Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Pharmacy 375

Pharmacometric Evaluation of Item Response Modeling to Inform Clinical Drug Development

LETICIA ARRINGTON

ACTA UNIVERSITATIS
UPSALIENSIS
2025

ISSN 1651-6192
ISBN 978-91-513-2438-8
urn:nbn:se:uu:diva-552893

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in A1:107a, BMC, Husargatan 3, Uppsala, Tuesday, 13 May 2025 at 13:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English. Faculty examiner: Professor of Clinical Pharmacy Sebastian Wicha (Institute of Pharmacy, University of Hamburg, Germany).

Abstract

Arrington, L. 2025. Pharmacometric Evaluation of Item Response Modeling to Inform Clinical Drug Development. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy* 375. 78 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2438-8.

Drug development is the process of advancing novel therapeutics to market to improve patient outcomes. However, in hard-to-treat diseases like neurodegenerative disorders there is a high failure rate in late-stage trials, creating significant unmet needs. This highlights the need for more sensitive endpoints, improved trial designs, or analytical methods to optimize data utilization. In many diseases, clinical outcome assessments (COAs) serve as clinical endpoints and are often reported as a composite score, potentially losing important information present at the item level. Alternatively, item response theory (IRT) leverages item-level data to describe the relationship between a subject's response on an item and their underlying ability, through item characteristic functions (ICFs), offering a more informed analysis of COAs. This thesis evaluates the robustness of IRT, estimation strategies and its applicability to model rating-scale-based COAs to facilitate model-informed drug development (MIDD).

For single time point analysis, our findings suggest at least 100 subjects and 20 items are generally sufficient. Comparison of Laplace and Gaussian-hermite quadrature (GHQ-EM) for the estimation of item parameters, indicated similar accuracy and precision with slight improvement in accuracy for GHQ-EM. IRT models in reduced assessments were relatively stable up to ~40-60% information remaining. However, removing items shifts the measured disease construct, which can affect the accurate assessment of disease progression and drug effect. The trade-offs in information lost or gained should be considered when shortening assessments. Comparison of two common estimation strategies for determining ICFs indicated similar performance, each providing different advantages. IRT was also effective in classifying disease (Parkinson's vs SWEDDs), showing comparable performance to artificial neural networks. Additionally, IRT demonstrated superior power for detecting symptomatic treatment effect in a short duration trial compared to traditional approaches, highlighting IRT's potential not only for endpoint analyses but as a strategic tool to optimize trial design. Greater public disclosure of applied IRT in real-time drug development, such as inclusion in trial protocols or in regulatory milestones could foster broader acceptance and wider adoption beyond ad-hoc analyses. In conclusion, this thesis presents a methodological foundation for successful implementation of IRT in a pharmacometric framework to facilitate MIDD and inform clinical decision-making.

Keywords: pharmacometrics, nonlinear mixed-effects models, item response theory, Parkinson's Disease, Alzheimer's Disease, composite score, clinical outcome assessments

Leticia Arrington, Department of Pharmacy, Box 580, Uppsala University, SE-75123 Uppsala, Sweden.

ISSN 1651-6192
ISBN 978-91-513-2438-8
URN urn:nbn:se:uu:diva-552893 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-552893)

To those who inspire

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. Arrington, L., Ueckert, S., Ahamadi, M., Macha, S., Karlsson, MO. (2020) Performance of longitudinal item response theory models in shortened or partial assessments. *J Pharmacokinet Pharmacodyn*, 47(5):461-471.

II. Arrington, L., Ueckert, S. (2024) Item response parameter estimation performance comparison using Gaussian quadrature and Laplace. *arXiv preprint server,* arXiv:2405.20164 [stat.ME]

III. Arrington, L., Karlsson, MO. (2024) Comparison of Two Methods for Determining Item Characteristic Functions and Latent Variable Time-Course for Pharmacometric Item Response Models. *AAPS J*, Jan 25;26(1):21.

IV. Arrington, L., Van Dijkman, S.C., Plan, EL., Karlsson, M.O. (2025) Item Response Modeling for Differentiation of Parkinson's patients and Subjects Without Evidence of Dopaminergic Deficit. *CPT: Pharmacometrics Syst. Pharmacol*. doi: 10.1002/psp4.70000. Epub ahead of print. PMID: 40045658

V. Arrington, L., Mehrotra, N., Hogan, J., Yee, K.L. Item Response Theory Pharmacometric Modeling to Support Proof of Concept Trial in Patients with mild-to-moderate Alzheimer's Disease. *In Manuscript*. Presented at ACoP13; Aurora, CO, USA; October 30-November 2, 2022.

