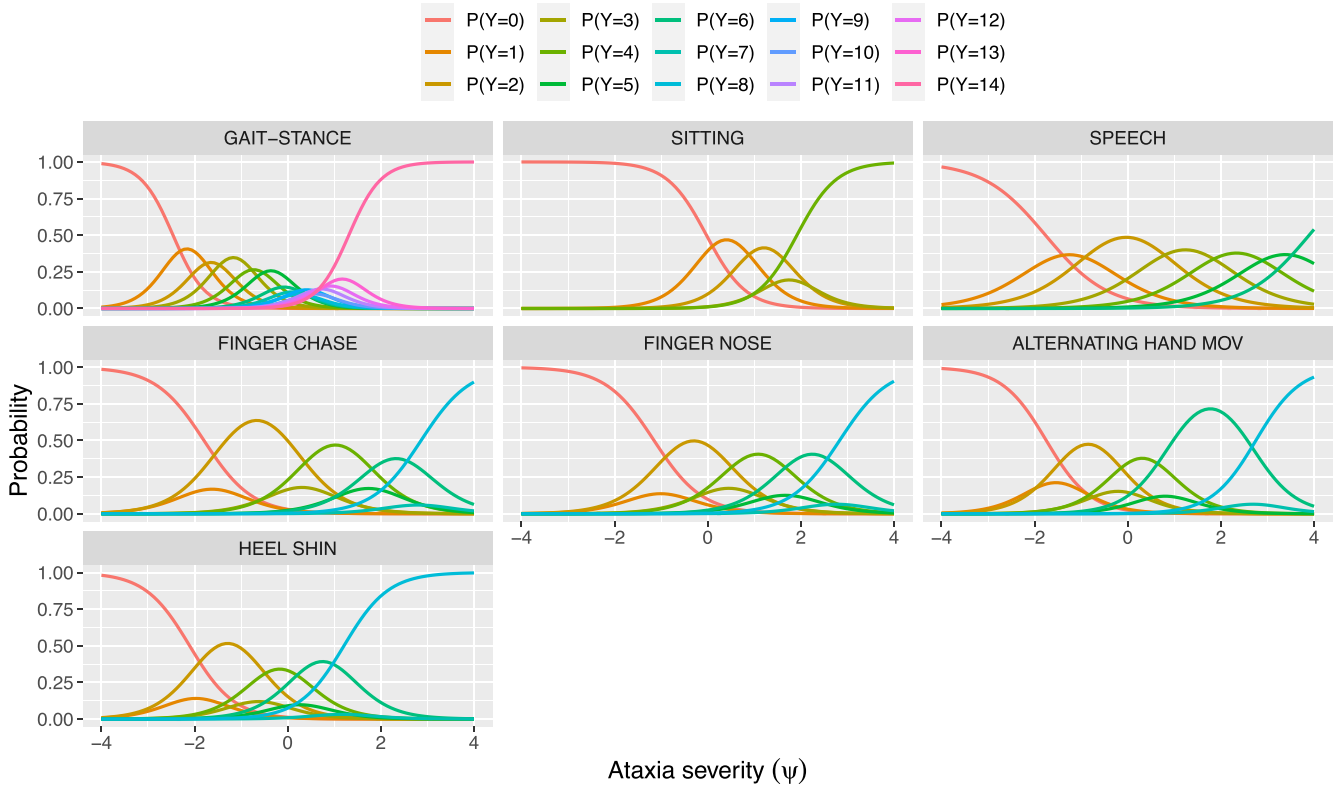

FIGURE 1 Item characteristic curves for the different items of the SARA scale, describing the probability of occurrence of each score Y for a patient with a given ataxia severity level. Ataxia severity represents a Z-score relative to the typical individual in the reference population, that is, the entire ARCA cohort.

the 95% confidence intervals of the 5th, 50th, and 95th percentiles of the simulated data responses and residuals (Figures 2 and 3), further indicating the lack of model misspecifications. Further studies show that the unidimensionality assumption is validated using exploratory and confirmatory factor analysis (Supplementary Material 5: Appendix S1).

Fisher item information

The Fisher information of the different SARA items as a function of ataxia severity is illustrated in Figure 4 along with the total item information in the population (I_j). In general, an item with high Fisher information will be able to determine a patient's disease severity precisely. The figure shows that the item information curves differ in both the amplitude of the peak and its location on the latent variable scale. The combined gait-stance item has the highest amplitude of information and the highest total population information (2.58), suggesting a higher informativeness and hence, higher sensitivity to changes in ataxia severity compared to other items (accompanied also by a higher discrimination value; 2.90). However, other items, for example, finger chase, finger nose, and sitting items, have a similar or higher (~1–1.5) information amount compared to gait-stance (~1) at more extreme ataxia severity levels ($\psi_i = 2$), thus enabling the differentiation between patients with higher ataxia severity. Generally, all SARA items add informational value with varying levels across the latent variable scale. Further analysis shows that the total population information of SARA decreases upon dropping different items from the assessment (eg, decrease by 5.5% and 7.9% when dropping gait and finger chase items, respectively) (Supplementary Material 6: Appendix S1).

