Application of Item Response Theory to Modeling of Expanded Disability Status Scale in Multiple Sclerosis.

Novakovic, A.M.; Krekels, E.H.; Munafo, A.; Ueckert, S.; Karlsson, M.O.

The AAPS Journal, Vol. 19, No. 1, January 2017 (© 2016)
DOI: 10.1208/s12248-016-9977-z

Research Article

Application of Item Response Theory to Modeling of Expanded Disability Status Scale in Multiple Sclerosis

A. M. Novakovic,[1,4] E. H. J. Krekels,[1,2] A. Munafo,[3] S. Ueckert,[1] and M. O. Karlsson[1]

Received 3 June 2016; accepted 15 August 2016; published online 15 September 2016

ABSTRACT. In this study, we report the development of the first item response theory (IRT) model within a pharmacometrics framework to characterize the disease progression in multiple sclerosis (MS), as measured by Expanded Disability Status Score (EDSS). Data were collected quarterly from a 96-week phase III clinical study by a blinder rater, involving 104,206 item-level observations from 1319 patients with relapsing-remitting MS (RRMS), treated with placebo or cladribine. Observed scores for each EDSS item were modeled describing the probability of a given score as a function of patients' (unobserved) disability using a logistic model. Longitudinal data from placebo arms were used to describe the disease progression over time, and the model was then extended to cladribine arms to characterize the drug effect. Sensitivity with respect to patient disability was calculated as Fisher information for each EDSS item, which were ranked according to the amount of information they contained. The IRT model was able to describe baseline and longitudinal EDSS data on item and total level. The final model suggested that cladribine treatment significantly slows disease-progression rate, with a 20% decrease in disease-progression rate compared to placebo, irrespective of exposure, and effects an additional exposure-dependent reduction in disability progression. Four out of eight items contained 80% of information for the given range of disabilities. This study has illustrated that IRT modeling is specifically suitable for accurate quantification of disease status and description and prediction of disease progression in phase 3 studies on RRMS, by integrating EDSS item-level data in a meaningful manner.

KEY WORDS: cladribine tablets; disease progression model; expanded disability status scale; item response theory; multiple sclerosis.

INTRODUCTION

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease of the central nervous system (1). MS affects over 1 million people worldwide, and it is the leading cause of non-traumatic disability in young adults (2). Over 80% of all patients present with relapsing-remitting MS (RRMS), which is characterized by unpredictable acute episodes of neurological dysfunction named relapses, followed by variable recovery and periods of clinical stability.

The heterogeneity of the MS patient population and complexity of its clinical course have offered challenges to the quantification of disease severity and progression. The clinical manifestations of the disease are extremely variable, even in an individual patient, ranging from motor and sensory problems to cognitive and affective disorders, which renders it necessary to use multidimensional outcome measures.

Since the 1960s, many scales for rating disability caused by MS have been proposed, but none has been entirely satisfactory (3). The Kurtzke Expanded Disability Status Scale (EDSS) remains the most widely used scoring system in MS. Its assessment is based on seven functional systems including vision, brainstem, pyramidal, cerebellar, sensory, bowel and bladder, mental (cerebral), and ambulation (500-m walk), and reliance on aid. The EDSS is a summarized measure which ranges from 0 (normal neurological exam) to 10 (death due to MS) in incremental steps of 0.5 (4). Despite its wide use and acceptance, there are several perceived problems with the use of the scale, such as limited inter-rater reproducibility, bimodal distribution of the scale, and potentially unequal steps, mostly due to its ordinal nature (5, 6). Its overall score is greatly weighted toward ambulation,

Alain Munafo's affiliation is part of Merck Serono S.A. Coinsins, Switzerland, an affiliate of Merck KGaA, Darmstadt, Germany.

Electronic supplementary material The online version of this article (doi:10.1208/s12248-016-9977-z) contains supplementary material, which is available to authorized users.

[1] Department of Pharmaceutical Biosciences, Uppsala University, Box 591, 751 24, Uppsala, Sweden.
[2] *Present Address:* Division of Pharmacology, Leiden Academic Centre of Drug Research, Leiden University, Leiden, Netherlands.
[3] Merck Institute for Pharmacometrics, Merck Serono, Lausanne, Switzerland.
[4] To whom correspondence should be addressed. (e-mail: ana.kalezic@farmbio.uu.se)

especially in higher scores (EDSS > 3.5) (7) and is rather insensitive to cognitive or upper limb dysfunctions. It is important to note that EDSS itself is rarely used as a clinical endpoint in MS clinical trial, but rather the EDSS-related endpoint: time to sustained EDSS progression.

