
Research Article

Modeling a Composite Score in Parkinson's Disease Using Item Response Theory

Gopichand Gottipati,¹ Mats O. Karlsson,¹ and Elodie L. Plan^{1,2}

Received 15 November 2016; accepted 7 February 2017; published online 28 February 2017

Abstract. In the current work, we present the methodology for development of an Item Response Theory model within a non-linear mixed effects framework to characterize the longitudinal changes of the Movement Disorder Society (sponsored revision) of Unified Parkinson's Disease Rating Scale (MDS-UPDRS) endpoint in Parkinson's disease (PD). The data were obtained from Parkinson's Progression Markers Initiative database and included 163,070 observations up to 48 months from 430 subjects belonging to *De Novo* PD cohort. The probability of obtaining a score, reported for each of the items in the questionnaire, was modeled as a function of the subject's disability. Initially, a single latent variable model was explored to characterize the disease progression over time. However, based on the understanding of the questionnaire set-up and the results of a residuals-based diagnostic tool, a three latent variable model with a mixture implementation was able to adequately describe longitudinal changes not only at the total score level but also at each individual item level. The linear progression rates obtained for the patient-reported items and the non-sided items were similar, each of which roughly take about 50 months for a typical subject to progress linearly from the baseline by one standard deviation. However for the sided items, it was found that the better side deteriorates quicker than the disabled side. This study presents a framework for analyzing MDS-UPDRS data, which can be adapted to more traditional UPDRS data collected in PD clinical trials and result in more efficient designs and analyses of such studies.

KEY WORDS: Parkinson's disease; Item Response Theory; Movement Disorder Society (sponsored revision) Unified Parkinson's Disease Rating Scale; disease progression.

INTRODUCTION

Parkinson's disease (PD) is a chronic neurodegenerative disorder affecting the central nervous system. The pathophysiological manifestation associated with the motor deficits in PD is the progressive loss of dopaminergic neurons of pars compacta of substantia nigra resulting in a significant decrease in the dopamine levels. Availability of physiological biomarkers or neuroimaging markers that can give indications about the disease status is one of essential prerequisites for studying disease progression. However, lack of such definitive markers in PD (1) has been a major challenge in the development of newer therapies. Among a number of rating scales used for the assessment in PD, the Unified

Parkinson's Disease Rating Scale (UPDRS), originally developed three decades ago, is still the mainstay. The composite score of the different components of this rating scale reflects the severity of the disease, i.e., higher score is indicative of a more severe disease.

While the original UPDRS emphasized on gradation between marked and severely disabled patients, more recent scientific advances lay emphasis on early prognosis of the disease and the need for developing therapies for early intervention and neuroprotection (2). In order to address this aspect (among others, such as correcting inconsistencies and resolving ambiguities, etc.), the Movement Disorder Society (MDS) sponsored a revision of the UPDRS version to adapt the scale such that it can detect smaller changes in the disease early and measure milder deficits. Consequently, the MDS-UPDRS version focuses on a broader but among the lower ranges in disability (such as differentiation of slight from mild deficits) rather than differentiating the gradations in advanced disability (such as severe from marked deficits) (2). The MDS-UPDRS questionnaire consists of four parts, namely, non-motor and motor aspects of experiences of daily living, motor examination, and motor complications. It has a total of 68 items, among which 2 items are binary, i.e., have

Electronic supplementary material The online version of this article (doi:10.1208/s12248-017-0058-8) contains supplementary material, which is available to authorized users.

¹Department of Pharmaceutical Biosciences, Uppsala University, Box 591, 75124, Uppsala, Sweden.

²To whom correspondence should be addressed. (e-mail: elodie.plan@farmbio.uu.se)

0/1 responses, and 66 are ordered categorical responses, most of which are rated between five categories, ranging from 0 indicating normal or no impairment to 4 indicating severe impairment, except for one item, “Hoehn and Yahr Stage” which has six categories (2). The two binary items, i.e., with two potential responses—“yes”/“no,” correspond to whether dyskinesias were present during the examination (item 60) and if those movements interfered with the ratings (item 61), the latter being evaluated only for those who answered “yes” previously.

Traditionally, analyses of PD trials are performed using the composite score, e.g., “total” UPDRS score (sum of parts I, II, and III). However, occasionally, only a portion(s) of the rating scale (e.g., only the motor component of UPDRS scale), dependent on the specific range of disease severity of the patient cohort (3,4), may be used. The ranges in the resulting scores may therefore vary greatly, making the comparability across the rating scales difficult and leveraging/integration of knowledge from multiple sources cumbersome. Furthermore, treating the total score as continuous may not be appropriate considering it results from summing up individual answers to questions of varying difficulty and category range.

Item Response Theory (IRT) has been reported to be a promising approach compared to the classical methods (5) in the development and validation of tests in patient-reported outcomes (PRO) research (6,7). It has also been extensively used in computerized adaptive educational testing applications (8). IRT is a statistical framework consisting of mathematical models that describe the relationship between an individual’s underlying *latent* (or hidden) variable to the pattern of responses to the items on the assessment scale and such a relationship is described by the Item Characteristic Curves (ICC). The recent application of the IRT methodology in Alzheimer’s disease (9–11), where it was shown to be a more advantageous approach than the traditional analysis of the composite score because of an improved utilization of the data at an individual item level, was later applied and further explored in other disease areas such as multiple sclerosis (12) and schizophrenia (13).

In this manuscript, we aimed to further the understanding of the disease progression characteristics in PD by describing the longitudinal changes in the MDS–UPDRS data using the IRT approach.

Although the questionnaire is designed to diagnose and evaluate the overall disease status, each subscale of the questionnaire reflects a specific aspect of the disease, e.g., motor and non-motor-related symptoms. Therefore, additional objectives of this work were to explore whether the scale items relate to one or several traits (i.e., the utility of multiple latent variables in the IRT framework) and to contribute to the development of appropriate diagnostic tools to assess these modeling considerations.