Analysis of genetic ARCA subpopulations

The forest plot in Figure 5 depicts the analysis results of 29 different genetic ARCA subpopulations (with three or more subjects) and 2 pooled groups of subpopulations (subpopulations with one subject ($n = 69$) and subjects with genetically undetermined diagnoses ($n = 304$)), that is, even including ultra-rare ARCA subpopulations where a primary IRT analysis would not be possible, per se. The plot shows the mean difference between the individual

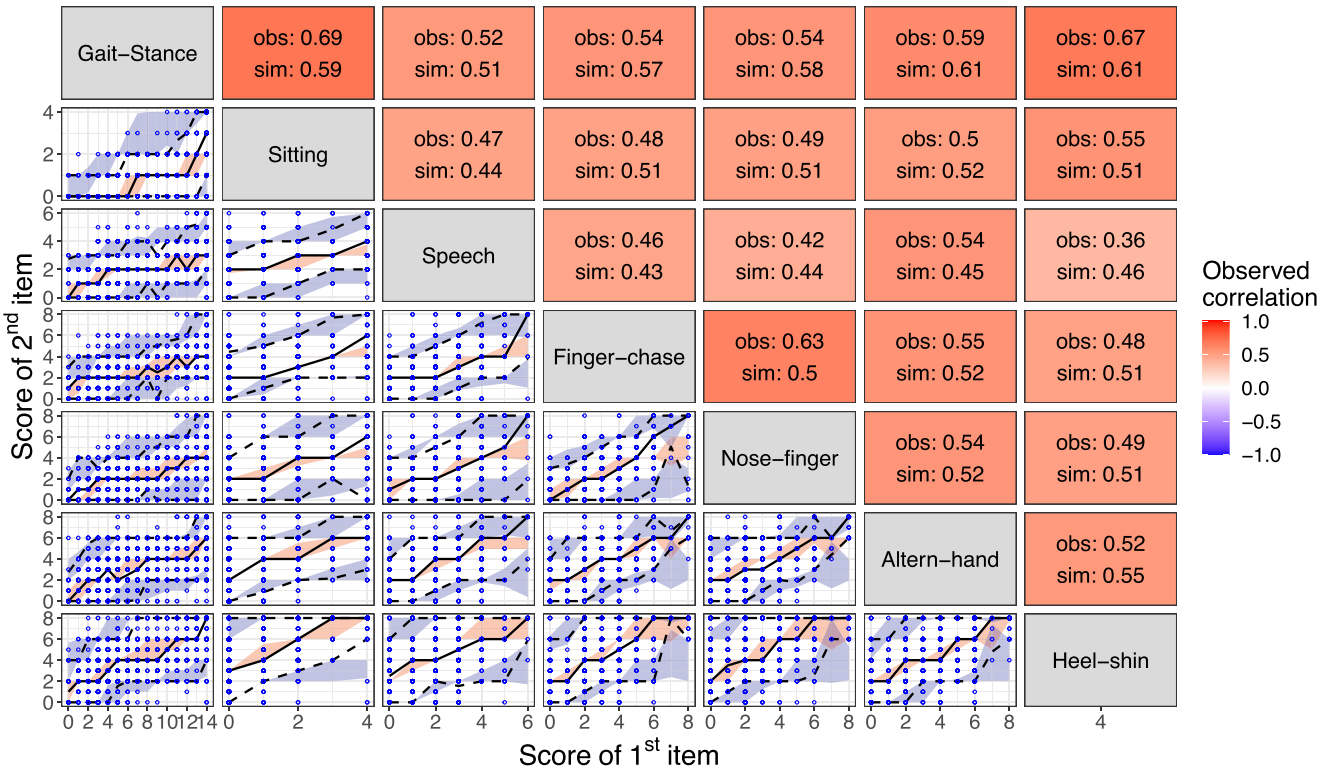

FIGURE 2 Data correlations between different item pairs (referred to as the first and second item) of the SARA scale for both original ARCA dataset and 100 simulated datasets. Upper matrix: item-pairs correlations for the observed dataset (with value-based color scaling) and item-pairs average correlations for the simulated datasets. Lower matrix: The black lines represent the median (solid lines) and the 5th and 95th percentiles (dashed lines) of the observed item scores (blue circles). The shaded areas represent the 95% confidence intervals of the corresponding percentiles for the simulated data (blue: upper and lower percentiles, pink: median percentile).

likelihood estimates of each genetic ARCA subpopulation and the entire ARCA population along with the 95% confidence interval. All confidence intervals of ARCA subpopulations (including those with two subjects ($n=17$) not shown in the figure) encompass zero, indicating the absence of evidence for any IRT model misfits on the subpopulation level. On the contrary, the confidence interval of the permuted group, the positive control group in this case, shows a worse fit than the original population; as one would anticipate since the relationship between the patient response to the different items and the underlying disease status has been altered upon randomly permuting the individual item scores among individuals. This permuted group behavior illustrates how an ARCA subpopulation with different SARA item patterns would perform in this analysis.

These results are supported by further analysis using one-way analysis of variance (ANOVA) with a null hypothesis stating that there is no difference in the mean individual likelihood estimates (iOFVs) across genetic subpopulations with $N \geq 3$ and with a significance level of 0.05. One-way ANOVA revealed that there was no statistically significant difference in the iOFVs means between the different subpopulations (p-value $=0.22$).

DISCUSSION

While the Scale for the Assessment and Rating of Ataxia (SARA) is the most widely used COA for assessing the severity and progression of cerebellar ataxias and is now considered a primary end point in several ataxia treatment trials, its underlying composite item measurement model that has not yet been tested. This is, however, needed to qualify it as a fit-for-purpose COA, as now also highlighted by the FDA.[19] Leveraging a large ataxia patient dataset, an IRT analysis was successfully implemented for the first time on SARA assessment to evaluate its metric properties and performance. This type of latent variable modeling is based on the reflective indicator model described in the FDA's guidance.[19] In this work, the adequacy of the SARA was tested based on multiple aspects including (i) SARA items' characteristics and Fisher information, (ii) the reasonableness of the unidimensionality assumption by showing that all items reflect a single latent variable which can be described as the relevant concept of interest (ie, ataxia severity), thus also allowing to combine responses from multiple items into a single score (latent variable) in the IRT model, and (iii) the SARA applicability across genetic ARCA subpopulations, including even ultra-rare ones.