Quantifying the disease severity in MS is important to monitor individual patients during their treatment and for evaluating experimental therapies in clinical trials. As increasing numbers of treatment options become available, sensitive clinical outcome measures that can detect small changes in disability that reliably reflect long-term changes in disease progression are required. Identifying effective treatments depends upon the availability of outcome measures that exhibit good sensitivity to rates of changes caused by the disease.

Traditionally, item response theory (IRT) models have been applied in educational testing to measure ability or proficiency and in psychological assessments to measure personality traits (8). Also, health outcome researchers have been employing IRT to questionnaire development, evaluation, and refinement (9). IRT is a statistical theory consisting of mathematical models expressing the probability of the particular response to a scale item as a function of an underlying trait, here disability of a person (10). IRT models are also referred to as latent trait models, because the latent "unobservable" trait of interest cannot be measured directly and is therefore assessed indirectly by scoring various items constructed to measure that underlying domain. Traditional scoring consists of summarizing all the information in one composite score, which might lead to loss of information captured in the individual item. The recent application of IRT to Alzheimer's disease has demonstrated that increased precision in cognitive assessment can be achieved by not only considering scores on item level, but also how those items function and the amount of information they contain for the studied population (11, 12).

Here, we report the development of the first IRT model within a NLME (non-linear mixed effect) framework in MS therapeutic area. Analysis was based on the data from CLARITY (CLAdRIbine Tablets treating multiple sclerosis orallY) study where cladribine was found to reduce, as compared to placebo, the risk of 3-month sustained progression, by 33 and 31% in the cladribine 3.5 and 5.25 mg/kg groups, respectively (13). Giovannoni *et al.* have reported that the administration of cladribine tablets have been found to be also efficient in regard to other studied clinical endpoints: annualized relapse rate (primary endpoint), percentage of relapse-free patients, and occurrence of magnetic resonance imaging (MRI) detected brain lesions. The current work investigates the possibility of quantification of MS disease progression and of effect of cladribine tablets. We also explore the information content of each item constituting EDSS.

MATERIALS AND METHODS

Patients and Study Design

Data from the CLARITY clinical trial were included in the analysis. CLARITY was a phase III randomized, multicenter, double blind, parallel group, controlled study, evaluating the efficacy and safety of 3.5 and 5.25 mg/kg cumulative doses of cladribine tablets over 96 weeks in patients with RRMS. Enrolled subjects had a diagnosed definite relapse-remitting form of multiple sclerosis, according to the McDonald criteria (14). Outcome assessments were conducted in identical fashion to all other major MS clinical trials. The blind was maintained by utilizing a treating physician who viewed clinical laboratory results, and assessed adverse events and safety information. Patients received neurological assessment at baseline and every 12 weeks thereafter for the duration of the study by an independent blinded evaluating physician. The additional details of the study protocol, subject characteristics, and study results can be found in the original publication (13).

Modeling Methodology

Analyses were performed in the software NONMEM 7.2.0, and Laplacian estimation method was applied for parameter estimation (15). The simulation-based diagnostics were realized using computer-intensive statistical methods available in the Perl-coded program PsN (16).

In addition to seven polychotomous items of functional systems with internal ranking, EDSS comprises measures of ambulation (0–500 m walk) and reliance on aid (0, 1, 2). According to neurostatus definition (www.neurostatus.net), it is the combination of ambulation and reliance on aid that is affecting the determination of EDSS and not one of those variables independently. This was used as a rational for combining those two variables in the IRT analysis. Thus, a polychotomous variable with 11 categories, called *ambaid* was defined as following: *ambaid* = 0: ambulation ≥ 500 m and aid = 0; *ambaid* = 1: 300 m ≤ ambulation ≤ 499 m and aid = 0; *ambaid* = 2: 200 m ≤ ambulation ≤ 299 m and aid = 0; *ambaid* = 3: 100 m ≤ ambulation ≤ 199 m and aid = 0; *ambaid* = 4: 5 m ≤ ambulation ≤ 99 m and aid = 0 or ambulation ≥ 50 m and aid = 1 or ambulation > 120 m and aid = 2; *ambaid* = 5: 10 m < ambulation ≤ 49 m and aid = 1 or 10 m ≤ ambulation ≤ 120 m and aid = 2; *ambaid* = 6: ambulation ≤ 5 m and use of standard wheelchair; *ambaid* = 7: ambulation of few steps requiring aid to transfer and use standard wheelchair with assistance or motorized wheelchair; *ambaid* = 8: patient is wheelchair bound and capable of "many" self-care; *ambaid* = 9: patient is bed bound and capable of "some" self-care; *ambaid* = 10: patient is bed bound and not capable of any self-care.