MATERIAL AND METHODS

Data

The data were obtained from Parkinson’s Progression Markers Initiative (PPMI) database (<http://www.ppmi-info.org/data>) version available as of November 2014. The

observations used in this work consisted of individual item level MDS–UPDRS records up to 48 months from 423 subjects belonging to the *De Novo* PD cohort. The subjects in this cohort corresponded to patients who were diagnosed with PD for 2 years or less, did not take PD medications (e.g., levodopa, dopamine agonists, MAO-B inhibitors, amantadine, etc.) for more than 60 days prior to the baseline visit, and were not currently on and did not expect to require PD medications within at least 6 months from the baseline visit. MDS–UPDRS observations were collected during the visits every 3 months up to 12 months and every 6 months after that, up to 48 months. The demographics of this cohort are listed in Table I and further information about the design aspects and the inclusion/exclusion criteria can be found at http://www.ppmi-info.org/wp-content/uploads/2015/01/PPMI-AM9-NOV-1-2014_clean.pdf. For the motor examination (part III) of the questionnaire, which considers if the subjects are receiving medications for treating the symptoms of PD and their clinical states, only the pre-dose assessments were included in the dataset. The few individuals who did not have information of which cohort they belonged to at the baseline visit (enrollment) were assigned to the cohort determined at the screening visit (1.5 months before the baseline visit).

Modeling

The IRT modeling approach was used to describe the relationship between the probability of the subjects’ responses for each item of MDS–UPDRS assessment and an unobservable (latent) disease status, termed as “disability.” Overall, the model development using a single latent variable

Table I. Demographics and PD Characteristics of the DeNoPD Cohort at Baseline Visit

Variable	DeNoPD cohort (<i>N</i> = 423)
Age (years)	
Mean (SD)	62 (9.7)
Range	(33, 85)
Gender <i>N</i> (%)	
Male	277 (65)
Female	146 (35)
Race— <i>N</i> (%)	
White	391 (92)
Black/African American	6 (1)
Asian	8 (2)
Other	18 (4)
Disease duration since diagnosis (months)	
Mean (SD)	6.7 (6.5)
Range	(0.4, 35.8)
MDS–UPDRS scores—mean (SD)	
Total score	32.4 (13.1)
Part I	5.6 (4.1)
Part II	5.9 (4.2)
Part III	20.9 (8.9)
Handed— <i>N</i> (%)	
Left	38 (9)
Right	375 (89)
Mixed	9 (2)

consisted of two components: (i) estimation of ICC parameters, discussed below, and (ii) characterization of the longitudinal changes in the “disability” as a consequence of disease progression.

Item Response Probabilistic Model

The item response model parameters were classified into item-specific parameters namely, a_j , b_j (described in detail below) for an item j and subject (denoted i)-specific parameter—“disability,” D_i . The probability that the subjects’ response was at least k (ranging between 0 and maximum of K), i.e., the cumulative probability, was modeled using a proportional odds, ordered categorical model, also referred to as 2PL (2 parameter logit) in the IRT literature (14). The functions used to characterize the ICC relationships and the probability of observing each individual score k , up to a maximum of K (i.e., either 4 or 5) were calculated by:

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

$$\begin{cases} P(Y_{ij} = 0) = 1 - P(Y_{ij} \geq 1) \\ P(Y_{ij} = k) = P(Y_{ij} \geq k) - P(Y_{ij} \geq k + 1) \\ P(Y_{ij} = K) = P(Y_{ij} \geq K) \end{cases}$$

where Y_{ij} is the subjects’ observed response to i th item with a response of at least k , a_j is the slope or discrimination parameter, and b_{jk} is the difficulty parameter, representing the disability at which there is a 50% probability of obtaining a positive response for that item. The initial estimates for the difficulty parameter (at individual category level) were constrained (i) to be non-decreasing for the higher score categories within each item (i.e., $b_{j,k+1} \geq b_{j,k}$); (ii) to an upper bound of 50 for all the score categories except the first (i.e., $b_{j,k=1}$); and (iii) to a fixed value of 50 for items in which there was no observed response within a certain category, usually the higher categories (e.g., $b_{j,k} = 3$ or 4). The implementation of the latter two constraints provided numerical stability in the model building process.

For the binary items, the probability of responding “yes” (i.e., response of 1) was also modeled as a function of disability using a 2PL model (14):

$$P(Y_{ij} = 1) = \frac{e^{a_j(D_i - b_j)}}{1 + e^{a_j(D_i - b_j)}}$$

$$P(Y_{ij} = 0) = 1 - P(Y_{ij} = 1)$$

where the item-specific parameters a_j and b_j are the discriminatory and difficulty parameters, respectively, as described above.

IRT Latent Variable Model

The item-specific parameters, a_j , and b_{jk} , characterizing the ICCs were modeled as fixed effects, while the subject-specific “disability” parameter (D_i) was modeled as a random effect. This “disability” scale is a hypothetical construct; it can take values between $-\infty$ to $+\infty$ and, at baseline, it was

assumed to follow a normal distribution with a mean of zero and a variance of 1 ($N(0, \omega^2 = 1)$).

When investigating longitudinal changes, the disease progression was implemented on the disability scale using a linear model (4) as a function of time since baseline visit:

$$D_i(t) = D_{i,0} + \text{Slope}_i \times t$$

where $D_{i,0}$ is a subject-specific random effect ($\theta_i^{\text{baseline}} + \eta_{\text{baseline}}$, assumed to be centered around a typical value of 0 and variance fixed to 1) characterizing the disability at baseline and Slope_i is the rate of disease progression, also a subject-specific parameter modeled through random effects ($\theta_i^{\text{slope}} + \eta_{\text{slope}}$). In order to facilitate the fixation of variance of disability at baseline (to 1) while allowing for estimating its correlation with the random effect on the slope, a transformation was necessary and it was achieved by implementing the Cholesky decomposition matrix.

