Longitudinal Analysis of Natural History Progression of Rare and Ultra-Rare Cerebellar Ataxias Using Item Response Theory

Alzahra Hamdan[1], Niels Hendrickx[2], Andrew C. Hooker[1], Xiaomei Chen[1], Emmanuelle Comets[2,3], Andreas Traschütz[4,5], Rebecca Schüle[4,6], ARCA Study Group[‡], EVIDENCE-RND Consortium[§], France Mentré[2], Matthis Synofzik[4,5,†] and Mats O. Karlsson[1,*]

Degenerative cerebellar ataxias comprise a heterogeneous group of rare and ultra-rare genetic diseases. While disease-modifying treatments are now on the horizon for many ataxias, robust trial designs and analysis methods are lacking. To better inform trial designs, we applied item response theory (IRT) modeling to evaluate the natural history progression of several ataxias, assessed with the widely used scale for assessment and rating of ataxia (SARA). A longitudinal IRT model was built utilizing real-world data from the large autosomal recessive cerebellar ataxia (ARCA) registry. Disease progression was evaluated for the overall cohort as well as for the 10 most common ARCA genotypes. Sample sizes were calculated for simulated trials with autosomal recessive spastic ataxia Charlevoix–Saguenay (ARSACS) and polymerase gamma (POLG) ataxia, as showcased, across multiple design and analysis scenarios. Longitudinal IRT models were able to describe the changes in the latent variable underlying SARA as a function of time since ataxia onset for both the overall ARCA cohort and the common genotypes. The typical progression rates varied across genotypes between relatively high in POLG (~0.98 SARA points/year at SARA = 20) and very low in COQ8A ataxia (~0.003 SARA points/year at SARA = 20). Smaller trial sizes were required in case of faster progression, longer trials (~75–90% less with 5 years vs. 2 years), and larger drug effects (~70–80% less with 100% vs. 50% inhibition). Simulating under the developed IRT model, the longitudinal IRT model had the highest power, with a well-controlled type I error, compared to total score models or end-of-treatment analyses. The established longitudinal IRT framework allows efficient utilization of natural history data and ultimately facilitates the design and analysis of treatment trials in rare and ultra-rare genetic ataxias.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ The small sample sizes in clinical trials hinder the therapeutic progress in rare and ultra-rare diseases such as genetic ataxias. A powerful analysis method is needed to provide robust evidence of disease progression and treatment effects in rare diseases. The ataxia severity, progression, and drug effects are commonly analyzed using the sum score of the scale for the assessment and rating of ataxia (SARA).

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Would the analysis of item-level data of SARA through item response theory (IRT) combined with pharmacometric approaches robustly define the natural disease progression and detect treatment effects for ataxias? To what extent can we utilize natural history data to inform trial designs in rare diseases?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ The disease progression profiles for 10 ataxia genotypes were defined along with their uncertainties using longitudinal IRT models. Clinical trial simulations showed that the use of IRT analysis results in a reduction in sample sizes needed for detecting drug effects. The impact of several design factors and analysis methods was also determined.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ The presented work provides a powerful methodology that makes sufficient use of natural history data to better inform trial designs in genetic ataxias, a type of rare neurological disease.

Rare diseases are defined as any disease with a particularly low prevalence, affecting less than 200,000 individuals in the United States (~6.6 in 10,000) or 5 in 10,000 individuals in Europe.[1] Ultra-rare diseases are those with a prevalence of < 1 in 50,000 or even with single registered cases.[2] Contrary to what the "rare disease" term implies, there are 6000–10,000 clinically defined rare

Received May 6, 2024; accepted September 23, 2024. doi:10.1002/cpt.3466

[1]Pharmacometrics Research Group, Department of Pharmacy, Uppsala University, Uppsala, Sweden; [2]Université Paris Cité, IAME, Inserm, Paris, France; [3]Univ Rennes, Inserm, EHESP, Irset – UMR_S 1085, Rennes, France; [4]Department of Neurodegenerative Diseases, Center for Neurology and Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany; [5]German Center for Neurodegenerative Diseases (DZNE) Tübingen, Tübingen, Germany; [6]Division of Neurodegenerative Diseases, Department of Neurology, Heidelberg University Hospital, Heidelberg, Germany.
*Correspondence: Mats O. Karlsson (mats.karlsson@farmaci.uu.se)
† Shared last authorship with M.O.K.
‡ For ARCA Study Group authors, see **Supplementary Material S1**.
§ For the EVIDENCE-RND members, see the Acknowledgment.

diseases that collectively affect 260–450 million people around the world.[1] Such "collectively not rare" diseases pose a significant health burden with urgent need for innovative treatments.

Autosomal recessive cerebellar ataxias (ARCAs) are a group of rare and ultra-rare neurodegenerative diseases characterized by progressive damage to the cerebellum and its associated tracts.[3] There are > 100 defined ARCA genes, many of which are being targeted with novel disease-modifying therapies.[4] However, due to the large disease heterogeneity and the small number of affected patients, powerful trial designs and analysis methodology that would facilitate the generation of robust evidence of disease progression and treatment effects are lacking, hence hindering the therapeutic progress in many rare diseases including ARCAs. Registry-based natural history studies constitute a key source for informing treatment trial designs. Aiming to apply this concept to the ARCAs, the multinational ARCA registry collects real-world data from large number of patients with recessive and/or early-onset ataxia.[5] Ataxia severity is measured using the scale for the assessment and rating of ataxia (SARA), the most widely used and recommended clinical outcome assessment (COA) for ataxia,[6,7] and is also now considered the primary endpoint in several ataxia treatment trials.[8–13]

Several studies have attempted to assess the longitudinal changes in SARA scores in multiple ARCA genotypes and estimate sample sizes for treatment trials.[14–19] These efforts, however, failed to consider the complex and discrete nature of COA data by treating the total SARA scores (sum scores) and/or individual item scores as continuous data. An alternative approach is to apply the item response theory (IRT) methodology, a statistical framework well equipped for the analysis of COA data on the item level. IRT describes the "true" unobserved (i.e., latent) trait underlying a certain COA (ataxia severity in case of SARA) that is reflected by the observed individual item responses. Such methodology acknowledges the differential characteristics of the individual items of a COA resulting in a sufficient utilization of all information and hence resulting in higher power.[20] Pharmacometric IRT methods have been increasingly used to detect disease progression and treatment effects in several neurological diseases by modeling longitudinal COA item-level data.[21–26]

In previous work, we provided evidence supporting the adequacy of SARA as a COA in ataxias using IRT methodology.[27,28] The developed IRT model was able to describe both the characteristics of SARA items and the cross-sectional item-level data from ARCA patients (1932 visits) with only one common underlying latent variable reflecting the ataxia severity. Leveraging longitudinal data from the ARCA registry, we here (i) extended the previously developed IRT framework to describe the longitudinal changes in ataxia severity captured through SARA; (ii) evaluated the ataxia severity and natural disease progression within and across a large number of ARCA genotypes; and (iii) performed sample size calculations for disease-modifying treatment trials for different ARCA genotypes, trial design scenarios, and analysis approaches.

METHODS

Data

Data were obtained from the ARCA registry, a large prospective multicenter registry (> 30 sites, > 15 countries) capturing longitudinal real-world data across a wide range of rare and ultra-rare ARCAs.[5] The registry complies with the European regulatory authorities' ethical standards and the General Data Protection Regulation (GDPR) and has been approved by the IRB at the University of Tübingen (598/2011BO1). Informed consents were obtained from patients before enrollment.[5] The dataset used in this study consists of 990 patients with a genetically confirmed ARCA and/or early-onset ataxia (onset age < 40 years, an ataxia population stratum known to be highly enriched for ARCAs)[14] comprising a total of 115 different ARCA genotypes. Patients were enrolled in the ARCA registry at any stage of ataxia and up to 65 years after ataxia onset, but most (~ 72%) were enrolled within 30 years after the onset of ataxia. Ataxia onset was defined as first onset of gait disturbance as reported by the patient. A total of 420 subjects were followed longitudinally with up to nine predominantly annual visits resulting in a total of 1932 assessments. The ataxia severity was evaluated using SARA, a clinician-reported outcome measure consisting of eight polychotomous items assessing a patient's movement, balance, speech, and coordinated arm and leg movements.[6] It results in a total composite score ranging from 0 (non-ataxia) to 40 (the most severe ataxia).

Analyses in this work were conducted on the entire ARCA population as well as the 10 most common genetic subpopulations in the dataset. Further description of the ARCA dataset and the 10 common genotypes is presented in **Table S1**.