The relationship between patients' response to an item and their level of disability, here called IRT disability, was modeled as ordered categorical data, and item characteristic curves (ICC) are used to quantify and visualize it (17). Observed scores for each EDSS item were modeled describing the probability of a given score as a function of patients' disability variable using a logistic model:

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

With b_j and a_j representing a point along the ICC of item j at which the probability of the positive response for a

dichotomous item is 50% and the slope of the ICC at that point, respectively, and D_i representing unobserved IRT disability of patient *i*. Cumulative probabilities for a score of M categories were modeled according to following equations (18):

$$\begin{aligned} P(Y_{ij}=0) &= 1-P(Y_{ij}>1) \\ P(Y_{ij}=k) &= P(Y_{ij}\geq k)-P(Y_{ij}\geq k+1) \\ P(Y_{ij}=M) &= P(Y_{ij}\geq M) \end{aligned}$$

Parameters a_j and b_j characterizing item specific parameters were modeled as fixed effects, while the IRT disability D was modeled as subject-specific random effect, assuming normal distribution with a mean of zero and fixed variance of 1, meaning that 68% of the population will be within the IRT disability range of (−1, 1). The assumed scale of D goes from $-\infty$ to $+\infty$, and it is relative to the studied population with a typical patient at baseline having an IRT disability of 0. In the case when scores of an item were not occurring in the available data, merging of scores with a closest observed score was performed.

Model development was conducted in five sequential steps: development of the baseline model; development of disease progression model based on placebo data; development of the exposure-response model based on data from patients on cladribine treatment; development of the covariate model; and model evaluation.

For the disease progression model, linear and non-linear (e.g., power and asymptotic) relationships were explored to describe the change in IRT disability over time. The disease progression model was then fixed to develop the exposure-response model. Linear, E_{max}, and sigmoidal E_{max} models were tested. Exposure-dependent and exposure-independent drug effects on disease progression were tested. A surrogate exposure measure based on cumulative dose (CumDose) and creatinine clearance (CL_{cr}) was used to drive the exposure-response relationship (19):

$$\text{Exps}_i = \frac{\text{CumDose} \times \text{Cl}_{\text{cr median}}}{\text{Cl}_{\text{cr}}}$$

After the drug model was developed, all model parameters were re-estimated simultaneously based on all available data.

Age and clinical covariates (disease duration and number of relapses in the preceding 12 months (EXNB)) were evaluated for their potential to account for the variability in baseline IRT disability and in slope of disease progression of the full model described above, using a full random effect models (FREM) approach (20). Covariates were introduced as observed variables, and their distribution was modeled as random effects. A full covariance matrix between random effects for parameters and covariates was estimated together with other model components. Coefficients for covariate-parameter relations were obtained from the ratio of covariance between parameter and covariate variability to the covariate variance.

Model discrimination between hierarchical models was primarily numerical and based on the likelihood ratio test of obtained objective function values (OFVs). For model selection, a significance level of $p < 0.05$ was used, with the degrees of freedom being equal to the difference in the number of parameters between two models.

Model evaluation was carried out through simulation-based diagnostics, mainly visual predictive checks (VPCs). Two hundred Monte Carlo simulation replicates of the original dataset with 95% prediction intervals were generated. Simulations were performed both on item level and on total score level. An algorithm was developed using the R program (21), to derive the total EDSS scores from individual item scores.

Calculation of Information Content

From the developed IRT model, the Fisher information for estimating a patient's IRT disability was calculated for each item constituting EDSS as minus the expectation of the second derivative of the log-likelihood. Subsequently, the information content for each item was computed for the studied population, and items were ranked according to the amount of information they contained.

Based on obtained item ranking, it was investigated whether a shorter version of the EDSS including only the most informative items would be able to identify patients with sustained progression equally well as the original scale. For this purpose, sustained progression was defined as a confirmed increase in EDSS after a period of at least 3 months with the increase defined in relation to the baseline, of ≥1.5 points if baseline EDSS was 0; ≥1 points if baseline EDSS was ≥1.0 and ≤4.5; ≥0.5 point if baseline EDSS was ≥5.0 (22). IRT disability status was determined based on all or on the subset of EDSS items, and then 200 simulations were performed using the developed model. The proportions of patients identified as progressing according to the original and shorten EDSS form were compared.

RESULTS

A total of 104,206 item level observations from 1319 patients were included in this analysis. A summary of study demographics is shown in Table I.

Model

Baseline Model

The final baseline model contained eight ordered categorical submodels in which a total of 42-item specific parameters were estimated (Supplemental Table 1). All parameters were successfully estimated with low uncertainty associated. The obtained ICCs are shown in Fig. 1, illustrating that a person with higher IRT disability has a higher probability of increased scores for each item. Noteworthy are the low slope parameter of 0.49 for *visual*, meaning that a large increase in IRT disability only yields a small increase in the probability for an increased score on this item, and the high slope parameter value of 3.5 for *ambaid* resulting in a high discrimination power in IRT disability around the b_i value of each score in this item.