Simultaneous Vs. Sequential Parameter Estimation Processes

The simultaneous approach consisted of a single step in which the ICCs and the longitudinal changes were estimated simultaneously. On the other hand, the sequential approach involved a two-step process. In the first step, the dataset was modified such that observations at each of the subject’s visits (i.e., from baseline to each of the scheduled visit) were treated as though they were of separate individuals and the ICCs were estimated at baseline as the reference and a shift (parameter) for the post-baseline distribution of disability. In the second step, the dataset was reconciled, the ICCs were fixed to their values obtained in the first step, and then, the longitudinal model parameters were estimated.

Furthermore, eta distributions, namely, t-distribution and box-cox transformation, were investigated for alternative distributions of the disability at baseline using the final models obtained from both approaches as means to validate the assumption of normality.

Model Building and Evaluation

All the analyses were performed using the software NONMEM version 7.3 (15). The parameter estimation was carried out using second-order conditional estimation with Laplacian approximation. Model selection between the alternative (nested) models was based on likelihood ratio test of the obtained OFV at a significance level of $p < 0.05$ and akaike information criteria (AIC) was used for evaluating non-nested models.

The fit obtained based on the model-predicted ICCs for each item were compared to the fit obtained from a generalized additive model (GAM) using a cross-validated cubic spline as a smoothing function in R (16). Additionally, the final models from both the approaches were evaluated using simulation-based diagnostics comparing the ICCs and the predicted individual disability estimates to simulate responses (200 replicates) to the observed responses (see details in Appendices III and VI).

Further simulation-based diagnostics were performed by visual predictive checks (VPCs) using PsN tools (17). Monte Carlo simulations of 200 datasets were generated using the final models, and 95% prediction intervals were obtained around the median, the 2.5th and the 97.5th percentiles at individual item level (the

proportion of the subjects within each score category) as well as at total score level (sum of the scores from all the MDS-UPDRS assessments) level and compared with the same metrics calculated from the original data.

The correlations between the item responses were also explored by calculating the residuals (RES) as described below

$$RES_{ij} = DV_{ij} - E_{ij}$$

$$E_{ij} = P(1) * 1 + P(2) * 2 + \dots + P(k) * k + \dots + P(K) * K$$

$$E_{ij} = \sum_{k=1}^K P(k) * k$$

Where DV_{ij} is the observed response of i th individual for j th item and E_{ij} is the respective weighted prediction from the ICCs of individual probabilities (as shown in Fig. 1) calculated based on the model-predicted disability (D_i) of each subject. The correlation matrix of residuals across the items was then plotted (Fig. 2), with correlation values ranging from -1 (indicated by blue color) to $+1$ (indicated by red).

Single Vs. Multiple Latent Variable(s)

The pattern of correlations observed from the residuals plot with a single latent variable was quintessential in informing the model building process in terms of visual diagnostics and formed the basis for exploring multiple latent variables and gave insight not only on the number but also the type(s) of variables. The IRT model building and evaluation using the multiple latent variables was performed in an identical manner as described with the single latent variable model. However, based on the results obtained from the single latent variable, only the simultaneous approach was explored. Additionally, the linear disease progression was implemented separately on each of the latent variable that was explored.

Four latent variables were tested, one each for (i) PR—for the items (#1–26) which characterized the Patient-Reported (self-administered) Responses, (ii) RSR—for the items (# 30, 32, 34, 36, 38, 40, 42, 50, 52, 54, 56) which characterized the Right Side Responses, (iii) LSR—for the items (# 31, 33, 35, 37, 39, 41, 43, 51, 53, 55, 57) which characterized the Left Side Responses, and (iv) NSR—for the rest of the items (27–29, 44–49, 58–68) which characterized the Non-Sided Responses, i.e., neither the left nor the right side. Furthermore, a three latent variables model with PR,

NSR, and SR—for Sided Responses, a common latent variable for the items (30–43 and 50–57) that evaluated the right and left sides, was explored with a mixture model. This was implemented in NONMEM using the \$MIXTURE subroutine to separate out and estimate the size (or proportion, as a fixed effect parameter) of the two subpopulations—one whose most disabled side initially (i.e., at baseline) was the right side or the other, whose most disabled side initially (i.e., at baseline) was the left side. Furthermore, a fixed effect shift parameter (associated with an exponential eta) was implemented to reflect the lower disability for the items assessing the initially better side, based on the assignment of the individual to one of the two subpopulations by the mixture. Lastly, owing to the ambiguity in the description of the first six items, as to whether they were self-administered (PR) or evaluated by the clinical investigator, the multiple latent variable models described above were also tested by re-assigning them to NSR instead of PR.

For the longitudinal changes, all the possible correlations between random effects, i.e., the distribution of disability at baseline and the slope of the disease progression for each of the latent variable, were investigated while implementing the Cholesky decomposition matrix as mentioned above.