Disease progression modeling

Cross-sectional IRT model of SARA items' scores. Upon designing a disease COA, multiple items are set to capture one or more disease aspects that cannot be measured directly (i.e., unobserved). In IRT, the underlying aspect(s) is/are called *latent variable(s)* and is/are reflected by the items' responses of the COA. This type of analysis is discussed in the FDA's recent guidance on fit-for-purpose COAs[29] and the technical specification guidance for COA data submissions using IRT.[30]

In previous work, we developed an IRT model for SARA that was able to describe the item-level data using only one common underlying *latent variable*.[27] In clinical terms, the *latent variable* underlying SARA is the *severity of ataxia*. In the IRT model, the mathematical relationship between the probability of obtaining a certain item's score and the patient's *latent variable* (i.e., *ataxia severity*) was modeled using logistic *item characteristic functions*. The model was built using data from the same ARCA dataset used in the current paper but treating each assessment visit as a distinct individual.

Based on the estimated item characteristic functions and patient's responses to the individual SARA items, the ataxia severity of the patient is summarized using one value—the latent variable. The latent variable scale is continuous, unbounded, and relative with standard normal distribution (mean = 0, standard deviation (SD) = 1) where the reference population is the one used for defining the item functions (the cross-sectional

ARCA patients for SARA). Hence, a "typical" ARCA patient has a latent variable value of 0, and 95% of the ARCA patients have a latent variable value between −2 (less ataxic) and +2 (more severe ataxia) (i.e., 2 SD away from the mean of zero). A translation from the latent variable scale to the expected scores for each SARA item (and SARA total score) is also possible using the estimated item characteristic functions as illustrated in **Figure 1**. The *x*-axis in **Figure 1** is the latent variable scale where the zero point and the unit size are assigned based on the reference population (i.e., mean and SD). It ranges between very mild to very high ataxia severity levels expressed as −4 to +4 latent variable values.

For further details on the assessment of SARA adequacy, the characteristics of SARA items, and IRT model diagnostics, see ref. [27]. Furthermore, a tutorial on IRT methodology, the role of IRT in clinical trial and natural history studies, and a guide to results interpretations are provided in **Supplementary Material S3**.

Longitudinal IRT model. In the current work, we extended the IRT model of SARA by adding a longitudinal component to describe the changes in the ataxia severity (assessed on latent variable scale) with time while fixing the parameters of the item characteristic functions to the estimates from the cross-sectional IRT model.[27] This modeling approach allowed for a precise estimation of the individual latent variables while avoiding potential bias in the item characteristic parameters, which was possible if simultaneously estimated in case of longitudinal model misspecifications.[31] In other words, the longitudinal IRT modeling includes the use of (i) SARA item-level data, (ii) the established item characteristic functions which allow expressing item-level data as individual latent variable estimates reflecting the ataxia severity, and (iii) a longitudinal model component to describe the disease progression profile of the ataxia population. The longitudinal model was implemented as follows:

$$\psi_{ij} = \psi_i^0 + \alpha_i t_j \quad (1)$$

where ψ_{ij} and ψ_i^0 are the latent variable values for individual i at visit j and baseline, respectively, and α_i is the disease progression rate (slope) for individual i. Model parameters were assumed to be subject specific and modeled through random effects (inter-individual variability with normal distribution). As described above, the latent variable scale is a hypothetical construct for the ataxia severity underlying SARA with a $N(0, 1)$ distribution relative to the reference population (i.e., ARCA cross-sectional data). The structure of the longitudinal IRT model was developed based on the entire ARCA cohort. Two options of t_j were considered upon modeling the time course of the disease progression: (i) time since inclusion in the registry and (ii) time since ataxia onset. Furthermore, a deviation from linearity was assessed by estimating a power parameter, θ_i, as described in Eq. 2. The estimation of a power term will allow deviation in disease progression rate in both directions—increasing or decreasing with time.

$$\psi_{ij} = \psi_i^0 + \alpha_i t_j^{\theta_i} \quad (2)$$

The best model was then used to describe the disease progression of the 10 most common ARCA genotypes. This was performed using one model in which distinct fixed-effect parameters (i.e., slope and intercept for a typical patient) were estimated for each of the 10 genotypes, while the inter-individual variability parameters were estimated using all patients' data from the 10 subpopulations due to the few numbers of individuals in many genotypes (**Table S1**).

Model building and evaluation. Model implementation and parameter estimation were performed using the nonlinear mixed-effects model software NONMEM version 7.5.0 with the importance sampling algorithm.[32] Model selection between alternative models was based on the likelihood ratio test (LRT) at a significance level of 0.05 for nested models and Bayesian information criterion for mixed-effect models (BIC)[33] for non-nested models in which penalty terms are

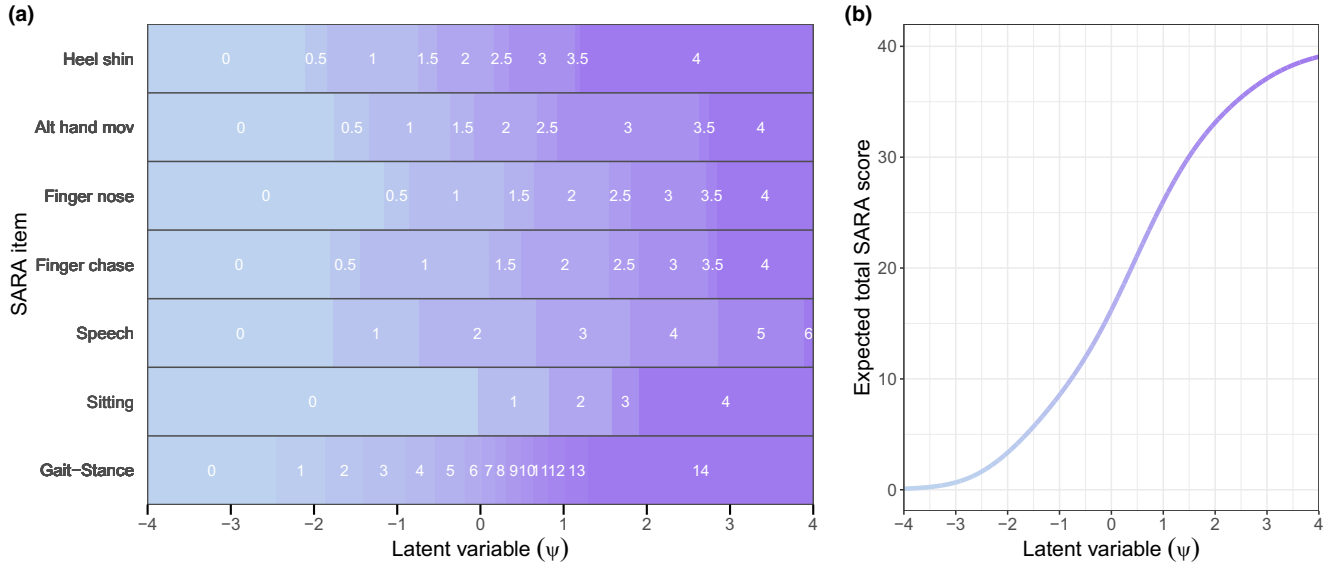

Figure 1 The relationship among SARA latent variable, individual item scores, and the total score. The latent variable scale (*x*-axis) is defined relative to a reference population (i.e., the zero point is the mean and unit size is the SD) and is unbounded (i.e., −∞ to ∞). The displayed latent variable range (−4 to 4) includes >99% of the reference population. (**a**) The expected score of the individual items from SARA at different latent variable estimates from SARA IRT model (cross-sectional model). The expected scores are based on the probabilities of item responses estimated using the item characteristic functions. (**b**) The expected total score of SARA at different latent variable estimates based on IRT-informed polynomial link functions. The linkage between a specific latent variable value and the expected item scores in **panel a** corresponds as well to an expected total score on the curve in **panel b**.

introduced to the likelihood function to reduce the risk of overfitting. In addition, the plausibility of parameter estimates and the model stability (i.e., convergence minimization) were considered during model selection. The ability of the longitudinal IRT model to describe the data was further evaluated using visual predictive check (VPC) implemented in Perl-speaks-NONMEM (PsN).[34,35] VPCs were performed for both item-level data and total SARA by transforming the latent variable values to item scores (or their sums for total SARA) through simulations ($n = 200$). The uncertainty in parameter estimates was evaluated using the sampling importance resampling (SIR) procedure in PsN.[36] SIR was selected due to its efficiency and improved performance in small datasets (as in our case for ARCAs) compared to other methods, such as covariance step in NONMEM or bootstrap.[36,37] The NONMEM model code for the longitudinal IRT model of the ARCA genotypes is shown in **Supplementary Material S4**.

Clinical trial simulations and sample size calculations

General framework. A stochastic simulation and estimation (SSE) procedure[38] was used to assess both type I error and the sample size required to detect a hypothetical drug effect with 80% power and at a significance level of 0.05. Specifically, the final longitudinal IRT model for the 10 subpopulations was used to simulate SARA scores on the item level for 500 trials for sample size calculation and 1500 trials for type I error check with 1000 subjects in each trial. The simulation step was followed by an estimation step where models with and without drug effect (i.e., models where a drug effect is estimated for treatment group and models where treatment allocation is disregarded) were estimated for each simulated trial. Then, the difference in objective function values (ΔOFV) for model pairs (i.e., with and without drug effect) was calculated and a two-sided LRT was performed for detecting significant drug effects (i.e., when ΔOFV > 3.84 for 1 degree of freedom and significance level of 0.05).