Figure 1 also highlights an expected score larger than 0 for the *sensory*, *mental*, and *visual* item for individuals

Table I. Summary of Patient Baseline Characteristics

Variable	CLARITY placebo	CLARITY cladribine 3.5 mg/kg	CLARITY cladribine 5.25 mg/kg
Age—year			
Median	38	38	39
Range	18–64	18–65	18–65
Body weight—kg			
Median	69	66.1	66.9
Range	40–119.7	40–117	41.2–120
Sex no. (%)			
Male	149 (34.3)	135 (31.4)	143 (31.5)
Female	286 (65.7)	295 (68.6)	311 (68.5)
Disease duration			
Median	7.1	5.7	7.2
Range	0.4–39.5	0.3–42.3	0.4–35.2
Age of disease onset			
Median	29.2	29.1	29.2
Range	7.7–55.9	7.9–57.6	5.8–57.9
EDSS			
Median	3	2.5	3
Range	0–5.5	0–6	0–5.5
EXBN			
Median	1	1	1
Range	1–5	1–5	0–5

EXBN annualized number of relapses 2 years prior to the study

considered healthy (IRT disability = −4); this can be explained by non-MS-related impairment of those functions, as those are not MS-specific symptoms.

For most of the items, score of 0 is the most frequently observed score. Also, the probability for a score of 0 drops quickly as the IRT disability increases, except for *ambaid* item where probability of having score of 0 remains 100% with increasing IRT disability until a certain level of IRT disability is reached. This is in line with common clinical knowledge that only patients with advanced stage of the disease will start experiencing impaired ambulation (EDSS higher than score of 4).

Probability curves for different scores of some items (e.g., *mental*) overlap over a range of IRT disability levels, indicating that a specific item does not differentiate well between those scores for a given range of IRT disability.

Figure in supplemental 1 shows that the frequency with which the score is observed at baseline is captured within the 95% prediction interval of the model.

Figure 2 depicts the relationship between EDSS and the corresponding IRT disability levels for each patient in the dataset at baseline. This figure shows that although there is a trend of increasing EDSS scores with increasing disease states, each EDSS score corresponds to a wide spectrum of underlying IRT disability scores and vice-versa.

Disease Progression Model Based on Placebo Data

Disease progression in patients receiving placebo treatment was best described with a power model with an estimated IRT disability at baseline (D_0), disease progression rate (α), and a power constant (pwr):

$$D_i(t) = D_{0,i} + \alpha_i \times t^{\text{pwr}}$$

A significant positive correlation of 0.59 ($p < 0.001$) was observed between baseline IRT disability and the disease progression rate, indicating that patients with higher IRT disability at baseline are likely to progress faster. Positive slope of disease progression, significantly different from zero ($p < 0.001$) was estimated. The estimated disease progression rates were on IRT disability scale. Simulations were also performed which translates those results to the EDSS scale, and according to these simulations, the typical patient in this dataset receiving placebo treatment will progress 0.16 EDSS points over 2 years.

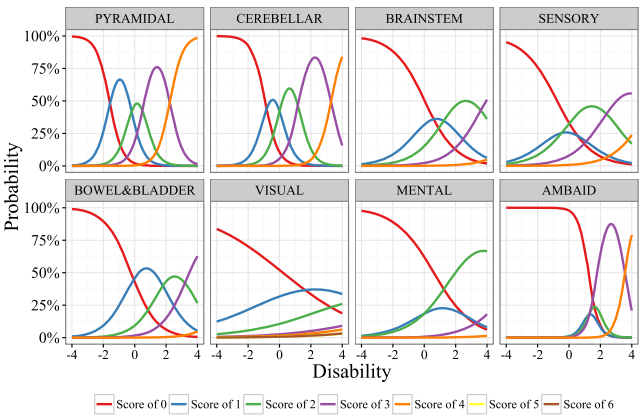

Fig. 1. Item characteristic curves per item: probability of occurrence of each score as a function of IRT disability at baseline (with positive values of disability indicating a higher disability than the disability of the typical patient)