RESULTS

Data

Overall, there were 163,070 observations from 430 individuals. The distribution of these responses is shown in Fig. 3. The ordinal items show a diverse pattern in the distribution of responses, e.g., items for which the responses were mostly in the lower score range (2 or less), i.e., skewed to the left (lower scores suggesting lower disability), which was the case for majority of the items, as well as a few items for which responses seemed to be normally distributed. This pattern of observed responses for the items in MDS-UPDRS scale in the current dataset is plausible because the subjects belong to DeNoPD cohort, who had been diagnosed with PD for 2 years or less at screening. The binary items, since only recorded by subjects who responded that they had dyskinesias (item 60, $n = 17$ observations) and were further evaluated if the movements interfered with the ratings (item 61), had, as expected, very low frequencies of responses. It can also be observed that items 63 (*time spent with dyskinesias*) and 64

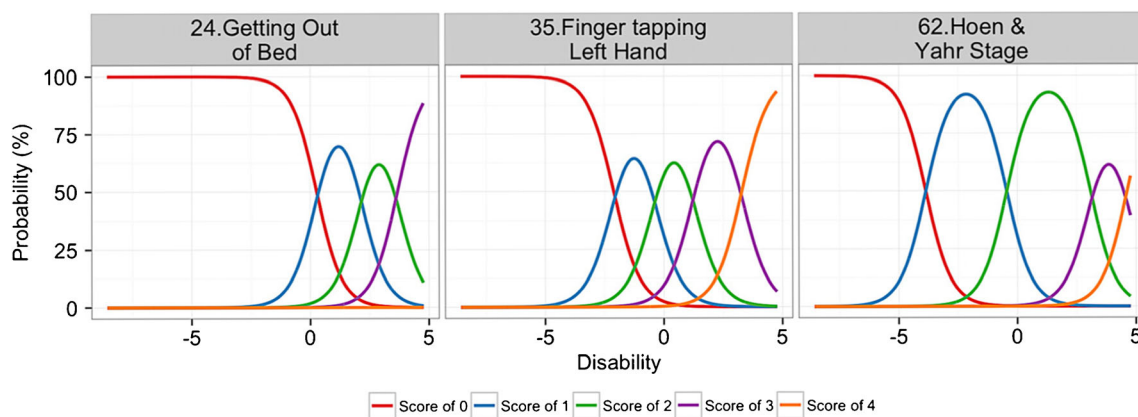


Fig. 1. Item characteristic curves showing the individual probability of obtaining scores in each category for items 24, 35, and 62, representative of most informative items within latent variables PR, SR, and NSR, respectively

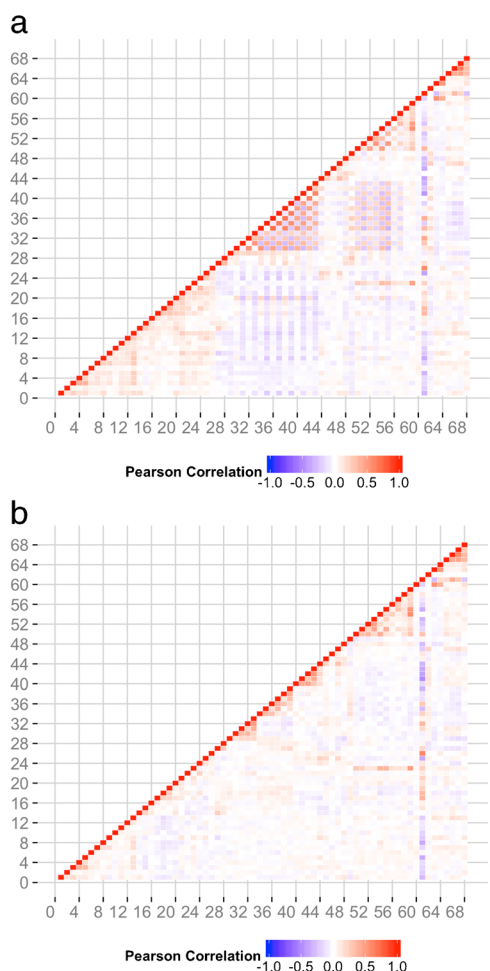


Fig. 2. Correlation between the residuals obtained using a single latent variable IRT model in panel **a** (top) and a three latent variables IRT model with mixture in panel **b** (bottom), across the 68 items of the score

(functional impact of dyskinesias), that also rely on item 60, showed category 0 responses mostly.

Single Latent Variable Model

All the item-specific parameters were successfully estimated with both the approaches. However, the simultaneous approach showed a better OFV value (268,092.95) than the sequential approach (268,159.99); final estimates of the former approach are provided in Appendix I. Overall, the individual EBE estimates of the disabilities of the subjects were in the range of -2.73 (healthier or less disabled) to 3.82 (less healthy or more disabled). When alternative shapes of distributions, e.g., t-distribution and box-cox transformation, were explored for the distribution of disability at baseline, neither the OFV change nor the parameters estimated for these alternative shapes were found to be significant, suggesting that the normality assumption was valid.

The longitudinal IRT model characterizes the temporal changes in the “disability” over time. Since it is a hypothetical construct, it is a dimensionless number, but can be interpreted in standard deviation terms, i.e., it takes 50 months for a typical individual to progress linearly (0.02 units/month) by 1 standard deviation relative to the disability at baseline. Subjects with a lower

disability at baseline were likely to progress faster due to a correlation of -0.25 between the random effects on baseline and on slope of disease progression. The goodness of fit plots and simulation-based diagnostics (provided in Appendices II and III) suggested that the model was able to adequately characterize the ICCs and the longitudinal changes in both the total score as well as the individual item level profiles.

The residuals plot (shown in Fig. 2, panel A) showed distinct patterns, for e.g., positive correlations between residuals of items 1–26, negative correlations (in general) between residuals of items 1–26 and 27–43. Furthermore, “checkered” pattern with alternating positive and negative correlations was observed between the residuals of items 30–43, and between 30–43 and 50–57, respectively. In order to address these patterns, multiple latent variables were explored. Another distinct line corresponding to the correlation of the residuals of item 61 with all the items was observed. This, however, was due to the very low frequency of responses that were dependent on the responses to item 60.