Type I error was evaluated by simulating trials with no drug effect and calculated empirically as the ratio of false-positive drug effect. The type I error rate was assumed to be adequate if it was within the 95% prediction interval of a binomial distribution with a success probability of 5% in 1500 trial replicates (4.0–6.2). For sample size calculation, trials were simulated with drug effect (30%, 50%, and 100%), and the power vs. sample size curves were automatically generated using the parametric power estimation (PPE) algorithm in PsN.[39] In PPE, the unknown non-centrality parameter (NCP) of the theoretical χ^2 distribution of the test statistic (i.e., ΔOFV) was estimated based on a reference design with a specific sample size using maximum-likelihood estimation. The NCP is then scaled to generate power for different sample sizes. For a more robust performance of the PPE algorithm in terms of defining the test statistic distribution and NCP extrapolations, a large sample size (1000 subjects) was selected using the reference design. A parametric bootstrap procedure is performed in the PPE to assess the uncertainty of the NCP and, hence, calculate the 95% confidence interval of the power. The sample size calculations were adjusted for 5% type I error rate based on the empirical cumulative distribution of the test statistic under the null hypothesis. This was done through adapting the degrees of freedom of the non-central χ^2 distribution in the PPE algorithm to the equivalent cut-off values for a 5% error rate.

Clinical trial simulations. In the above SSE procedure, clinical trial simulations were performed for a randomized, placebo-controlled (1:1 allocation ratio) trial with a parallel design with a focus on two ARCA subpopulations: autosomal recessive spastic ataxia Charlevoix–Saguenay (ARSACS) and polymerase gamma (POLG)-associated ataxia. These two subpopulations were selected as showcase examples for relatively more common rare ataxia (ARSACS) and rare ataxia with a high progression rate (POLG). Multiple trial scenarios were conducted addressing different trial durations, inclusion criteria, and magnitudes of treatment effect resulting in a total of 32 scenarios for each ARCA subpopulation. Specifically, 2- and 5-year trials were evaluated with assessment frequency of every 6 months. Four simulated groups of patients with different ataxia duration (i.e., time since onset) at the start of the trials were assessed: 0–10 years, 10–20 years, 20–30 years, and a heterogeneous group with 0–30 years of ataxia. Hypothetical disease-modifying effects were introduced in the simulations with 0%, 30%, 50%, and 100% reduction levels in the ataxia progression rate, as follows:

$$\psi_{ij} = \psi_i^0 + \alpha_i \mathrm{TSO}_{0i} + \alpha_i t_j (1 - \mathrm{TE}) \tag{3}$$

where TSO_{0i} is the time since onset at the start of clinical trial for individual i, t_j is the time since the start of trial for visit j, and TE is the treatment effect magnitude (0 in the placebo group and 0.3, 0.5, or 1 in the treatment group). Simulations using the longitudinal IRT model generate discrete data for each item in SARA allowing for subsequent trial analyses using item-level data as in IRT models or analyses on total SARA score by calculating the sum of the items' scores.

Pharmacometric analysis methods. The power and type I error of different types of analysis methods to detect a treatment effect was assessed across the different clinical trial simulation scenarios: (i and ii) both longitudinal and end-of-treatment (EoT) analyses using the longitudinal IRT model, (iii and iv) both longitudinal and EoT analyses using IRT-informed total score model, (v) linear mixed-effects model of the total SARA score, and (vi) four parameters logistic growth model of the total SARA score previously developed on the ARSACS data with mixed effects on all parameters.[40] The total SARA scores are treated as continuous in the total score models.

In the IRT-informed total score model, polynomial link functions were used to translate between the latent variable and the expected mean and SD of the total scores.[41] These link functions were estimated based on the base IRT model of SARA in a previous work[27] and discussed in detail in **Supplementary Material S5**. The IRT-informed total score model allows the use of total score data to describe the disease progression and drug effect on the latent variable scale (same as in Eq. 3, where TE is estimated). Then, the individual latent variable estimates are translated to predicted total scores as follows:

$$Y_{ij} = E\left(Y \mid \psi_{ij}\right) + \mathrm{SD}\left(Y \mid \psi_{ij}\right) * \varepsilon_{ij} \tag{4}$$

where Y_{ij} is the total score for individual i at study visit j, $E\left(Y \mid \psi_{ij}\right)$ and $SD\left(Y \mid \psi_{ij}\right)$ are the expected total score value and SD based on the predetermined polynomials, respectively, of the total score given the estimated latent variable at a certain visit, and ε_{ij} is residual unexplained variability with $N\left(0, \sigma^2\right)$ which allows for a deviation from the predicted total score variance.

The linear mixed-effect model and the four-parameter logistic mixed-effects model of total SARA scores are described in Eqs. 5 and 6, respectively.

$$Y_{ij} = y_i^0 + \alpha_i \mathrm{TSO}_{0i} + \alpha_i t_j (1 - \mathrm{TE}) + \varepsilon_{ij} \tag{5}$$

$$Y_{ij} = \delta_i + \frac{\gamma_i}{1 + e^{\beta_i - \alpha_i \mathrm{TSO}_{0i} - \alpha_i (1-\mathrm{TE}) * t_j}} + \varepsilon_{ij} \tag{6}$$

With δ_i the total SARA score at the time of minus infinity, γ_i the amplitude of the logistic function, α_i the disease progression rate (1/year), β_i the scale parameter, and ε_{ij} is the residual error for individual i at study visit j, which is normally distributed with variance σ^2.

RESULTS

Longitudinal IRT model for the entire ARCA dataset

The longitudinal profiles of the individual latent variables were better described as a function of time since ataxia onset rather than time since inclusion with a ΔBIC of 131. Moreover, no deviation from linearity was observed upon estimating a power model with a power estimate close to 1 (0.95) and ΔOFV of only 0.76 ($p = 0.38$, df = 1), meaning that the more parsimonious linear model is sufficient to describe the longitudinal changes in the latent variable. It should be noted that the time since onset was calculated as the difference between the age at onset and the patient's age at certain visits; however, the age at ataxia onset was missing for around 6% of the patients in the dataset. Missing values were imputed during model estimation using a model-based approach. Further details are described in **Supplementary Material S6**.

Based on the final longitudinal IRT model (Eq. 1), a typical individual in the ARCA dataset has an ataxia severity, measured on the latent variable scale, at the age of onset ($\psi^0_{typical}$) of −0.69 (SE = 0.05) and a linear progression rate ($\alpha_{typical}$) of 0.035 (SE = 0.002) latent variable units per year. The inter-individual variabilities were estimated with SD of 0.86 (SE = 0.09) and 0.02 (SE = 0.0002) for ataxia severity at onset and progression rate, respectively. A negative correlation of −0.25 was estimated between the random effects, meaning that individuals with lower ataxia severity at onset were more likely to progress faster. These results translate to a total SARA score of 11 points at the age of onset as the expected mean of the entire ARCA population and a progression rate between 0.24 and 0.35 SARA points/year for total SARA of 11–20 points (further details are described in **Supplementary Material S5**). Categorical VPCs for individual items and total SARA VPC showed a good ability of the model to describe SARA scores from the entire ARCA dataset (**Supplementary Material S7**).

Disease progression of ARCA genotypes

As a quick exploratory check for heterogeneity among ARCA genotypes in terms of ataxia progression, total SARA scores were plotted vs. time since ataxia onset (**Figure 2**). Visual examination of the exploratory plots indicated potential differences in the disease progression profiles of different ARCA genotypes (e.g., POLG and SYNE1 plots). The differences in the progression profiles between the 10 most common genotypes were incorporated in the longitudinal IRT model by estimating distinct model parameters (ataxia severity at onset and progression rate—in latent variable terms) for each ARCA subpopulation. The typical estimates along with their 95% confidence intervals (CIs) are illustrated in **Figure 3a**.The annual progression rate of latent variable units per year was highest in