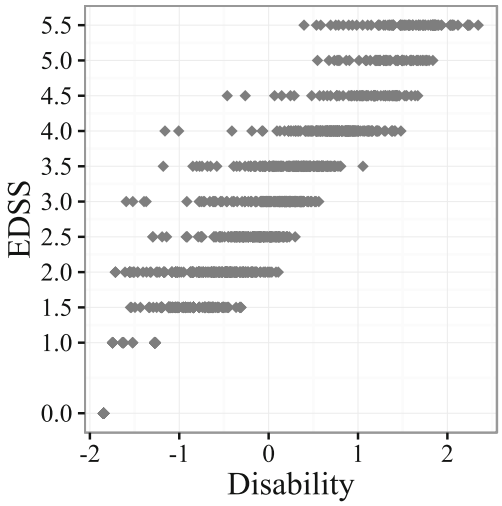

Fig. 2. Observed EDSS scores and estimates of IRT disability at baseline

Exposure-Response Model

The disease progression as well as the drug effect in patients receiving cladribine treatment was best described according to the following equation:

$$D_i(t) = D_{0,i} + \alpha_i \times t^{\mathrm{pwr}} \times (1 - \mathrm{Eff}D) - \frac{\mathrm{Emax} \times \mathrm{Exps}_i}{\mathrm{Exps}_{50} + \mathrm{Exps}_i}$$

with IRT disability at baseline (D_0), disease progression rate (α), power constant (pwr), maximal exposure-dependent drug effect (Emax), exposure needed for half maximal effect (Exps_{50}), and constant exposure-independent drug effect (Eff$D$).

The effect of cladribine on IRT disability was best described using both exposure-dependent and exposure-independent drug effects. The final model suggests that cladribine treatment significantly ($p < 0.001$) slows disease-progression rate, with a 20% decrease in disease progression rate compared to placebo, irrespective of exposure in the investigated cumulative dose range (20–600 mg). The model also describes an exposure dependent decrease in IRT disability in patients treated with cladribine tablets with a cumulative dose of 407 mg being needed for half maximal (exposure-dependent) effect in a typical patient, which would translate for a typical patient receiving a typical dose of 240 mg in 45% reduction of disease progression.

Covariate Model

Covariate analysis revealed that baseline IRT disability was correlated with age, duration of disease, and EXNB by coefficients of 0.027, 0.037, and 0.075, respectively. This means for instance that a typical patient of 58 years, who is 20 years older than the population's mean of 38 years, will have a baseline IRT disability that is 0.54 (i.e., 20*0.027) units higher on the disability scale, then the IRT disability of a typical patient with the mean age in this population. Similarly, there is a 7.5% increase in baseline IRT disability per number of relapses (>1) in the year previous to the study. Coefficients for covariates effects on the slope of disease progression were 0.0053, 0.0054, and 0.05 for age, duration of disease, and EXNB, respectively.

Table II. Population Parameter Estimates from the Final Model

	Parameter estimates (RSE%[a])
Disease progression slope	0.087 (6.5)
Disease progression power	0.707 (3.6)
ω^2 Slope	0.199 (6.9)
Emax exposure-dependent drug effect	0.171 (5.3)
ω^2 Emax	2.20 (9.2)
Exp50 exposure-dependent drug effect	406.8 (4.8)
Constant exposure-independent drug effect	0.228 (4.5)
Correlation dis_0/slope	0.113 (18.6)

[a] Relative standard errors from bootstrap ($n = 100$) in NONMEM

Final parameter estimates for this model are shown in Table II. RSE values in all model parameters were below 20%, meaning the parameters could be estimated from the data with high certainty.

Model Evaluation

Simulations from the developed IRT model were performed in order to demonstrate the predictive ability of the final model. The item level VPC, with the example of brainstem item, in Fig. 3 shows that for the duration of the trial there is a good agreement between observed and predicted scores. VPCs for the remaining seven items can be found in Supplemental 2. Moreover, Fig. 4 shows the observed and model-predicted total EDSS scores coincide over time for each treatment arm.

Calculation of Information Content

Fisher information content as a function of IRT disability is shown for each item in Fig. 5. The shaded area indicates the

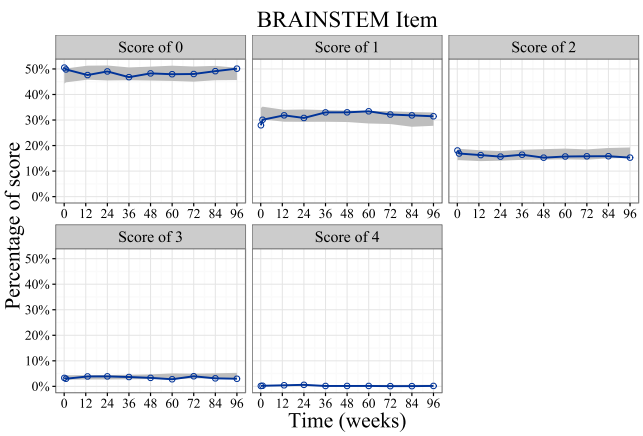

Fig. 3. Visual predictive checks (VPCs) describing the time-courses of each score for the brainstem item. Median (*blue solid line*) of the observed data is compared to the 95% prediction interval (*gray shaded area*) for the simulated data