Multiple Latent Variables

Of the four latent variables (namely PR, RSR, LSR, and NSR) model and the three latent variables (namely PR, SR, and NSR) model with a mixture to identify which of the two subpopulations (most disabled side being either right or left side) the subject is more likely to belong to, the latter offered a better description of the data. This choice was driven by the correlation between the variance of the distribution of disability at baseline for RSR and LSR being low (-0.15); therefore, a common latent variable SR seemed a more rational structure. The ICC parameters of the final longitudinal model with three latent variables and a mixture were successfully estimated (listed in Appendix IV). Owing to the time costs associated with the current bootstrap techniques together with the model complexity and the data size, standard errors could not be evaluated.

The proportion of subjects whose right side was the most disabled based on the mixture estimate was 58%. The linear progression rates of PR and NSR were similar, i.e., around 0.02 units/month, or about 50 months for the typical subject to progress linearly from the disability at baseline by 1 standard deviation. However, the progression rates of the SR varied depending on if the items evaluated the most disabled side (either the right or the left side) initially (i.e., at baseline) or if the items evaluated initially the better side.

For the items evaluating the most disabled side initially, the disease progression occurred slower (slope = 0.0072 units/month), compared to items evaluating initially the better side (slope = 0.030 units/month). This suggests that as the initially better side deteriorates quicker, its difference in disability with the initially most disabled side becomes smaller as time progresses. The shift parameter that was used to reflect the lower disability for the initially better side was estimated to be 2.11 with a SD of 0.60.

The latent variables at baseline were correlated: 0.57 for PR and NSR, 0.41 for PR and SR, and 0.57 for NSR and SR. Furthermore, the correlations between the variance of the latent variable at baseline and the slope of the linear progression of the same latent variable as well as with the other latent variables were found to be negative, suggesting that subjects with lower disability at baseline were likely to progress faster.

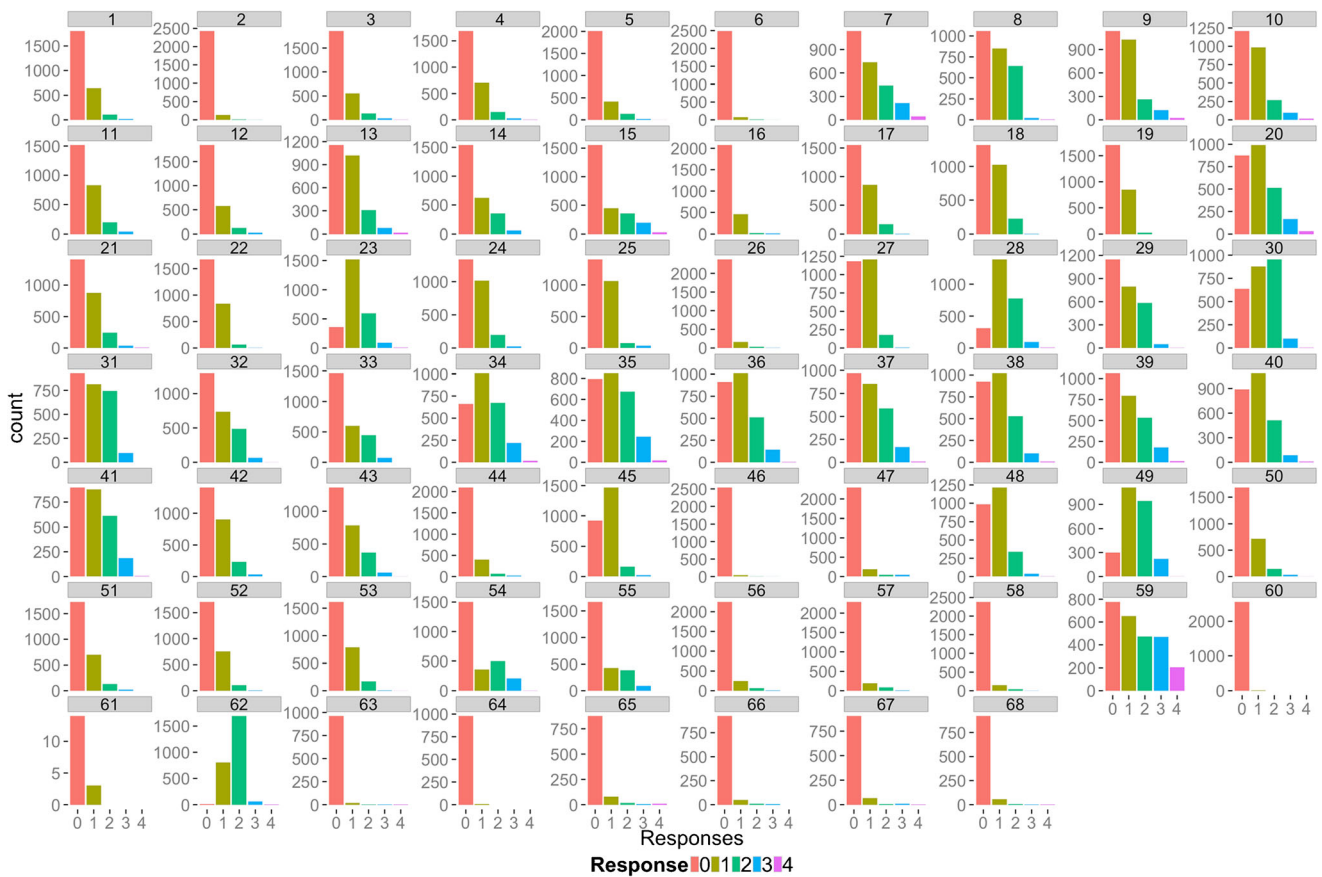


Fig. 3. Distributions of observed item responses in the DeNoPD cohort

The goodness of fit plots for the final three latent variable model with a mixture implementation are shown in Figs. 4 and 5. The ICCs for the cumulative probabilities for three items, namely 24 (*getting out of bed*), 35 (*finger tapping of the left hand*), and 62 (*Hoen & Yahr stage*) representative of most informative items (as calculated (10) based on Fischer Information Matrix, FIM) in each of the latent variable categories, namely PR, SR, and NSR, respectively, shown in Fig. 4 (and for rest of the items in Appendix V), suggest that the IRT model fit seems to be in good agreement with the GAM-cross-validated spline function fit. Additionally, it can also be observed that the item-specific parameters, namely the slope, a_j (or discrimination parameter) and the difficulty parameter, b_{jk} , differ between the score categories.