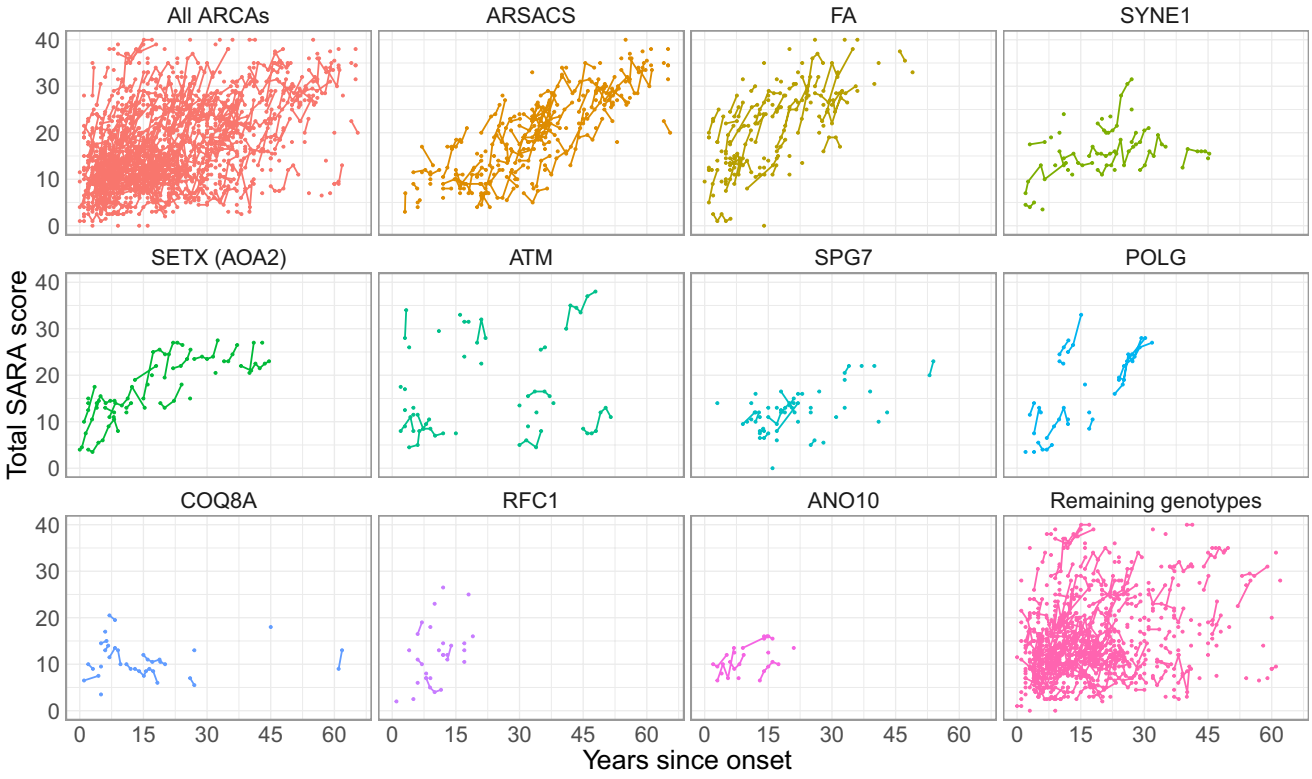

Figure 2 Longitudinal total SARA data of the entire ARCA dataset and the 10 most common genotypes. AOA2, ataxia with oculomotor apraxia type 2; ARCA, autosomal recessive cerebellar ataxia; ARSACS, autosomal recessive spastic ataxia of Charlevoix–Saguenay; ATM, ataxia telangiectasia; FA, Friedreich's ataxia; SYNE1 /SETX/SPG7/RFC1/COQ8A/POLG/ANO10, ataxia related to respective gene.

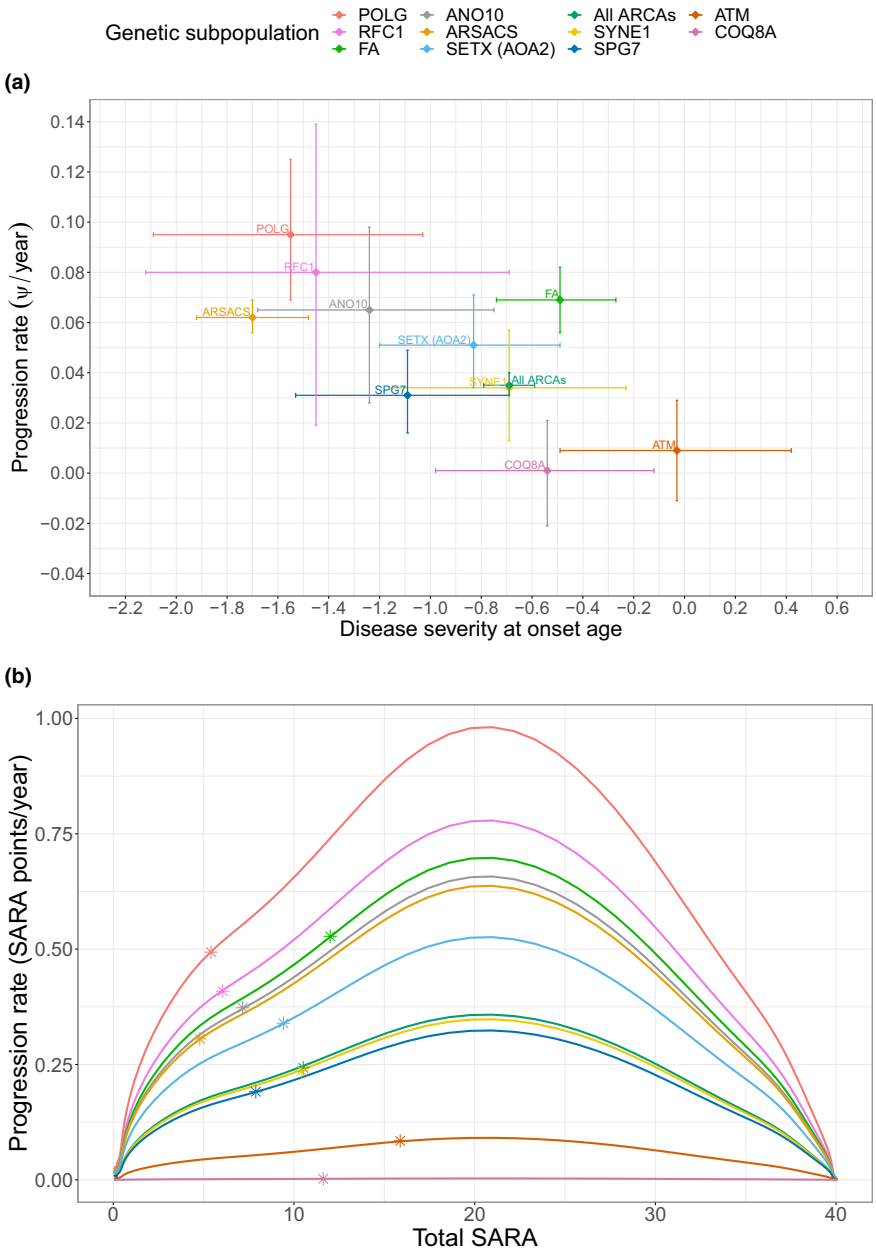

Figure 3 (**a**) Longitudinal IRT model parameter estimates for the 10 most common subpopulations in ARCA dataset as well as the entire ARCA population (treated as one group). The lines represent the 95% confidence interval for the disease severity at onset estimates (horizontal lines) and for the progression rate estimates (vertical lines). (**b**) Back-transformed progression rate values (from the latent variable scale as in (**a**) to total SARA scores) as a function of ataxia severity expressed as total SARA. The stars represent the progression rate and total SARA at the age of onset. AOA2, ataxia with oculomotor apraxia type 2; ARCA, autosomal recessive cerebellar ataxia; ARSACS, autosomal recessive spastic ataxia of Charlevoix–Saguenay; ATM, ataxia telangiectasia; FA, Friedreich's ataxia; and SYNE1 /SETX/SPG7/RFC1/COQ8A/POLG/ANO10, ataxia related to respective gene.

POLG (0.095/year, SE: 0.01) and lowest in ATM (0.009/year, SE: 0.01) and COQ8A (0.001/year, SE: 0.01) in which the CIs encompassed zero. Differences in the ataxia severity at onset are seen with lowest severity in ARSACS (−1.70, SE: 0.12) and POLG (−1.55, SE: 0.27) and highest in ATM (−0.03, SE: 0.23) corresponding to expected mean score of total SARA of 5 in ARSACS and POLG and 16 in ataxia telangiectasia (ATM) at age of onset. To facilitate the interpretation of modeling results, we transformed the model estimates described in latent variable units (e.g., ψ/year) to the equivalent values in SARA total score units (e.g., SARA points/year). The translated results are shown in **Figure 3b** and described in detail in **Supplementary Material S5**.

It is worth noting that there are overlaps between the parameters' confidence intervals of different subpopulations (**Figure 3a**); for example, ARSACS and FA progression rates, which means that the "true" values might be similar between overlapping subpopulations. The trajectories of individual patients vary according to their genotype-specific typical parameters as well as the inter-individual

variability estimates of 0.71 SD (SE = 0.08) for ataxia severity at onset and 0.02 SD (SE = 8×10^{-5}) for progression rate (on latent variable scale). A negative correlation was estimated between the random effects in the 10-subpopulations' model (−0.5).