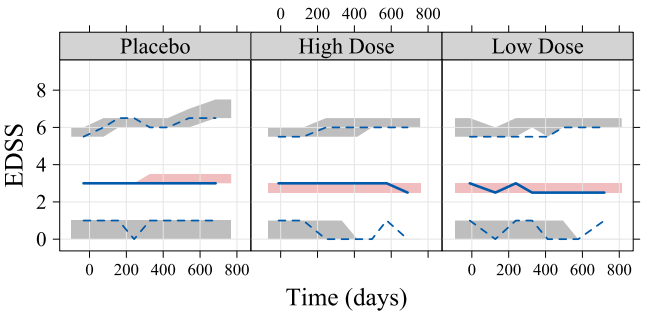

Fig. 4. VPC for the final model describing the change in EDSS *vs* time, stratified by treatment arm. Median (*blue solid line*), 2.75th, and 97.25th percentiles (*blue dashed lines*) of the observed data are compared to the 95% confidence intervals for the simulated data's 2.75th, 97.25th percentiles (*gray shaded areas*), and median (*red shaded area*)

interval of IRT disability containing 95% of the study population. The information content varies considerably across items as is evident from differences in the location of the maxima of the information curves as well as differences in their amplitude. As an item is most informative around its *b* value, the most "difficult" parameter, *ambaid*, contains most information concerning the highly disabled subgroup of studied patients' population.

Items were ranked based on their expected Fisher information for the range of IRT disabilities in the current study population. The *cerebellar* item was found to contain most of information, followed by *pyramidal* and *ambaid* items. As shown in Table III, four (*cerebellar*, *pyramidal*, *ambaid*, *bowel, and bladder*) out of eight items contained 80% of information for the given range of IRT disabilities. Noteworthy is the *visual* item that is found to contain least information among EDSS items, which is also visualized by the flat ICCs for this item in Fig. 1.

From this, the EDSS4 scale, based only on the four most informative items, was derived and then evaluated by computing the ratios of patients classified as progressing for EDSS4 (shortened version) and EDSS8 (original version). Based on simulations, proportions of progressing patients were very similar independent of used scale (95% CI [0.92, 1.06]).

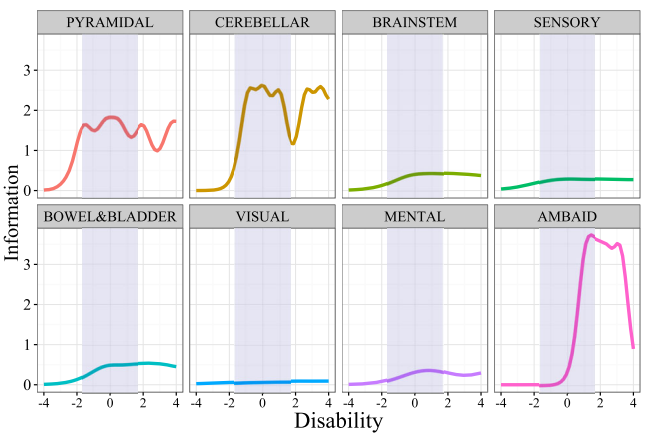

Fig. 5. Information content for EDSS items *versus* IRT disability. The shaded areas indicate the disability range for 95% of studied population

Table III. Ranking of EDSS Components by Information Content in Studied Population

Item	Information	% total	Cumulative % total
Cerebellar	2.18	33.9	33.9
Pyramidal	1.62	25.3	59.2
Ambaid	1.14	17.8	77
Bowel and bladder	0.45	7	84
Brainstem	0.37	5.8	89.8
Cerebral	0.29	4.5	94.3
Sensory	0.28	4.4	98.7
Visual	0.08	1.2	100

DISCUSSION

Using the data from a phase III clinical trial, IRT methodology has been successfully implemented for the first time to model EDSS in patients with RRMS. The model reported here was developed using data from a clinical trial investigating the effect of cladribine tablets on RRMS. The drug effect model is certainly specific to cladribine, but the implementation of IRT methodology to EDSS as well as the description of time-course of disease progression has broader applicability, beyond cladribine tablets.

Traditional approaches to analyze questionnaire-based scales generally disregard the underlying nature of the data and usually regard only summary scores. In the past, EDSS has been modeled either as a continuous variable (19) or as an ordered categorical variable with considerable simplification of the scale (20 categories combined into 5–6 categories) (23, 24). Instead of modeling changes in the composite score over time, application of IRT allows derivation of underlying/unobserved latent variable from observed subscores and model the changes in that latent variable over time. The IRT methodology has been applied here to order categorical data, but it has been shown by Ueckert *et al.* that it is equally suitable for other types of non-continuous data, such as binary or count data (12).