The ICCs for the individual probabilities of the respective items shown in Fig. 1 seem to overlap. However, the pattern of the overlap is very different for different items suggesting that the items vary in the informative content and in their relation to the (respective) underlying latent variable.

The residuals plot (shown in Fig. 2, panel B) illustrated a significant improvement in the dependence pattern observed with the single latent variable model. The distinct line corresponding to the correlation of the residuals of item 61 with all the items was still observed, owing to the very low frequency of responses that were dependent on the responses on item 60.

Further simulation-based diagnostics are shown in Fig. 5: the top row and bottom left panels show simulated item responses belonging to each of the three latent variables, which were summed

up to get total score for that respective latent variable and compared to the respective observed scores, while in the bottom right panel, all the simulated item responses for items 1 to 68 were summed up to get the total MDS-UPDRS score and compared to the total of the observed scores for items 1 to 68, respectively. The solid lines, representing the median, 2.5th and 97.5th percentiles, seemed to be within the shaded areas, represented by the 95% confidence intervals based on the model prediction, suggesting that the model simulations were in good agreement with the observations for most of the time points. The visual predictive checks at the individual item level are shown in Appendix VII.

DISCUSSION

This work applied IRT methodology to describe the longitudinal changes in MDS-UPDRS records from the *De Novo* Parkinson's disease patient cohort. To achieve this, it was necessary to explore the use of multiple latent variables and development of diagnostic tools to assess it. The questionnaire-type endpoints are traditionally analyzed using the total (composite) of sub-scores on a continuous scale. However, the use of such an analysis method can be potentially limiting or misleading due to implicit assumptions such as (i) the importance of the sub-score relative to the total score being ignored, (ii) the difficulty/discrimination between the items in the questionnaire being ignored, and (iii) ignoring that data may be missing and that certain subjects may have intentionally refused to answer an item because it was difficult, thus