Clinical trial simulations

Based on the genotype-specific longitudinal IRT model, we performed clinical trial simulations and followed sample size calculations for different trial designs and different analysis methods according to the general framework illustrated in **Figure 4**. As use case examples, we focused on two genotypes as examples for: (i) a common ARCA genotype with the largest subpopulation (ARSACS); and (ii) a highly progressive genotype (POLG). As a substitute for ataxia severity as inclusion criteria, we studied different disease duration groups at the start of the trial which generally represented mild-to-moderate ataxia stages. Mapping between the disease duration and the predicted total SARA scores is shown in **Supplementary Material S8**.

Type I error assessment (**Supplementary Material S9**) corresponding to the significance level of 5% and a target range of (4.0–6.2) (i.e., the 95% prediction interval based on the binomial distribution) showed a well-controlled type I error in IRT models including the EoT analysis in all the studied scenarios for both POLG and ARSACS. Detailed results of type I error are given in **Supplementary Material S9**, which shows that inflation in type I error was seen in both total score models for the early (0–10 years) and the heterogenous (0–30 years) ataxia scenarios. **Supplementary Material S10** shows that this was linked to model misspecification between the simulation and analysis models. Indeed, when simulating and estimating under the logistic total score model, the type I error was controlled for the logistic model in all scenarios, while it was inflated for the linear model in the early, late, and heterogenous inclusion scenarios, as the linear model is less suited to the logistic increase in total scores.

Figure 5 (ARSACS) and **Figure 6** (POLG) show the minimum sample size required (and 95% CI) to detect a drug effect with 80% power in multiple trial scenarios and using the different analysis methods (the logistic total score model was tested in ARSACS only). Among the different studied factors, the trial duration had the largest effect with around 75–90% reduction in sample size upon increasing trial duration from 2 to 5 years. Even though a 5-year duration is not realistic for a clinical trial, it serves as a theoretical lower limit for sample sizes needed in each trial scenario. Compared to the best-case scenarios (100% inhibition of disease progression), 3- to 5- and 10- to 16-fold larger sample sizes are needed in case of 50% and 30% drug effects, respectively. Generally, smaller sample sizes are needed in POLG than in ARSACS (e.g., 42 vs. 94 subjects needed for 5-year trials with 50% drug effect in heterogenous groups analyzed using longitudinal IRT model). The required sample size varied between the different disease duration groups, with a general decrease for trial populations with an increase in ataxia duration and for heterogenous populations with largely varied ataxia duration.

Figures 5 and 6 also reveal some evident differences between the analysis methods. Utilizing the full longitudinal data with longitudinal IRT models required the smallest sample sizes in all tested scenarios (e.g., ~ 20% fewer ARSACS subjects than that in linear total score model). Next in the rank are the IRT-informed and linear total score models which required approximately similar sample sizes for most of the 5-year trials. For the 2-year trials, the linear total score model required smaller sample size than the IRT-informed model for most cases except for the early groups with 0–10 years of ataxia. The performance of the logistic total score model (ran only in ARSACS) was dependent on the tested disease duration group, worst in early groups (0–10 years), especially for 2-year trials, but more robust on other scenarios with closer sample sizes to the IRT-informed and linear total score

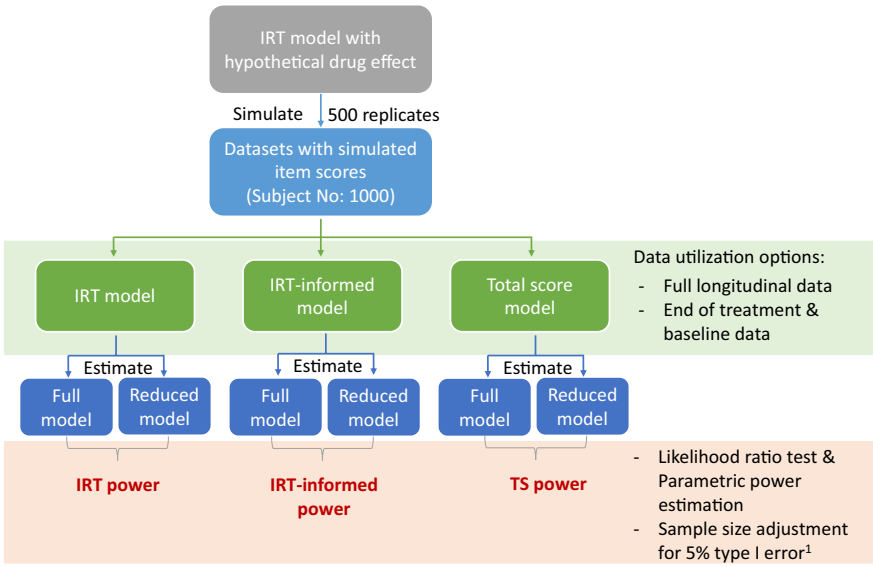

Figure 4 General workflow of the clinical trial simulations and power/sample size calculations. [1]Based on the empirical cumulative distribution of the test statistic (Δ OFV) under the null hypothesis (0% drug effect) for 1500 simulations. Type I error assessments were performed following the same workflow but with 1500 simulations and empirical calculations of false-positive drug effects.

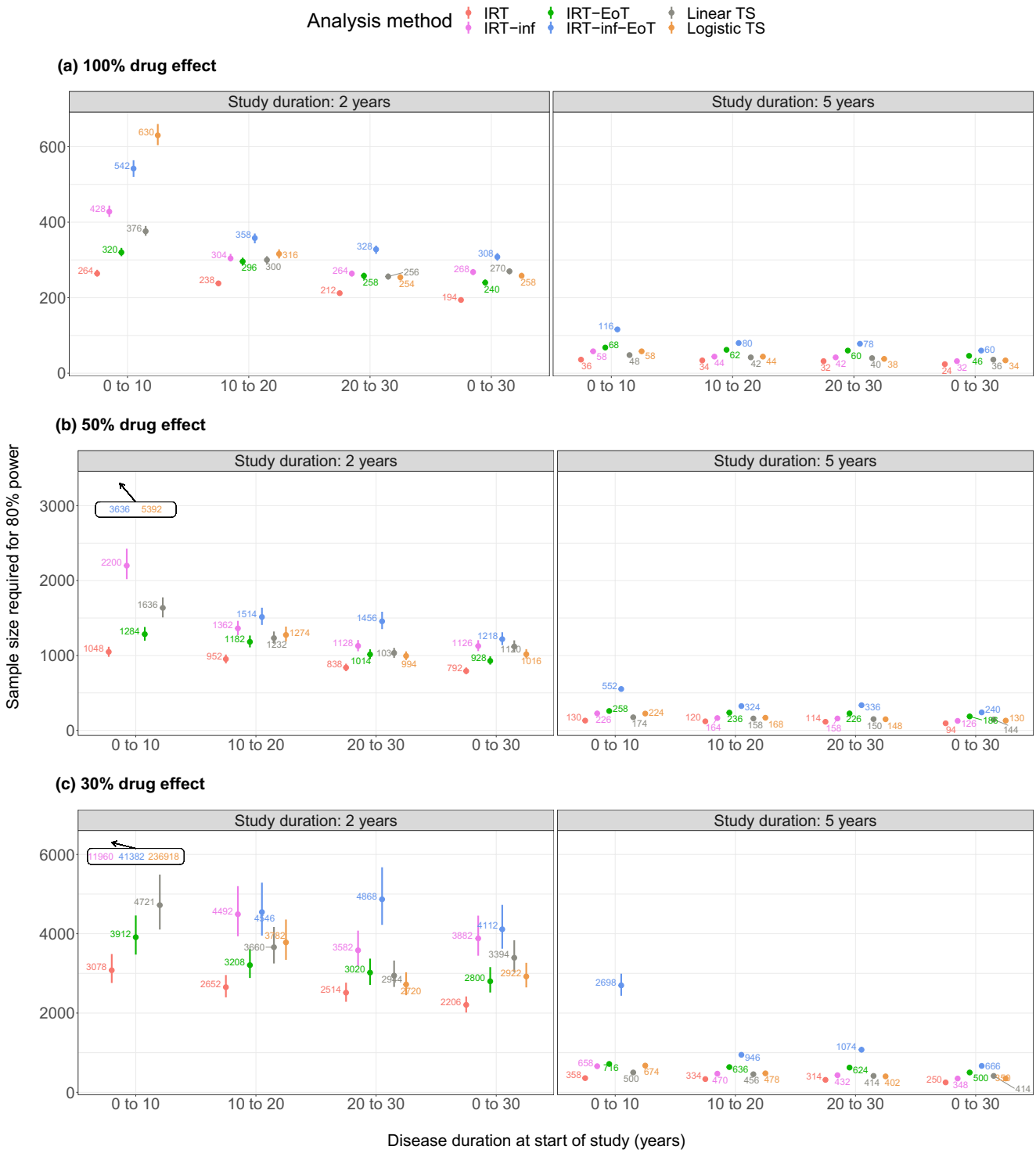

Figure 5 Sample size needed to obtain 80% power of detecting 100% (**a**), 50% (**b**), and 30% (**c**) drug effects in **ARSACS** subpopulation. The vertical lines represent the sample size range corresponding to the 95% confidence interval of the estimated non-centrality parameter (very tight relative to the plot scales if not visualized). IRT, Item response theory model; IRT-inf, IRT informed total score model; IRT-EoT, IRT model with only baseline and end-of-treatment data; IRT-inf-EoT, IRT-informed model with only baseline and end-of-treatment data; Linear TS, linear total score model; Logistic-TS, logistic total score model.