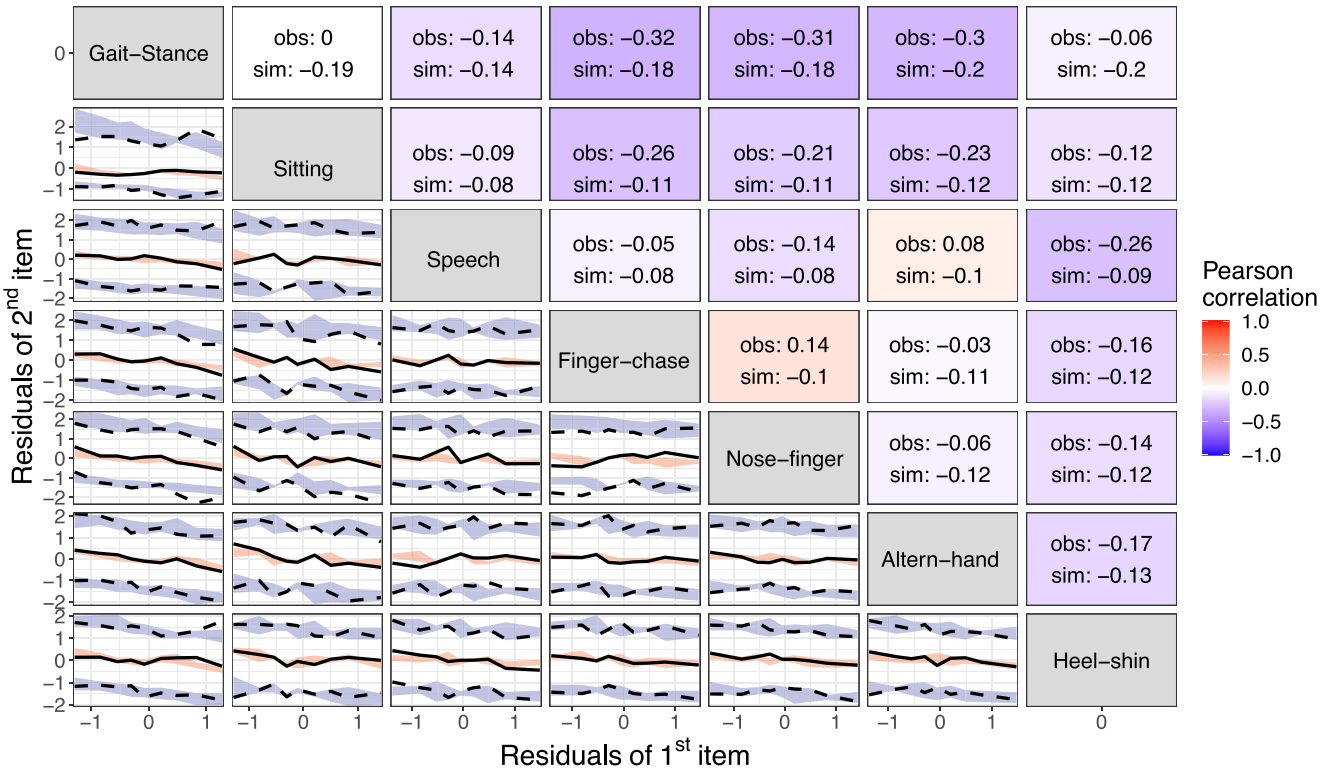

FIGURE 3 Residual correlations between different item pairs (referred to as the 1st and 2nd item) of the SARA scale for both original ARCA dataset and 100 simulated datasets. Upper matrix: item-pairs residual correlations for the observed dataset (with value-based color scaling) and item-pairs average residual correlations for the simulated datasets. Lower matrix: The black lines represent the median (solid lines) and the 5th and 95th percentiles (dashed lines) of the observed item scores. The shaded areas represent the 95% confidence intervals of the corresponding percentiles for the simulated data (blue: upper and lower, pink: median).

Composite clinical outcome assessments are traditionally analyzed as single aggregated total scores with disregard to the assessment complexity and its individual components. Multiple drawbacks are associated with the use of such analytical approach, as (i) it does not incorporate the differential informativeness of different items, (ii) it ignores the differential contribution of subscores to the total score, and (iii) it disregards the discrete and bounded nature of the scale. On the contrary, the IRT methodology acknowledges the interaction of the studied individuals and the assessment components by relating items' responses probabilistically to an underlying unobserved variable (disease severity). The resulting item-level information can be utilized in evaluating and refining assessments[21]—as now discussed for the SARA.[3] Furthermore, IRT models can be used to describe the change in disease severity over time in longitudinal assessment score data. Recent publications have shown that longitudinal IRT models result in higher power to detect treatment effects compared to total score models.[22,25,27] This is of timely need for the SARA where, while appreciating it as a key end point, several studies have shown that its power to detect treatment effects by standard analysis is too low for most trials in real-world ataxia populations.[18,40]

We developed a unidimensional IRT model leveraging all available data from the individual items using a graded response model with two parameters. The model was able to characterize SARA items including the differential discriminative ability of different items, which would not have been possible with a more parsimonious one-parameter model. While an additional "guess" parameter can account for the probability of getting the item correct by guessing alone, a sensitivity analysis has indicated the absence of such characteristics in SARA items (data not shown).