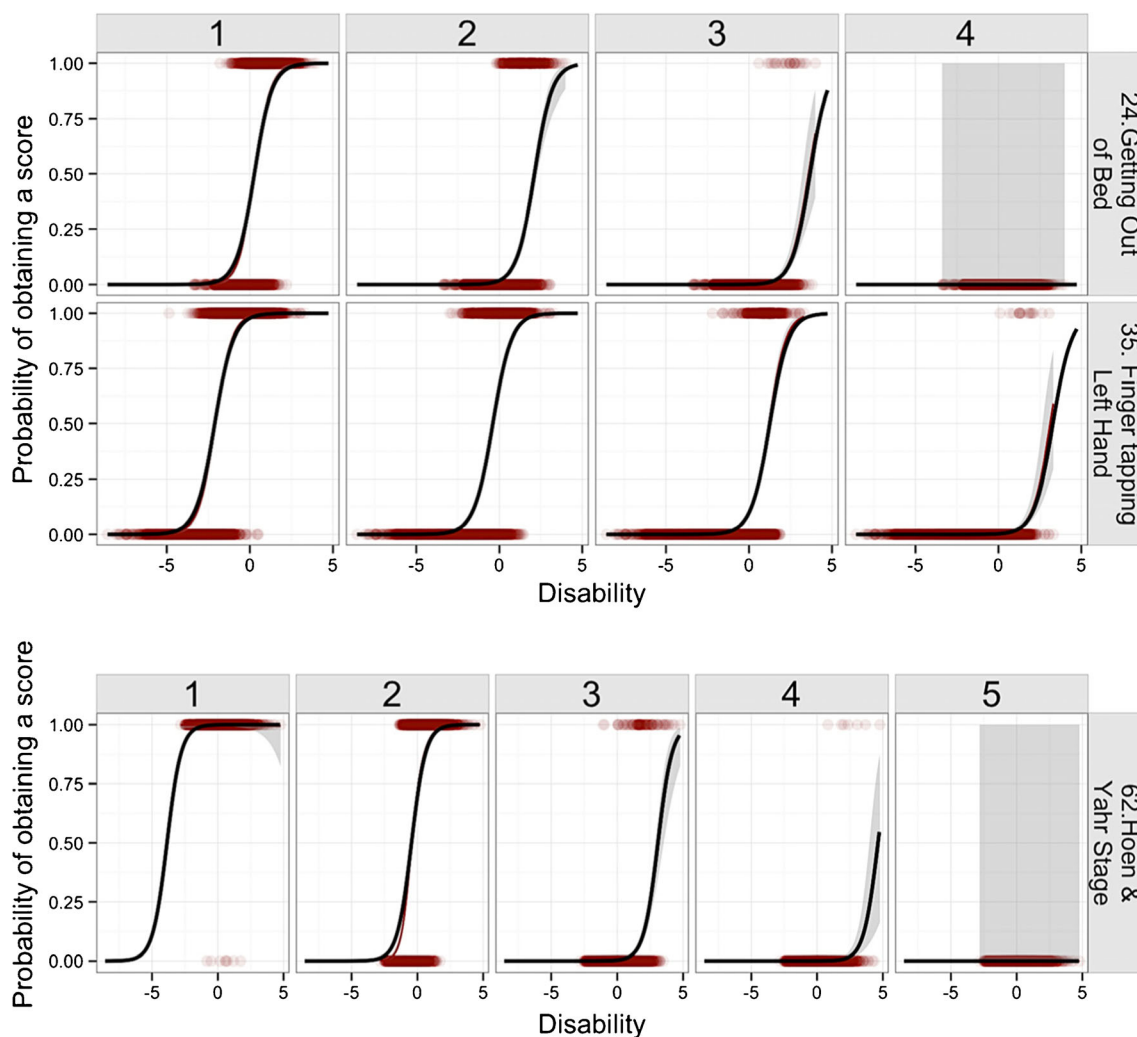


Fig. 4. Item Characteristic Curve (ICC) fits of the cumulative probabilities along with the generalized additive model (GAM) diagnostics for items 24, 35, and 62, each representative of most informative items within latent variables PR, SR, and NSR, respectively. *Red dots* indicate the observed scores. Panel with “1” (*left most column*) shows values with scores 0 and ≥ 1 ; panel with “2” (*second column from left*) shows values with scores 1 and ≥ 2 ; panel with “3” (*third column from left*) shows values with scores 2 and ≥ 3 ; panel with “4” (*fourth column from right*) shows values with scores 3 and ≥ 4 ; panel with “5” (*last column*) shows values with scores of 5. The ICC curves from the IRT model fit (*black line*) is compared to the fit of a GAM with cross-validated cubic spline as a smoothing function (*dark red line* and the associated 95% confidence interval is shown in *gray*)

incorporating a bias if imputed (e.g., as 0 or mean value). IRT offers an alternative approach by considering all the individual item level responses and relating them to an underlying hidden latent variable, defined as “disability” in this work, while modeling the longitudinal changes in the disability scale.

This methodology was applied to scales composed of sub-scores of different types of data, e.g., ordered categorical, binary, counts. Having its origins in psychometrics and its first pharmacometrics application in Alzheimer’s disease, this approach has since then been used in schizophrenia, multiple sclerosis (ms), amyotrophic lateral sclerosis (ALS), and oncology. In this work, IRT was successfully applied to the scale MDS–UPDRS used in PD, composed of ordered categorical and binary items and describing the longitudinal changes in the disability.

There are two steps while modeling questionnaire data in the IRT framework: (1) estimation of the item response model

parameters, also referred to as ICCs, and (2) estimation of the longitudinal aspects of the disease progression. This can be performed in a sequential or a simultaneous manner. Although both approaches use the entire data, the way it is handled is slightly different: in step 1 of the sequential approach, the data is treated such that each occasion is a separate individual for the estimation of ICCs, thus efficiently utilizing all the information to inform the ICC parameters. Then, with step 2, the data within each individual was reconciled, to address the temporal aspects, which are ignored in step 1, thus potentially avoiding potential misspecifications. The simultaneous approach on the other hand aims to capture both ICCs as well as temporal changes at the same time. In the current IRT work, simultaneous fit seemed to give better description of the data over the sequential fit.

A new diagnostic tool assessing the pattern in residuals suggested the need to explore multiple latent variables, as a single latent variable may not be sufficient to address the different aspects

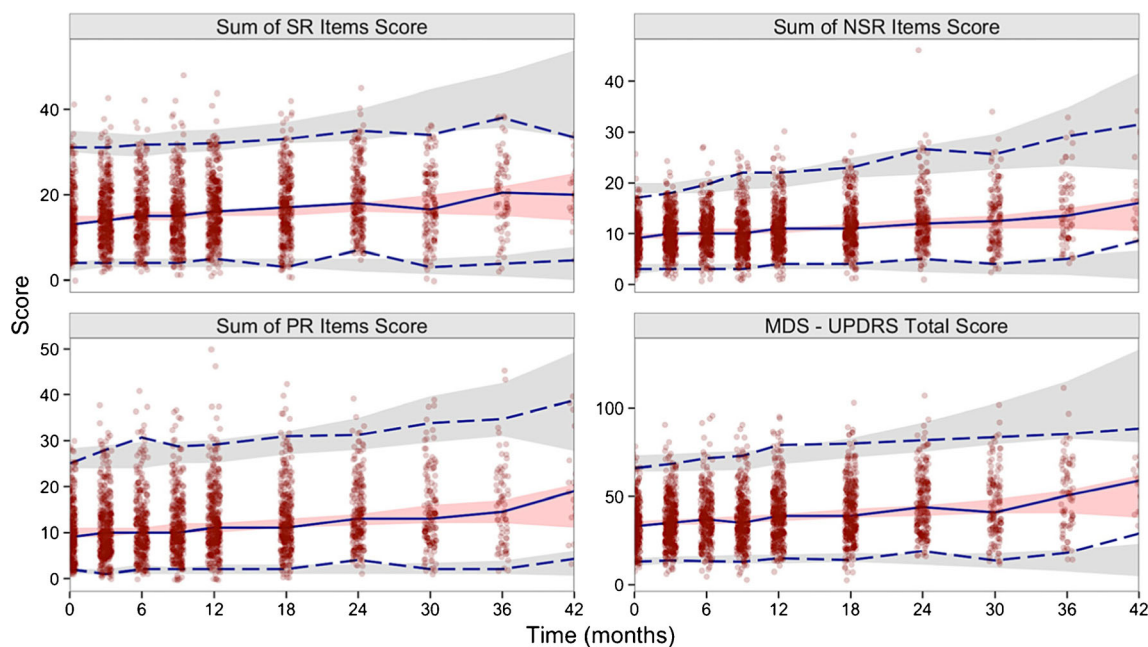


Fig. 5. Visual predictive checks comparing the median, 2.5th and 97.5th quantiles (blue lines) of the observed data (points) with the respective confidence intervals (shaded areas) based on the final longitudinal model with three latent variables and a mixture

of the disease. Based on how the questionnaire was set up, a three latent variable model with a mixture determining whether subjects were affected predominantly on the left side or the right side was developed. The model evaluation fits suggested that the model captured the longitudinal changes adequately. Further, there was a significant improvement in the pattern of the residuals compared to the obvious patterns observed in the case of the single latent variable. While not much literature is available where multiple latent variables were tested within the IRT framework, the exploratory IRT analysis by Verma *et al.* (18) reported that ADAS-cog may not be unidimensional, and explored multiple latent variables.