model estimates. **Supplementary Material S10** again links this with the simulation setting: when simulating with a logistic total score model, the power of the logistic total score model in all scenarios was similar or higher than when simulating under the longitudinal IRT model, except in the early inclusion scenario where very low power is linked to slow progression in the ARSACS population that could not be separated from the residual variability with the logistic total score model. Finally, more

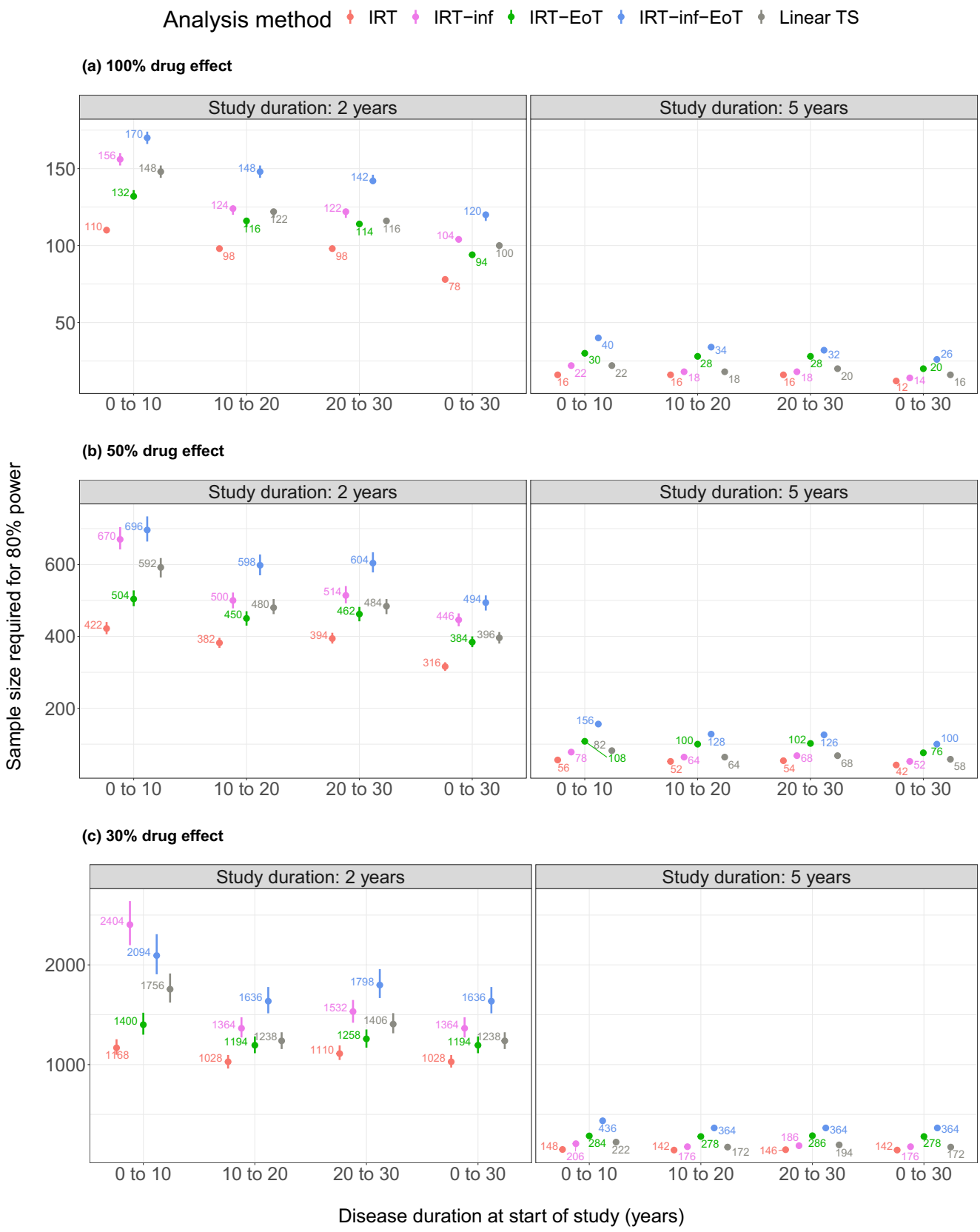

Figure 6 Sample size needed to obtain 80% power of detecting 100% (**a**), 50% (**b**), and 30% (**c**) drug effects in **POLG** subpopulation. The vertical lines represent the sample size range corresponding to the 95% confidence interval of the estimated non-centrality parameter (very tight relative to the plot scales if not visualized). IRT, Item response theory model; IRT-inf, IRT-informed total score model; IRT-EoT, IRT model with only baseline and end-of-treatment data; IRT-inf-EoT, IRT-informed model with only baseline and end-of-treatment data; Linear TS, Linear total score model.

subjects were needed when using only baseline and EoT data for both longitudinal IRT and IRT-informed models with around 20% and 90% larger samples for 2- and 5-year trials, respectively.

DISCUSSION

There is a pressing need for robust trial designs and analysis approaches to facilitate therapeutic progress in rare neurological diseases, for example, genetic ataxias. Utilizing real-world data from the largest natural history database for ARCAs, the ARCA registry, and combined with the innovative pharmacometric IRT methodology, we evaluated the disease progression in several ARCAs as use cases with which we assessed a series of trial designs and analysis approaches. For the first time ever, longitudinal IRT models were successfully implemented in rare neurological disease modeling. In the example of genetic ataxias, they were used to analyze the longitudinal changes in ataxia severity, measured by SARA, focusing on 10 ARCA genotypes as exemplary showcases for rare and ultra-rare ataxias.

The adequacy of SARA as a COA for ataxias was shown in previous work using the IRT methodology to describe the individual item characteristics.[27] Based on the established item characteristic functions, we built a longitudinal IRT model that described disease progression in ARCAs. Differential disease progression profiles were identified for the 10 most common ARCAs through IRT modeling (**Figure 3**). The highest progression rate was seen in POLG, followed by RFC1, Friedreich's ataxia (FA), ANO10, ARSACS, SETX, SYNE1, and SPG7 (in order). For ataxia telangiectasia (ATM genotype) and COQ8A, the progression rate estimates were very low and not significantly different from zero limiting the feasibility of trial simulations and sample size calculation in these two ARCA populations with the available data and results. Clinical trial simulations were performed to evaluate factors affecting sample sizes needed to detect hypothetical drug effects, including (i) trial duration, (ii) disease progression rate, (iii) disease duration at inclusion, (iv) magnitude of drug effect, (v) longitudinal analysis or EoT, and (vi) analysis model (i.e., IRT based or total score based). Larger sample sizes were needed in case of shorter trials (e.g., 2-years vs. 5-years), slower disease progression (e.g., slower progression in ARSACS vs. faster progression in POLG), shorter ataxia durations (e.g., 0–10 years vs. 20–30 years), smaller drug effects (e.g., 30% vs. 50%), less utilization of the longitudinal data (EoT analysis vs. full longitudinal data analysis), and less utilization of SARA data (total score models vs. IRT models).

The application of pharmacometric IRT models to describe longitudinal SARA data has multiple benefits. First, the IRT methodology takes into consideration the discrete and bounded nature of COA data which are disregarded in traditional total score-based analyses. Second, the differential informativeness of individual items is inherently acknowledged in IRT models, which deems the description of disease progression in terms of latent variable changes less sensitive to potentially noisy or slowly progressing items compared to total score analyses. On the other hand, the interpretation of the estimated longitudinal IRT parameters can be challenging due to the uncommon range of the latent variable and the use of a reference population ($N(0, 1)$). To ease the interpretation, we used IRT-informed link functions to translate estimates from the latent variable scale to the total SARA scores as shown in **Figure 3b** and Supplementary Material **S5**.