The resulting ICCs showed that SARA is able to detect changes in item responses upon changing ataxia severity. This is further corroborated when comparing the IRT characteristics of SARA to previously published IRT models for other neurological diseases COAs, for example, MDS-UPDRS, ADAS-Cog, and EDSS, where SARA shows a good discrimination ability, with mean discrimination values of about 2.7-, 1.2-, and 1.3-folds higher than the aforementioned COAs, respectively.[23,25,27] The difficulty parameters for single categories were both (i) ordered so the difference between adjacent categories reflects a change in the latent variable, and (ii) sufficiently spaced with little overlap which indicates that the

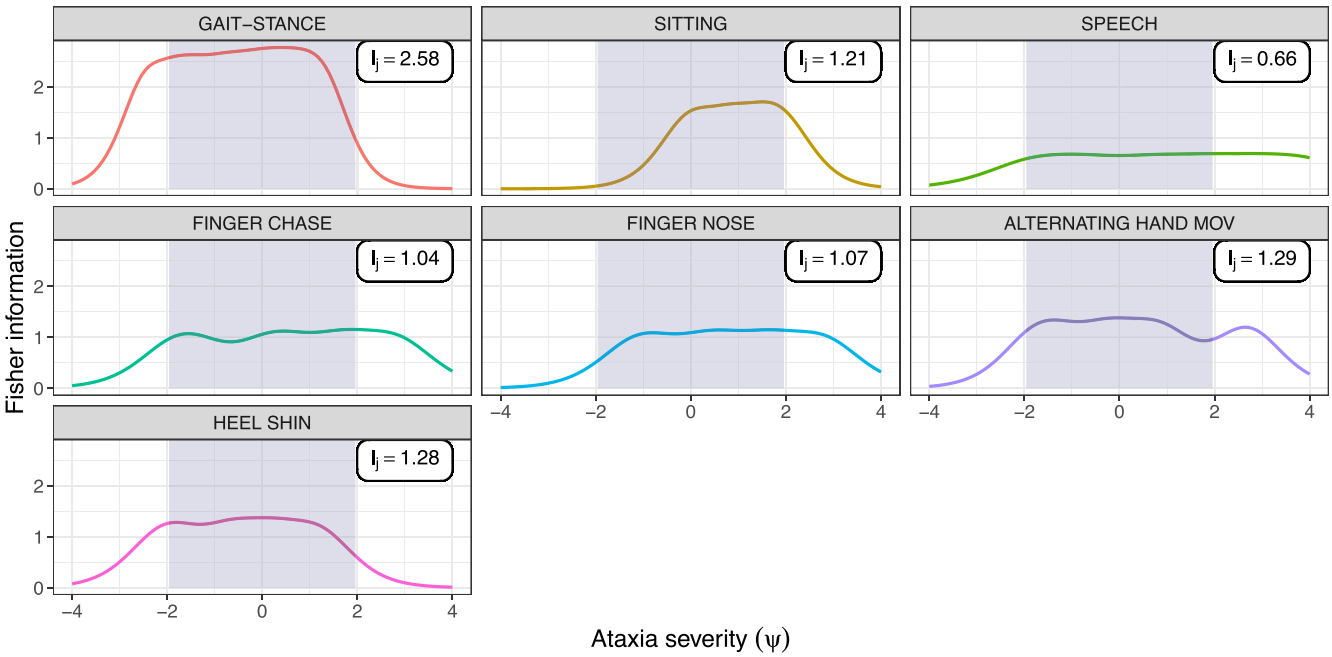

FIGURE 4 Item information curves for the different items in the SARA scale as a function of ataxia disease severity (ψ). The shaded areas indicate the ataxia severity interval for 95% of the studied population. The reported I_j is the total population information for each item; that is, the integration of item information over the full ψ range.

individual categories of SARA items were well designed allowing clinicians to properly rate the patient's performance in each task. The adequacy of SARA was also supported by the high Fisher information values of all items, 1.14 mean population item information compared to the mean of 0.8 for EDSS assessment,[27] and hence their high sensitivity to the disease severity changes. The varying importance of items across the full latent variable range emphasizes the contribution of all SARA items in capturing the underlying disease status. Such positive contribution of a SARA item based on IRT analysis is not necessarily associated with similar behavior in a classical total score analysis, as shown in previously published studies, which demonstrated that some SARA items (in particular, some appendicular items) perform poorly in total score analysis.[17,18] This discrepancy could be attributed to the differential informativeness of the SARA items that is recognized in IRT analysis.

The analysis of genetic ARCA subpopulations suggests that the ARCA population heterogeneity[2,14] is not manifesting as different item patterns in SARA assessment of a magnitude that is discernable in the present ARCA dataset. Unexpected behavior of items would result if the ICFs have differed across subpopulations; for example, if a good scoring in gait-stance item is associated with a bad scoring in speech item at certain latent variable levels. Such altered relationships were indeed observed in the group with permuted item scores, which here served as a "positive control group". Yet no difference between any of the ARCA subpopulations and the entire ARCA population was observed, as indicated by the 95% confidence interval of the mean difference in iOFV in the forest plot. This indicates the applicability of one IRT model across all different genetic ARCA subpopulations, that is, even including ultra-rare ARCA subpopulations where such a primary IRT analysis would not be possible due to small amount of data. This indicates that the SARA is measuring the same construct across disease subpopulations.