In the present work in PD, a mixture was needed on one of the latent variables. It suggested that there was 58% probability that a subject belongs to the subpopulation in which the right side is the most likely to be affected.

While a negative correlation was observed between baseline and slope in the current analysis, a positive correlation has been reported in the literature when a linear model was used to characterize longitudinal changes in total score UPDRS data (19). Plausible explanations for this discrepancy include the difference in scale (MDS-UPDRS vs. UPDRS) and structural model (IRT vs. continuous).

Additionally, PD medication (recorded within the following classes: (a) dopamine replacement, (b) COMT inhibitors, (c) dopamine agonists, (d) MAO-B inhibitors, (e) propranolol, (f) anti-cholinergics, and (g) others) was allowed after entry, thus potentially affecting disease progression. Since the potential effect of these medications on the disease progression was not quantified as part of this study, it must be taken into account in the interpretation of the disease progression, which is therefore a combination between true disease progression, placebo effect, and treatment effect.

One of the promising advantages of the IRT methodology is that it allows for pooling of data across multiple studies and even across the different variants of the assessments.

Therefore, the IRT model developed with MDS-UPDRS data can be adapted and used as a tool to analyze more traditional UPDRS data collected in PD clinical trials. Furthermore, the ICCs generated from this work can be used as informative priors to analyze PD trial data and thus effectively integrate knowledge from multiple sources.

CONCLUSION

A longitudinal IRT model with three latent variables was successfully developed to describe disease progression of PD in *De Novo* subjects adequately. This model-based approach developed using MDS-UPDRS data offers an improved utilization of MDS-UPDRS data, not only at the total score level but at the individual item level. Additionally, a new diagnostic tool assessing the pattern of the residuals gave insights into exploring multiple latent variable models when the single latent variable model may not be appropriate. This framework can be adapted to handle data from other clinical endpoints in PD and therefore allowing integration from wide sources.

ACKNOWLEDGEMENTS

Data used in the preparation of this article were obtained from the PPMI database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI—a public-private partnership—is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Inc.; Eli Lilly and Company and its subsidiary company Avid Radiopharmaceuticals; Biogen Idec Inc.; Bristol-Myers Squibb Company; Covance, Inc.; F-Hoffmann- La Roche Ltd and its subsidiary company Genetech, Inc.; GE Healthcare; GlaxoSmithKline plc.; H. Lundbeck A/S; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Pfizer Inc.; Piramal Imaging; and UCB.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest Authors declare that they have no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

REFERENCES

- Wirdefeldt K, Adami H-O, Cole P, Trichopoulos D, Mandel J. Epidemiology and etiology of Parkinson's disease: a review of the evidence. *Eur J Epidemiol*. 2011;26 Suppl 1:S1–58.
- Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stebbins GT, *et al*. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): process, format, and clinimetric testing plan. *Mov Disord*. 2007;22(1):41–7.
- Vu TC, Nutt JG, Holford NHG. Progression of motor and nonmotor features of Parkinson's disease and their response to treatment. *Br J Clin Pharmacol*. 2012;74(2):267–83.
- Chan PL, Holford NH. Drug treatment effects on disease progression. *Annu Rev Pharmacol Toxicol*. 2001;41:625–59.
- Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. *Educ Meas Issues Pract*. 2005;12(3):38–47.
- 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.
- Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care*. 2000;38(9 Suppl):II28–42.
- Brennan RL, editor. Educational measurement. 4th ed. Westport: Praeger Publishers; 2006.
- Ard MC, Galasko DR, Edland SD. Improved statistical power of Alzheimer clinical trials by item-response theory: proof of concept by application to the activities of daily living scale. *Alzheimer Dis Assoc Disord*. 2013;27(2):187–91.
- 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;31(8):2152–65.
- Balsis S, Unger AA, Bengtson JF, Geraci L, Doody RS. Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: a comparison of item response theory-based scores and total scores. *Alzheimers Dement*. 2012;8(4):288–94.
- Novakovic AM, Krekels EHJ, Munafò A, Ueckert S, Karlsson MO. Application of item response theory to modeling of expanded disability status scale in multiple sclerosis. *AAPS J*. 2017;19(1):172–9.
- Krekels EHJ, Kalezić A, Friberg L, Vermeulen AM, Karlsson MO. Item response theory for analyzing placebo and drug treatment in phase 3 studies of schizophrenia. In: PAGE 23 (2014) Abstracts of the Annual Meeting of the Population Approach Group in Europe. 2014.
- DeMars C. Item response theory. Oxford, New York: Oxford University Press; 2006.
- Beal SL, Sheiner LB, Boeckmann A, Bauer RJ. NONMEM user's guides (1989–2009). Ellicott City, MD, USA; 2009.
- Wood S. Introducing GAMs. In: Generalized additive models: an introduction with R. Hall/CRC Press; 2006.
- Keizer RJ, Karlsson MO, Hooker A. Modeling and simulation workbench for NONMEM: tutorial on Pirana, PsN, and Xpose. *CPT Pharm Syst Pharmacol*. 2013;2, e50.
- Verma N, Markey MK. Item response analysis of Alzheimer's Disease Assessment Scale. *Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf*. 2014;2014:2476–9.
- Holford NHG, Chan PLS, Nutt JG, Kiebertz K, Shoulson I, Parkinson Study Group. Disease progression and pharmacodynamics in Parkinson disease—evidence for functional protection with levodopa and other treatments. *J Pharmacokinet Pharmacodyn*. 2006;33(3):281–311.