Our disease progression findings (**Figure 3**) are generally in line (e.g., ranking of genotypes based on their estimated progression rates) with previous results reported by Traschütz *et al*. who used largely the same datasets for eight ARCA common genotypes.[15] The only exception is that we found a relatively high progression rate in ARSACS. This could, however, be attributed to the larger dataset (173 vs. 140 patients) with longer follow-ups and more follow-ups in advanced disease patients in the dataset in our analysis. Contrary to previous research, RFC1 and COQ8A genotypes had lower progression rates which could be explained by the differences in analysis approaches and dataset sizes.[16,17]

The small sample sizes and the limited longitudinal data in natural history studies are the two major obstacles to robust longitudinal analysis for rare neurological diseases such as genetic ataxias. To mitigate these challenges: (i) we described the disease progression over *time since onset of ataxia* instead of *time since the first visit*, allowing a better utilization of cross-sectional data by informing the progression rate estimation; (ii) an SIR procedure was used to estimate parameters uncertainty which is more suitable for small datasets comparing to other commonly used methods (e.g., covariance matrix or bootstrap).[36,37] In spite of this, the limited amount of data still hinders precise estimation of disease progression in some ARCA populations. One example is the large uncertainty in parameter estimates for those ARCA populations with few patients and/or number of observations, as can be seen in the case of ANO10 (12 patients, 27 visits) and RFC1 (25 patients, 31 visits), which have the widest parameters' CI, comparing to ARSACS (173 patients, 349 visits) and FA (110 patients, 200 visits) with the narrowest CI (see **Table S1** and **Figure 3**). Another example is the potential negative impact of the limited and widely spread data in the ataxia telangiectasia population (ATM genotype), especially in the early years after onset (**Figure 2**), on the ability to detect longitudinal changes and estimate model parameters (i.e., inaccurate estimation of baseline or slope might affect the ability to accurately estimate the other parameter).

While the trial simulations focused on early and intermediate stages of ataxia—as being targeted in trials for neurodegenerative diseases[42]—, the calculated sample sizes varied between the different disease duration groups (i.e., ataxia severity). This intriguing result may be related to the differential informativeness of SARA across the disease severity range as shown in **Figure 7** (e.g., lowest sample sizes and highest informativeness for late ataxia groups, i.e., group of 20–30 years). Consistent with the IRT literature for other diseases,[23,26,43] applying longitudinal IRT modeling in simulated ataxia trials resulted in improved power of detecting drug effects without type I inflation. It is somewhat surprising that despite the higher information utilization in the IRT-informed model, its power was comparable to the total score model in most scenarios (except 0–30 years subgroups) according to **Figure 5**. A possible explanation is that a simple linear model can be sufficient to describe the link between latent variable and expected SARA scores at small

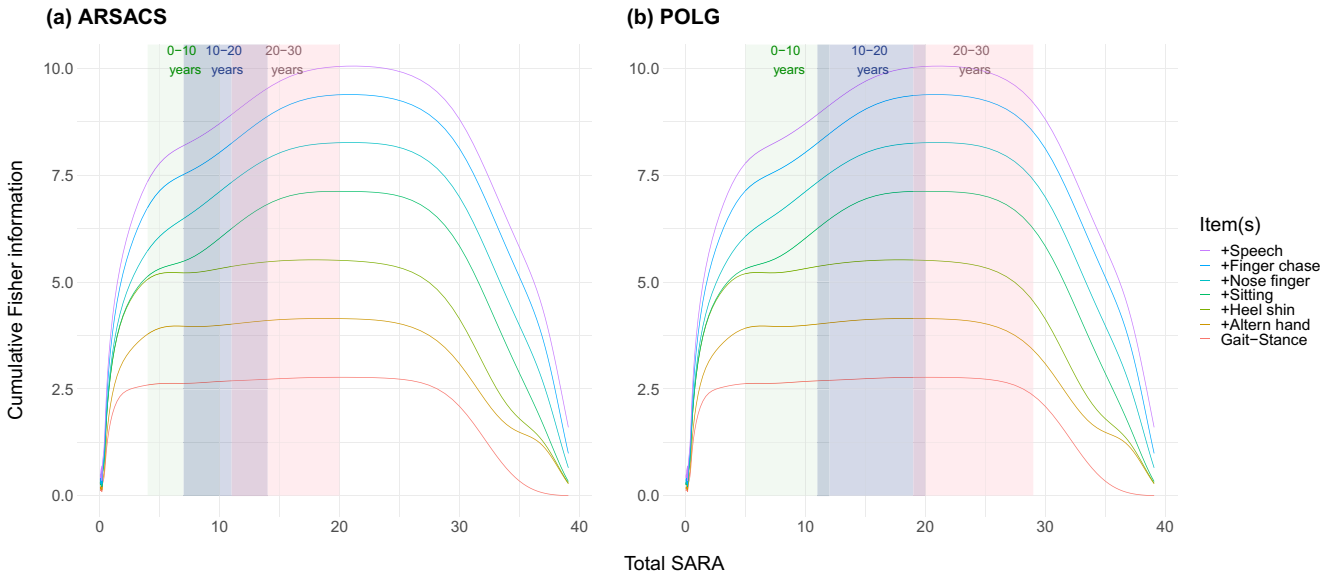

Figure 7 The cumulative Fisher information of SARA items (starting from the most informative item, gait stance, to the lowest one, speech) vs. the total SARA score (back transformed from the latent variable using IRT-informed link functions. The shaded areas represent the interquartile range of total SARA scores (based on IRT model simulations, $n=500$) for different disease duration groups of ARSACS (**a**) and POLG (**b**) genetic subpopulations. The Fisher information content of the different items of SARA assessment as a function of the underlying latent variable was evaluated in previous research.[27]

intervals (i.e., linear regions of non-linear curve, see **Figure 1b**). It is important to bear in mind the potential bias in the simulation study results due to the use of the IRT model for both simulating and analyzing data. This fact can partially explain the inflated type I error in other analysis models, including the logistic total score model, due to model misspecifications. Further simulations under the logistic total score model are shown in **Supplementary Material S10**. When simulating and estimating under the logistic total score model, type I error is preserved in every scenario and the power was similar to or higher than when simulating under the longitudinal IRT model for most inclusion criteria scenarios. The only exception was in case of the early inclusion criterion scenario where a lower power was found when simulating under the logistic total score model.

Potential limitations of present studies should be noted. First, using the time since (patient-reported) ataxia onset as the independent variable in the model can be a source of noise in the data due to (i) the inaccuracy in recollecting the age at the onset of ataxia, and (ii) the unclarity of the definition of the age of onset (despite tying it to the onset of first gait disturbances) which might result in increased uncertainty in model estimates. Second, it is likely that the correction approach for type I error in sample size calculations has potentially introduced further noise to the results due to potential violation of key assumptions in the PPE algorithm (e.g., extrapolation to very large sample sizes as the case of total score models for the early ARSACS groups in **Figure 5**). Future research should consider a careful control for type I errors with a very large number of simulations.

It is important to bear in mind that our simulations and sample size calculations aimed to establish a thorough understanding of the impact of trial designs and analysis factors on the statistical power. Some studied trial scenarios (e.g., complete inhibition of disease progression and 5-year trials) were conducted for theoretical illustrative purposes and not to be adopted as such in real ataxia trials. Clinical trials for other neurological diseases showed modest disease-modifying effects (e.g., 27–35% in approved treatments for Alzheimer's disease[44,45]). For recessive ataxias, omaveloxolone drug slowed progression in Friedreich ataxia patients by 50% each year for 3 years compared to corresponding data from an external control group in a natural history study.[46] Nonetheless, our results can provide insights about feasible sample sizes and the choice of design factors and analysis methods. In future, further trial designs that are more feasible for rare and ultra-rare diseases will be evaluated.

CONCLUSION

A pharmacometric longitudinal IRT framework was developed to describe the disease progression of multiple ARCAs using item-level SARA data. This methodology allowed more effective and precise analyses of the disease progression and treatment effects in ARCAs requiring smaller sample sizes compared to other analysis methods (i.e., analysis on the total score). Implementation of natural history data analyses can better inform trial designs and facilitate planning for future treatment trials for neurological diseases, such as genetic ataxias.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

Members of the Evidence-RND consortium include Nicole Maria Heussen, Ralf-Dieter Hilgers, Thomas Klockgether, Yevgen Ryeznik, and Oleksandr Sverdlov.

CONFLICTS OF INTEREST

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, Biogen, GenOrph, AviadoBio, Biohaven, Zevra, Solaxa, 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. Dr. Mentré has received consultancy fees from Pharmetheus and Ipsen, all unrelated to this manuscript. Dr Comets has received consultancy fees from Sanofi, unrelated to this manuscript. All other authors declared no competing interests in this work.

FUNDING

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 (no. 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.). The computations of models were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement nos. 2022-06725 and 2018-05973.

AUTHOR CONTRIBUTIONS

A.H., N.H., A.C.H., X.C., E.C., A.T., R.S., F.M., M.S., and M.O.K. wrote the manuscript; A.H., A.C.H., X.C., E.C., A.T., R.S., F.M., M.S., and M.O.K. designed the research; and A.H. and N.H. performed the research and analyzed the data.

DATA AVAILABILITY STATEMENT

Data supporting the study findings are available on request.