Potential limitations of the present studies—accompanied by limitations related to properties of the SARA in some cases—should be noted. First, the gait and stance items could not be individually characterized in the IRT model since we combined both into one item to handle the high correlation between them. One could argue that multiple latent variables could be included in the model, however, this was not possible due to the small number of items in the assessment. Alternatively, reducing the SARA by removing either gait or stance (or measuring the appendicular items on only one body side) could also be suggested. Nonetheless, at least for time reasons, such reduction is unnecessary since the SARA assessment has an overall few items, and takes only, on average, 15 min to be administered.[4] Furthermore, additional findings from our study show that removing one of the items, gait and stance, leads to at least slight reduction in the total Fisher information in the population (Table S4). Second, a normal distribution for the underlying latent variable was assumed in our model. The

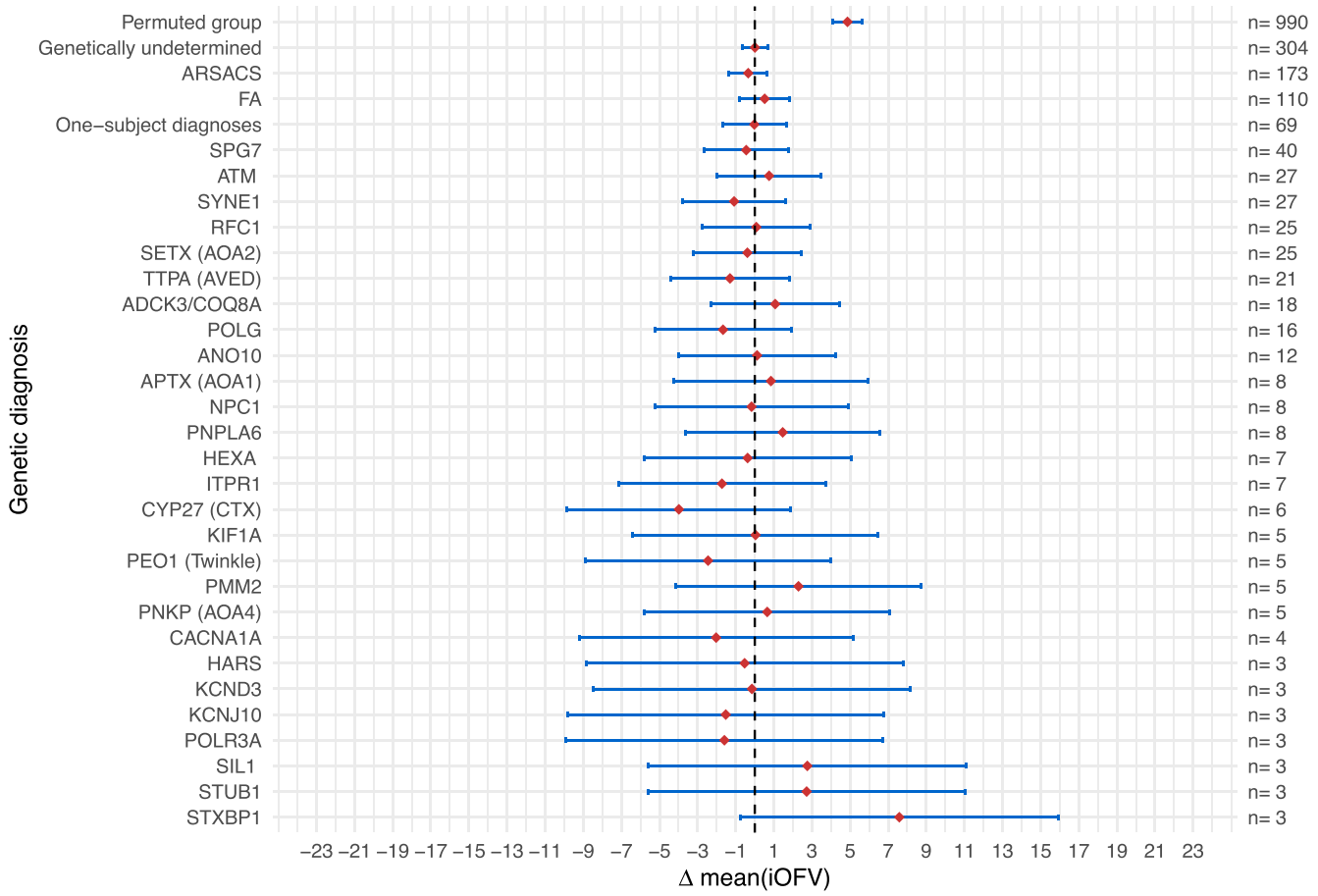

FIGURE 5 Difference in means of iOFVs (red points); $\text{mean}(\text{iOFV}_{\text{subpop}}) - \text{mean}(\text{iOFV}_{\text{all pop}})$, between different genetic ARCA subpopulations and the entire ARCA population, along with the 95% confidence intervals (error bars). *n* is the number of subjects in each subpopulation group. One-subject diagnoses is a pool of diagnoses comprising only one subject. Genetically undetermined: the pool of patients with genetically yet unidentified diagnoses, despite extensive genetic work-up. Permuted group is the hypothetical population where the item subscores of the original dataset were permuted among individuals. The results of subpopulations with $n \leq 2$ are not illustrated in the figure.

latent variable scale is a hypothetical construct with assumed distribution that will be dependent on the choices of item model. A normal distribution has been assumed in all IRT models of other COAs that we have encountered.[22–27,38,41] Nevertheless, if the normality assumption is inappropriate, it will result in model misspecification. In the assessment of model fit, no such misspecification was evident (Figures 2 and 3, Supplementary Material 7: Appendix S1). Another potential limitation is that data were treated as cross-sectional upon estimating the ICFs, as per the standard practices, which ignores the potential correlation between repeated observations of a certain subject. However, such phenomenon does not necessarily affect the estimated ICFs. Still, longitudinal analysis and validation of the IRT characteristics of the SARA are warranted, which will be addressed in future research. The small sample set in the ultra-rare ARCA subpopulations hinders the ability to implement IRT models for single ARCA subpopulations due to the lack of observed responses in multiple-item response categories (Figure S2). While such small sample sizes are inherent in these ultra-rare ARCAs, future research is needed to develop improved study designs for detecting treatment effects in small populations and rare diseases.

Overall, our results demonstrate a good performance of SARA assessment based on the item-level analysis using IRT. The developed IRT framework will facilitate the assessment of disease progression and treatment effects and guide trial designs and sample size calculations, thus improving trial designs in upcoming ataxia treatment trials in rare genetic ataxia populations.