The effectiveness of therapeutic interventions can be determined, only if accurate quantification of disease severity is possible. Central to the patient, the most important therapeutic aim of any disease modifying treatment of MS is to prevent or postpone long-term disability. In phase III trials, various surrogate measures such as relative reduction in annualized relapse rate and risk of 3-month sustained progression have been used as predictors for this disability, but there is limited evidence that those changes reflect true irreversible accumulation of disability at long-term scale (25). Both analyses of CLARITY trial data, our IRT analysis of EDSS subitems and traditional statistical analysis of time to sustained progression have found that cladribine tablets have an effect on the studied endpoints "disease progression" and "risk of 3-month sustained progression of disability". However, using time to first confirmed disability progression as an endpoint in MS clinical development does not allow for a description of disease progression, as we know that disability progression does not stop after the first event. In contrast, our model can be used to understand time-course of disease and effect of the treatment, and a role of the individual components of EDSS. It can also be used for clinical trial simulations.

Another aspect of the slowly progressing and highly variable nature of the disease is that patients may remain at the same score for a prolonged period of time. According to the model developed by Savic *et al.*(19) for instance, a typical patient will experience 0.14 EDSS units increase in disease severity over 2 years on placebo treatment and only 0.024 points increase when treated with a 3.5 mg/kg cumulated dose of cladribine. Using IRT, disability may be determined more accurately than with the composite EDSS score. Results shown in Fig. 2 for instance reveal that each EDSS score corresponds to a wide range of underlying IRT disabilities, indicating that total scores are relatively imprecise measures of underlying IRT disability. Better quantification of disease severity will also improve our assessment of disease progression and treatment effect. Comparable results were obtained for ADAS-cog score in Alzheimer's disease (11).

Full EDSS assessment takes over 40 min to be performed by a neurologist, hampering its use in everyday clinical practice. Increased efficiency could be achieved with optimal selection of the most informative subset of items. Here, we use Fisher information as a measure of item information content as it directly relates to the expected variance of the individual latent variable estimates. Conceptually, we are able to choose the items that have the largest signal to noise ratio, i.e., where a functional change relates most directly to a change in disease state. We have shown that 80% of the information about underlying disease status in MS, in the studied population, is quantified in only four of the eight EDSS items, namely *cerebellar*, *pyramidal*, *ambaid*, and *bowel and bladder*. Simulations have demonstrated that our proposed shortened scale performs equally well as the full EDSS scale when it comes to determining a clinically meaningful measure, the ratio of patients experiencing a 3-month sustained progression. With this example, we have just demonstrated how a rational subselection of items can be made if one wants to simplify the test. This could be taken even further by turning it into a dynamic process—the answer to the first item evaluation directs which item to investigate next.

MS affects functional systems of the EDSS differently as identified by Healy *et al.* (25). They have demonstrated that the time to sustained progression varied widely across the EDSS items; it was the fastest for the pyramidal and sensory scales and the slowest for brainstem and visual scales. Identification of subgroups of patients more likely to experience substantial worsening of the disease, by focusing on specific sensitive items, will increase the difference in drug effect between groups, if one is in fact present. Thus, the insight into information content on item level achieved through IRT analysis could be used as a valuable tool, in combination with other study enrichment strategies.

Despite its weaknesses, the extensive use of EDSS in patients with RRMS is likely to be continued. Current treatments have been authorized based on clinical trials using EDSS as one of the endpoints, and EMA requires new therapies to be compared to existing ones by using the same outcome measures to demonstrate their effectiveness (26). Also, on the individual patient level, there is a need for continuity in use of outcome measures in order to ensure the long-term records of disease severity (27). Moreover, the clinical course of RRMS can vary tremendously, and it is likely that different outcome measures are demanded in different stages of the disease (3). Establishing and quantifying the relationship between different outcome measures has been proven challenging in the past (25, 28).

One of the ways to enable the direct comparison of results obtained on different scales for disease severity would be by the application of IRT. Ueckert *et al.* have demonstrated the possibility of jointly analyzing different ADAS-cog variants, without any recalculation or normalization of measured scores. In the field of MS, this approach could be utilized to bridge between the diversity of scores that are used for quantification of disease progression, as IRT disability levels of patients can be easily compared once outcome measures on the different scales have been mapped to overall IRT disability. This approach will also allow evaluation of performance of one assessment method relative to another.

CONCLUSION

Accurate quantification of disease status and description and prediction of disease progression is essential for drug development. For chronic diseases with slow progression such as multiple sclerosis, this is especially pertinent, due to the high costs of long-term clinical trials required to establish treatment efficacy. This study has illustrated that IRT modeling is specifically suitable for this purpose in phase 3 studies on RRMS, by integrating EDSS item level data in a meaningful manner instead of aggregating information by deriving a composite score.