1. Lee, C.E., Singleton, K.S., Wallin, M. & Faundez, V. Rare genetic diseases: nature's experiments on human development. *iScience* **23**, 101123 (2020).
2. Smith, C.I.E., Bergman, P. & Hagey, D.W. Estimating the number of diseases – the concept of rare, ultra-rare, and hyper-rare. *iScience* **25**, 104698 (2022).
3. Anheim, M., Tranchant, C. & Koenig, M. The autosomal recessive cerebellar ataxias. *N. Engl. J. Med.* **366**, 636–646 (2012).
4. Synofzik, M., Puccio, H., Mochel, F. & Schöls, L. Autosomal recessive cerebellar ataxias: paving the way toward targeted molecular therapies. *Neuron* **101**, 560–583 (2019).
5. Traschütz, A. *et al*. The ARCA registry: a collaborative global platform for advancing trial readiness in autosomal recessive cerebellar ataxias. *Front. Neurol.* **12**, 677551 (2021).
6. Schmitz-Hübsch, T. *et al*. Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology* **66**, 1717–1720 (2006).
7. Klockgether, T. *et al*. Consensus recommendations for clinical outcome assessments and registry development in ataxias: Ataxia Global Initiative (AGI) working group expert guidance. *Cerebellum* **23**, 924–930 (2023).
8. Feil, K. *et al*. Safety and efficacy of acetyl-DL-leucine in certain types of cerebellar ataxia: the ALCAT randomized clinical crossover trial. *JAMA Netw. Open* **4**, e2135841 (2021).
9. Lei, L.F. *et al*. Safety and efficacy of valproic acid treatment in SCA3/MJD patients. *Park. Relat. Disord.* **26**, 55–61 (2016).
10. França, C. *et al*. Effects of cerebellar transcranial magnetic stimulation on ataxias: a randomized trial. *Park. Relat. Disord.* **80**, 1–6 (2020).
11. Coarelli, G. *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.* **21**, 225–233 (2022).
12. Benussi, A. *et al*. Cerebello-spinal tDCS in ataxia: a randomized, double-blind, sham-controlled, crossover trial. *Neurology* **91**, e1090–e1101 (2018).
13. Fields, T. *et al*. N-acetyl-L-leucine for Niemann-pick type C: a multinational double-blind randomized placebo-controlled crossover study. *Trials* **24**, 361 (2023).
14. Traschütz, A. *et al*. Autosomal recessive cerebellar ataxias in Europe: frequency, onset, and severity in 677 patients. *Mov. Disord.* **38**, 1109–1112 (2023).
15. Traschütz, A. *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.* **94**, 470–485 (2023).
16. Traschütz, A. *et al*. Clinico-genetic, imaging and molecular delineation of COQ8A-ataxia: a multicenter study of 59 patients. *Ann. Neurol.* **88**, 251–263 (2020).
17. Traschütz, A. *et al*. Natural history, phenotypic Spectrum, and discriminative features of multisystemic RFC1 disease. *Neurology* **96**, e1369–e1382 (2021).
18. Reetz, K. *et al*. Progression characteristics of the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS): a 4-year cohort study. *Lancet Neurol.* **20**, 362–372 (2021).
19. Pandolfo, M. Neurologic outcomes in Friedreich ataxia: study of a single-site cohort. *Neurol Genet.* **6**, e415 (2020).
20. Ueckert, S. Modeling composite assessment data using item response theory. *CPT Pharmacometrics Syst. Pharmacol.* **7**, 205–218 (2018).
21. Arrington, L., Ueckert, S., Ahamadi, M., Macha, S. & Karlsson, M.O. Performance of longitudinal item response theory models in shortened or partial assessments. *J. Pharmacokinet. Pharmacodyn.* **47**, 461–471 (2020).
22. Balsis, S., Unger, A.A., Benge, J.F., Geraci, L. & Doody, R.S. Gaining precision on the Alzheimer's disease assessment scale-cognitive: a comparison of item response theory-based scores and total scores. *Alzheimers Dement.* **8**, 288–294 (2012).
23. 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.* **34**, 2109–2118 (2017).
24. Gottipati, G., Karlsson, M.O. & Plan, E.L. Modeling a composite score in Parkinson's disease using item response theory. *AAPS J.* **19**, 837–845 (2017).
25. Novakovic, A.M., Krekels, E.H.J., Munafo, A., Ueckert, S. & Karlsson, M.O. Application of item response theory to modeling of expanded disability status scale in multiple sclerosis. *AAPS J.* **19**, 172–179 (2017).
26. Ueckert, S. *et al*. Improved utilization of ADAS-cog assessment data through item response theory based pharmacometric modeling. *Pharm. Res.* **31**, 2152–2165 (2014).
27. Hamdan, A. *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.* **13**, 1327–1340 (2024).
28. Hamdan, A. *et al*. Item Response Theory Analysis of the Scale for the Assessment and Rating of Ataxia in Autosomal Recessive Cerebellar Ataxias. In p. Abstr 10626. <www.page-meeting.org/?abstract=10626>.
29. 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. Changes in rheumatoid factor activity during the course of sarcoidosis, 44, 60, 67 <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-selecting-developing-or-modifying-fit-purpose-clinical-outcome> (2022) Accessed Feb 2, 2023.
30. US Food and Drug Administration. US Food and Drug Administration FDA; Submitting Clinical Trial Datasets and Documentation for Clinical Outcome Assessments Using Item Response Theory <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-clinical-trial-datasets-and-documentation-clinical-outcome-assessments-using-item> (2023). Accessed July 7, 2024.
31. Arrington, L. & Karlsson, M.O. Comparison of two methods for determining item characteristic functions and latent variable time-course for Pharmacometric item response models. *AAPS J.* **26**, 21 (2024).
32. Bauer, R.J. NONMEM tutorial part II: estimation methods and advanced examples. *CPT Pharmacometrics Syst. Pharmacol.* **8**, 538–556 (2019).
33. Delattre, M., Lavielle, M. & Poursat, M.A. A note on BIC in mixed-effects models. *Elect. J. Stat.* **8**, 456–475 (2014).
34. Bergstrand, M., Hooker, A.C., Wallin, J.E. & Karlsson, M.O. Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. *AAPS J.* **13**, 143–151 (2011).
35. Keizer, R., Karlsson, M. & Hooker, A. Modeling and simulation workbench for NONMEM: tutorial on Pirana, PsN, and Xpose. *CPT Pharmacometrics Syst. Pharmacol.* **2**, e50 (2013).
36. Dosne, A.G., Bergstrand, M., Harling, K. & Karlsson, M.O. Improving the estimation of parameter uncertainty distributions in nonlinear mixed effects models using sampling importance resampling. *J. Pharmacokinet. Pharmacodyn.* **43**, 583–596 (2016).
37. Dosne, A.G., Bergstrand, M. & Karlsson, M.O. An automated sampling importance resampling procedure for estimating parameter uncertainty. *J. Pharmacokinet. Pharmacodyn.* **44**, 509–520 (2017).
38. Karlsson, M.O. & Nordgren, R. Perl-speaks-NONMEM: User documentation: SSE user guide <https://uupharmacometrics.github.io/PsN/docs.html> Accessed December 11, 2023.
39. Ueckert, S., Karlsson, M.O. & Hooker, A.C. Accelerating Monte Carlo power studies through parametric power estimation. *J. Pharmacokinet. Pharmacodyn.* **43**, 223–234 (2016).
40. Hendrickx, N. *et al*. Prediction of individual disease progression including parameter uncertainty in rare neurodegenerative diseases: the example of Autosomal-Recessive Spastic Ataxia Charlevoix Saguenay (ARSACS). *AAPS J.* **26**(3), 57 (2024).
41. Wellhagen, G.J., Ueckert, S., Kjellsson, M.C. & Karlsson, M.O. An item response theory-informed strategy to model total score data from composite scales. *AAPS J.* **23**, 45 (2021).
42. Benatar, M. *et al*. Preventing amyotrophic lateral sclerosis: insights from pre-symptomatic neurodegenerative diseases. *Brain* **145**, 27–44 (2022).
43. Chen, C., Jönsson, S., Yang, S., Plan, E.L. & Karlsson, M.O. Detecting placebo and drug effects on Parkinson's disease symptoms by longitudinal item-score models. *CPT: Pharm Syst Pharmacol* **10**, 309–317 (2021).
44. Lilly's Kisunla™ (donanemab-azbt) Approved by the FDA for the Treatment of Early Symptomatic Alzheimer's Disease | Eli Lilly and Company <https://investor.lilly.com/news-releases/news-release-details/lillys-kisunlatm-donanemab-azbt-approved-fda-treatment-early> Accessed July 24, 2024.
45. LECANEMAB CONFIRMATORY PHASE 3 CLARITY AD STUDY MET PRIMARY ENDPOINT, SHOWING HIGHLY STATISTICALLY SIGNIFICANT REDUCTION OF CLINICAL DECLINE IN LARGE GLOBAL CLINICAL STUDY OF 1,795 PARTICIPANTS WITH EARLY ALZHEIMER'S DISEASE | Biogen <https://investors.biogen.com/news-releases/news-release-details/lecanemab-confirmatory-phase-3-clarity-ad-study-met-primary> Accessed July 24, 2024.
46. Lynch, D.R. *et al*. Propensity matched comparison of omaveloxolone treatment to Friedreich ataxia natural history data. *Ann. Clin. Transl. Neurol.* **11**, 4–16 (2024).