AUTHOR CONTRIBUTIONS

All authors wrote the manuscript and designed the research. A.H. performed the research and analyzed the data.

ACKNOWLEDGMENTS

Members of the Evidence-RND consortium include Emmanuelle Comets, Niels Hendrickx, Nicole Maria Heussen, Ralf-Dieter Hilgers, Thomas Klockgether, France Mentré, Yevgen Ryeznik, Oleksandr Sverdlov.

FUNDING INFORMATION

This work was funded by the European Joint Programme on Rare Diseases (EJP RD) Joint Transnational Call 2019 for the EJP RD WP20 Innovation Statistics consortium "EVIDENCE-RND" focusing on "Innovative Statistical Methodologies to Improve Rare Diseases Clinical Trials in Limited Populations" under the EJP RD Grant Agreement (n°825575) (to M.K, R.S., and M.S.). Moreover, work in this project was supported by the Clinician Scientist program "PRECISE.net" funded by the Else Kröner-Fresenius-Stiftung (to M.S., R.S., and A.T). This work was also financially supported by the Swedish Research Council Grant 2018-03317 (to M.O.K.).

CONFLICT OF INTEREST STATEMENT

Dr. Klockgether is receiving research support from the Bundesministerium für Bildung und Forschung (BMBF), the National Institutes of Health (NIH), and Servier. Within the last 24 months, he has received consulting fees from Biogen, UCB, and Vico Therapeutics, all unrelated to the present manuscript. Dr. Synofzik has received consultancy honoraria from Ionis, UCB, Prevail, Orphazyme, Servier, Reata, GenOrph, AviadoBio, Biohaven, Zevra, and Lilly, all unrelated to the present manuscript. Drs Hooker and Karlsson have received consultancy fees from and own stock in Pharmetheus, all unrelated to this manuscript. As an Associate Editor *for CPT: Pharmacometrics and Systems Pharmacology*, Andrew Hooker was not involved in the review or decision process for this paper. All other authors declared no competing interests in this work.

DATA AVAILABILITY STATEMENT

Data supporting the study findings are available on request from the corresponding author.

ORCID

Alzahra Hamdan https://orcid.org/0009-0002-3102-2707
Andrew C. Hooker https://orcid.org/0000-0002-2676-5912
Xiaomei Chen https://orcid.org/0009-0002-0663-0532
Andreas Traschütz https://orcid.org/0000-0002-8165-5898
Rebecca Schüle https://orcid.org/0000-0002-7781-2766
Matthis Synofzik https://orcid.org/0000-0002-2280-7273
Mats O. Karlsson https://orcid.org/0000-0003-1258-8297

REFERENCES

1. Anheim M, Tranchant C, Koenig M. The autosomal recessive cerebellar ataxias. *N Engl J Med*. 2012;366(7):636-646.
2. Synofzik M, Puccio H, Mochel F, Schöls L. Autosomal recessive cerebellar ataxias: paving the way toward targeted molecular therapies. *Neuron*. 2019;101(4):560-583.
3. Klockgether T, Ashizawa T, Brais B, et al. Paving the way toward meaningful trials in ataxias: an ataxia global initiative perspective. *Mov Disord*. 2022;37(6):1125-1130.
4. Schmitz-Hübsch T, du Montcel ST, Baliko L, et al. Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology*. 2006;66(11):1717-1720.
5. Perez-Lloret S, van de Warrenburg B, Rossi M, et al. Assessment of ataxia rating scales and cerebellar functional tests: critique and recommendations. *Mov Disord*. 2021;36(2):283-297.
6. Feil K, Adrion C, Boesch S, et al. Safety and efficacy of acetyl-DL-leucine in certain types of cerebellar ataxia: the ALCAT randomized clinical crossover trial. *JAMA Netw Open*. 2021;4(12):e2135841.
7. Lei LF, Yang GP, Wang JL, et al. Safety and efficacy of valproic acid treatment in SCA3/MJD patients. *Parkinsonism Relat Disord*. 2016;26:55-61.
8. França C, de Andrade DC, Silva V, et al. Effects of cerebellar transcranial magnetic stimulation on ataxias: a randomized trial. *Parkinsonism Relat Disord*. 2020;80:1-6.
9. Coarelli G, Heinzmann A, Ewenczyk C, et al. Safety and efficacy of riluzole in spinocerebellar ataxia type 2 in France (ATRIL): a multicentre, randomised, double-blind, placebo-controlled trial. *Lancet Neurol*. 2022;21(3):225-233.
10. Benussi A, Dell'Era V, Cantoni V, et al. Cerebello-spinal tDCS in ataxia: a randomized, double-blind, sham-controlled, crossover trial. *Neurology*. 2018;91(12):e1090-e1101.
11. Fields T, Bremova TM, Billington I, et al. N-acetyl-L-leucine for Niemann-pick type C: a multinational double-blind randomized placebo-controlled crossover study. *Trials*. 2023;24(1):361.
12. Jacobi H, du Montcel ST, Bauer P, et al. Long-term disease progression in spinocerebellar ataxia types 1, 2, 3, and 6: a longitudinal cohort study. *Lancet Neurol*. 2015;14(11):1101-1108.
13. Reetz K, Dogan I, Hilgers RD, et al. Progression characteristics of the European Friedreich's ataxia consortium for translational studies (EFACTS): a 4-year cohort study. *Lancet Neurol*. 2021;20(5):362-372.
14. Traschütz A, Adarmes-Gomez AD, Anheim M, et al. Autosomal recessive cerebellar ataxias in Europe: frequency, onset, and severity in 677 patients. *Mov Disord*. 2023;38(6):1109-1112.
15. Weyer A, Abele M, Schmitz-Hübsch T, et al. Reliability and validity of the scale for the assessment and rating of ataxia: a study in 64 ataxia patients. *Mov Disord*. 2007;22(11):1633-1637.
16. Bürk K, Mälzig U, Wolf S, et al. Comparison of three clinical rating scales in Friedreich ataxia (FRDA). *Mov Disord*. 2009;24(12):1779-1784.
17. Maas RPPWM, Teerenstra S, Lima M, et al. Differential temporal dynamics of axial and appendicular ataxia in SCA3. *Mov Disord*. 2022;37(9):1850-1860.
18. Traschütz A, Adarmes-Gómez AD, Anheim M, et al. Responsiveness of the scale for the assessment and rating of