ACKNOWLEDGMENTS

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115156, resources of which are composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution. The DDMoRe project is also supported by financial contribution from Academic and SME partners. This work does not necessarily represent the view of all DDMoRe partners.

COMPLIANCE WITH ETHICAL STANDARDS

Trial protocols, amendments, and subject informed consent forms were reviewed by a national, regional, or investigational site ethic committees or an Institutional Review Boards. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki and the International Conference on Harmonization Tripartite Guidelines for Good Clinical Practice.

Conflicts of Interest The authors declare that they have no conflict of interests.

REFERENCES

1. Goldman MD, Motl RW, Rudick RA. Possible clinical outcome measures for clinical trials in patients with multiple sclerosis. Ther Adv Neurol Disord. 2010;3:229–39.
2. Tullman MJ. Overview of the epidemiology, diagnosis, and disease progression associated with multiple sclerosis. Am J Manag Care. 2013;19:S15–20.
3. Amato MP, Portaccio E. Clinical outcome measures in multiple sclerosis. J Neurol Sci. 2007;259:118–22.
4. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology. 1983;33:1444–52.
5. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain. 2000;123(Pt 5):1027–40.
6. Weinshenker BG, Bass B, Rice GP, Noseworthy J, Carriere W, Baskerville J, *et al*. The natural history of multiple sclerosis: a geographically based study. 2. Predictive value of the early clinical course. Brain. 1989;112(Pt 6):1419–28.
7. Sharrack B, Hughes RA. Clinical scales for multiple sclerosis. J Neurol Sci. 1996;135:1–9.
8. Chang CH, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof. 2005;28:264–82.
9. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, *et al*. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15:708–15.
10. Baker FB. The basics of item response theory. 2001.
11. 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;8:288–94.
12. Ueckert S, Plan EL, Ito K, Karlsson MO, Corrigan B, Hooker AC, *et al*. Improved utilization of ADAS-Cog assessment data through item response theory based pharmacometric modeling. Pharm Res 2014.
13. Giovannoni G, Comi G, Cook S, Rammohan K, Rieckmann P, Soelberg Sorensen P, *et al*. A placebo-controlled trial of oral cladribine for relapsing multiple sclerosis. N Engl J Med. 2010;362:416–26.
14. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, *et al*. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol. 2001;50:121–7.
15. Beal SL, Sheiner LB, Boeckmann A, Bauer RJ. NONMEM user's guides (1989–2009). Ellicott City: Icon Development Solutions; 2009.
16. Keizer RJ, Karlsson MO, Hooker A. Modeling and Simulation Workbench for NONMEM: tutorial on pirana, PsN, and Xpose. CPT Pharmacometrics Syst Pharmacol. 2013;2:e50.
17. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007;16 Suppl 1:5–18.
18. Kjellsson MC, Zingmark PH, Jonsson EN, Karlsson MO. Comparison of proportional and differential odds models for mixed-effects analysis of categorical data. J Pharmacokinet Pharmacodyn. 2008;35:483–501.
19. Savic RM, Munafo A, Karlsson MO. Disease progression model for multiple sclerosis and effect of cladribine tablets therapy on clinical endpoints. San Diego: ACOP; 2011.
20. Karlsson MO. A full model approach based on the covariance matrix of parameters and covariates. PAGE 2012.
21. Team RC. R: A Language and Environment for Statistical Computing. 2014.
22. Ellison GW, Myers LW, Leake BD, Mickey MR, Ke D, Syndulko K, *et al*. Design strategies in multiple-sclerosis clinical-trials. Ann Neurol. 1994;36:S108–12.
23. Mandel M, Mercier F, Eckert B, Chin P, Betensky RA. Estimating time to disease progression comparing transition models and survival methods—an analysis of multiple sclerosis data. Biometrics. 2013;69:225–34.
24. Healy B, Chitnis T, Engler D. Improving power to detect disease progression in multiple sclerosis through alternative analysis strategies. J Neurol. 2011;258:1812–9.
25. Healy BC, Engler D, Glanz B, Musallam A, Chitnis T. Assessment of definitions of sustained disease progression in relapsing-remitting multiple sclerosis. Mult Scler Int. 2013;2013:189624.
26. Draft of Guideline on clinical investigation of medicinal products for the treatment of Multiple Sclerosis. 2012.
27. Mumford CJ, Compston A. Problems with rating scales for multiple sclerosis: a novel approach—the CAMBS score. J Neurol. 1993;240:209–15.
28. Confavreux C, Vukusic S, Moreau T, Adeleine P. Relapses and progression of disability in multiple sclerosis. N Engl J Med. 2000;343:1430–8.