ataxia and natural history in 884 recessive and early onset ataxia patients. *Ann Neurol*. 2023;94(3):470-485.
19. US Food and Drug Administration, U.S. Food and Drug Administration, FDA. Patient-focused drug development: selecting, developing, or modifying fit-for-purpose clinical outcome assessments. [cited 2023 Feb 2]. 2022 Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-selecting-developing-or-modifying-fit-purpose-clinical-outcome
20. Ueckert S. Modeling composite assessment data using item response theory. *CPT Pharmacometrics Syst Pharmacol*. 2018;7(4):205-218.
21. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res*. 2007;16:5-18.
22. Buatois S, Retout S, Frey N, Ueckert S. Item response theory as an efficient tool to describe a heterogeneous clinical rating scale in De novo idiopathic Parkinson's disease patients. *Pharm Res*. 2017;34(10):2109-2118.
23. Gottipati G, Karlsson MO, Plan EL. Modeling a composite score in Parkinson's disease using item response theory. *AAPS J*. 2017;19(3):837-845.
24. Arrington L, Ueckert S, Ahamadi M, Macha S, Karlsson MO. Performance of longitudinal item response theory models in shortened or partial assessments. *J Pharmacokinet Pharmacodyn*. 2020;47(5):461-471.
25. Ueckert S, Plan EL, Ito K, et al. Improved utilization of ADAS-cog assessment data through item response theory based Pharmacometric modeling. *Pharm Res*. 2014;31(8):2152-2165.
26. Balsis S, Unger AA, Benge JF, Geraci L, Doody RS. Gaining precision on the Alzheimer's disease assessment scale-cognitive: a comparison of item response theory-based scores and total scores. *Alzheimers Dement*. 2012;8(4):288-294.
27. Novakovic AM, Krekels EHJ, Munafo A, Ueckert S, Karlsson MO. Application of item response theory to modeling of expanded disability status scale in multiple sclerosis. *AAPS J*. 2017;19(1):172-179.
28. Bacci ED, Staniewska D, Coyne KS, et al. Item response theory analysis of the amyotrophic lateral sclerosis functional rating scale-revised in the pooled resource open-access ALS clinical trials database. *Amyotroph Lateral Scler Frontotemporal Degener*. 2016;17(3–4):157-167.
29. Synofzik M, van Roon-Mom WMC, Marckmann G, et al. Preparing n-of-1 antisense oligonucleotide treatments for rare neurological diseases in Europe: genetic, regulatory, and ethical perspectives. *Nucleic Acid Ther*. 2022;32(2):83-94.
30. Traschütz A, Schirinzi T, Laugwitz L, et al. Clinico-genetic, imaging and molecular delineation of COQ8A-ataxia: a multicenter study of 59 patients. *Ann Neurol*. 2020;88(2):251-263.
31. Traschütz A, Reich S, Adarmes AD, et al. The ARCA registry: a collaborative global platform for advancing trial readiness in autosomal recessive cerebellar ataxias. *Front Neurol*. 2021;12:677551.
32. Coutelier M, Hammer MB, Stevanin G, et al. Efficacy of exome-targeted capture sequencing to detect mutations in known cerebellar ataxia genes. *JAMA Neurol*. 2018;75(5):591-599.
33. Gama MTD, Braga-Neto P, Rangel DM, et al. Autosomal recessive cerebellar ataxias in South America: a multicenter study of 1338 patients. *Mov Disord*. 2022;37(8):1773-1774.
34. Beal SL, Sheiner LB, Boeckmann A, Bauer RJ. *NONMEM user's Guides (1989–2009)*. Icon Development Solutions; 2009.
35. Nordgren R, Ueckert S. Uppsala University, Pharmacometrics Research Group. 2022 [cited 2023 Jan 3]. Available from: https://github.com/UUPharmacometrics/piraid
36. Bergstrand M, Hooker AC, Wallin JE, Karlsson MO. Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. *AAPS J*. 2011;13(2):143-151.
37. Hayes LJ, Berry G. Comparing the part with the whole: should overlap be ignored in public health measures? *J Public Health*. 2006;28(3):278-282.
38. Lyauk YK, Jonker DM, Lund TM, Hooker AC, Karlsson MO. Item response theory modeling of the international prostate symptom score in patients with lower urinary tract symptoms associated with benign prostatic hyperplasia. *AAPS J*. 2020;22(5):115.
39. Wellhagen GJ, Ueckert S, Kjellsson MC, Karlsson MO. An item response theory-informed strategy to model Total score data from composite scales. *AAPS J*. 2021;23(3):45.
40. Ilg W, Müller B, Faber J, et al. Digital gait biomarkers allow to capture 1-year longitudinal change in spinocerebellar ataxia type 3. *Mov Disord*. 2022;37(11):2295-2301.
41. Haem E, Doostfatemeh M, Firouzabadi N, Ghazanfari N, Karlsson MO. A longitudinal item response model for aberrant behavior checklist (ABC) data from children with autism. *J Pharmacokinet Pharmacodyn*. 2020;47(3):241-253.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hamdan A, Hooker AC, Chen X, et al. Item performance of the scale for the assessment and rating of ataxia in rare and ultra-rare genetic ataxias. *CPT Pharmacometrics Syst Pharmacol*. 2024;13:1327-1340. doi:10.1002/psp4.13162

APPENDIX

Study group contributions of the ARCA study group:

Name, Title	Affiliation(s)	Contact
Astrid D. Adarmes-Gomez, MD	1. Unidad de Trastornos del Movimiento, Servicio de Neurología y Neurofisiología Clínica, Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville, Spain 2. Centro de Investigación Biomédica en Red Sobre Enfermedades Neurodegenerativas (CIBERNED), Spain	astridadarmes@hotmail.com
Mathieu Anheim, MD, PhD	1. Service de Neurologie, Hôpitaux Universitaires de Strasbourg, Hôpital de Hautepierre, 1 avenue Molière, Strasbourg, 67098, Cedex, France 2. Fédération de Médecine Translationnelle de Strasbourg (FMTS), Université de Strasbourg, Strasbourg, France 3. Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), INSERM-U964/CNRS-UMR7104/Université de Strasbourg, Illkirch, France	mathieu.anheim@chru-strasbourg.fr
Jonathan Baets, MD, PhD	1. Translational Neurosciences, Faculty of Medicine and Health Sciences, UAntwerpen, Antwerp, Belgium 2. Laboratory of Neuromuscular Pathology, Institute Born-Bunge, University of Antwerp, Antwerp, Belgium 3. Neuromuscular Reference Centre, Department of Neurology, Antwerp University Hospital, Antwerp, Belgium	jonathan.baets@uantwerpen.be
Bernard Brais, MD, PhD	Department of Neurology, McGill University, Montreal Neurological Institute, Montréal, Quebec, Canada	bernard.brais@mcgill.ca
Cynthia Gagnon, PhD	Centre de recherche du CHUS et du Centre de santé et des services sociaux du Saguenay–Lac-St-Jea, Faculté de médecine, Université de Sherbrooke, Québec, Canada	cynthia.gagnon4@usherbrooke.ca
Janina Gburek-Augustat, MD	Division of Neuropaediatrics, Hospital for Children and Adolescents, University of Leipzig, Leipzig, Germany	janina.gburek-augustat@medizin.uni-leipzig.de
Sarah Doss, MD	1. Department of Neurology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany 2. Department of Neurological Sciences, University of Nebraska Medical Center (UNMC), Omaha	sarah.doss@unmc.edu
Hasmet A. Hanagasi, MD	Behavioral Neurology and Movement Disorders Unit, Department of Neurology, Istanbul Faculty of Medicine, Istanbul University, 34,093, Istanbul, Turkey	hasmet@yahoo.com
Christoph Kamm, MD	Department of Neurology, University of Rostock, Rostock, Germany	christoph.kamm@med.uni-rostock.de
Peter Klivenyi, MD	Interdisciplinary Excellence Centre, Department of Neurology, Faculty of Medicine, Albert Szent-Györgyi Clinical Center, University of Szeged, Szeged, Hungary	klivenyi.peter@med.u-szeged.hu
Marcus Grobe-Einsler, MD	1. Department of Neurology, University Hospital Bonn, Bonn, Germany 2. German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany	marcus.grobe-einsler@ukbonn.de
Thomas Klopstock, MD	1. Department of Neurology, Friedrich-Baur-Institute, Ludwig-Maximilians-University of Munich, Munich, Germany 2. German Center for Neurodegenerative Diseases (DZNE), Munich, Germany 3. Munich Cluster for Systems Neurology (SyNergy), Munich, Germany	thomas.klopstock@med.uni-muenchen.de
Martina Minnerop, MD	1. Institute of Neuroscience and Medicine (INM-1), Research Centre Juelich, Juelich, Germany 2. Department of Neurology, Center for Movement Disorders and Neuromodulation, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany 3. Institute of Clinical Neuroscience and Medical Psychology, Medical Faculty, Heinrich-Heine University, Germany	martina.minnerop@med.uni-duesseldorf.de

Name, Title	Affiliation(s)	Contact
Alexander Münchau, MD	Institute of Systems Motor Science, Neurogenetics, University of Lübeck, Lübeck, Germany	alexander.muenchau@neuro.uni-luebeck.de
Mathilde Renaud, MD	1. Service de Génétique Clinique, CHRU de Nancy, Nancy, France 2. INSERM-U1256 NGERE, Université de Lorraine, Nancy, France	m.renaud2@chru-nancy.fr
Filippo M. Santorelli, MD	IRCCS Fondazione Stella Maris, Pisa, Italy	filippo3364@gmail.com
Dagmar Timmann, MD	Department of Neurology and Center for Translational Neuro- and Behavioral Sciences, Essen University Hospital, University of Duisburg-Essen, Essen, Germany	dagmar.timmann-braun@uk-essen.de
Stefan Vielhaber, MD	1. Department of Neurology, Otto-von-Guericke University, Magdeburg, Germany 2. German Center for Neurodegenerative Diseases (DZNE) within the Helmholtz Association, Magdeburg, Germany 3. Center for Behavioral Brain Sciences, Magdeburg, Germany	stefan.vielhaber@med.ovgu.de
Bart P. van de Warrenburg, MD, PhD	Department of Neurology, Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Centre, Nijmegen, The Netherlands	bart.vandewarrenburg@radboudumc.nl
Ginevra Zanni, MD, PhD	Unit of Neuromuscular and Neurodegenerative Disorders, Department of Neurosciences, Bambino Gesù Childrens' Hospital, IRCCS, Rome Italy	ginevra.zanni@opbg.net