Contents

Abbreviations

ADAS-COG	Alzheimer's disease assessment scale cognitive subscale
ANN	Artificial neural network
APOE	Apolipoprotein E
AUC	Area under the curve
COA	Clinical outcome assessment
DeNoPD	De novo Parkinson’s Disease
EDA	Exploratory data analysis
EM	Expectation maximum
EMA	European Medicines Agency
FDA	Food and Drug Administration
FOCE	First order conditional estimation
GAM	Generalized additive model
GHQ-EM	Gaussian Hermite Quadrature Expectation Maximum
GPC	Generalized partial credit
GRM	Graded response model
ICC	Item characteristic curve
ICF	Item characteristic function
IRT	Item response theory
LV	Latent variable
MCC	Matthews correlation coefficient
MDS-UPDRS	Movement Disorder Society Unified Parkinson's Disease Rating Scale
MIDD	Model informed drug discovery and development
mirt	Multidimensional IRT
NLMEM	Nonlinear mixed effect model
NONMEM	Non-linear mixed effect model software
OFV	Objective function value
PANSS	Positive and Negative Syndrome Scale
PD	Parkinson's Disease
PDUFA	Prescription drug users fee act
PPMI	Parkinson’s Progression Markers Initiative

PsN	Perl speaks NONMEM
PRO	Patient reported outcome
RMSE	Root mean square error
ROC	Receiver operating characteristic curves
rRMSE	Robust root mean square error
SCM	Stepwise covariate modeling
SD	Standard deviation
se	Standard error
SWEDD	Scans Without Evidence of Dopaminergic Deficit
VPC	Visual predictive checks

1 Introduction

1.1 Drug Development

The drug development process aims to develop safe and efficacious novel pharmaceutical compounds into marketed therapeutics that improve patient's lives. This highly regulated process is divided into several stages: discovery, preclinical research, clinical research, regulatory approval and post marketing.

First, a target molecule is identified and optimized in the discovery phase then the candidate progresses into preclinical development. The preclinical stage is designed to collect information on various aspects of the molecule including molecular properties, physiochemical properties, pharmacology and toxicology via in-vitro and in-vivo studies. The preclinical experiments provide vital information on proof of mechanism and provides insights to inform dose in first in human trials. The clinical stage of development, in which investigational drug clinical trials are performed, is divided into early and late stage. The phases of clinical development can also be viewed as "learn-confirm" cycles[1]. Phase 1 assesses pharmacokinetics, safety and tolerability of the doses administered. Phase 1 begins with First-in-human trials (i.e., single ascending dose) which typically evaluates these attributes in a small number of healthy volunteers, however in more severe diseases (i.e., oncology) the drug may be evaluated in the patient population especially if there are adverse effects that would not be safe for healthy volunteers. Additional clinical pharmacology studies may be required depending on the drug's modality and ADME properties to further evaluate pharmacokinetics and safety across different conditions, including special populations. In phase 2a safety and early markers of efficacy are evaluated in a larger number of patients to confirm learnings from Phase 1. In phase 2b the focus shifts to more rigorous assessment of benefit/risk to determine if further investment is warranted (i.e., go/no go decision to late phase development). Phase 3 aims to confirm sufficient efficacy and safety in the larger target patient population. Phase 3 trials (i.e., registrational trials) are conducted in a randomized manner at the target dose with a positive control; earlier phase trials can also be conducted in a randomized fashion depending on the research questions and indication. All trials should be designed following best practices and requirements set by regulatory agencies. The results from these registrational trials along with the totality

of data across phases are used to support the formal application for marketing authorization for the drug. In drug development clinical trials are focused on the safety and efficacy of an investigational drug through controlled interventions. On the other hand, observational trials offer a different approach to understanding the potential impact of an approved drug or therapy in a natural setting without introducing any additional interventions and are typically run by either other entities outside of the pharmaceutical company or as part of post-approval activities. This information is gathered to establish the real word effects of a drug, patient outcomes or the natural progression of a disease.

Executing a robust clinical trial requires selecting an appropriate patient population that is representative of the target group, as well as ensuring critical design elements, such as well-defined endpoints (e.g., clinical outcome, biomarkers, surrogate markers) and adequate sample size, to effectively evaluate the drug's effect. These components are essential to validate the trial findings and achieve the desired power for detecting a clinically meaningful difference at a given level of significance.

Clinical trial inclusion and exclusion criteria are designed to optimize the ability to attribute the study outcomes to the tested treatment and minimize the impact of confounding variables, therefore reducing bias. Selecting the appropriate trial population in earlier stages of development increases the probability of success in Phase 3 development. In certain disease indications such as infectious disease, selection of the appropriate trial population may be relatively straight forward. However, in other disease areas such as neurodegenerative diseases, with heterogenous patient populations, identifying the appropriate population and validated targets for treatment is particularly challenging. The lack of reliable biomarkers, coupled with variations in disease manifestation and progression, further complicates this process. Significant diagnostic advances have emerged in the recent years, but tools like imaging (e.g., MRI, PET) and fluid biomarkers (e.g.,CSF) remain costly and can be invasive. This highlights the need for convenient, cost-effective, complementary approaches to improve disease evaluation and trial outcomes.

According to a 2024 article titled "Costs of Drug Development and Research and Development Intensity in the US", which conducted an economic evaluation of drug development costs in the United States from 2000 to 2018, the average cost of developing a drug is $172.7 million[1, 2]. In addition, this cost "increased to $515.8 million when cost of failures was included in the calculation". For certain drugs, this cost can be significantly higher[3]. Given the cost of development and the approximate 50% failure rate in late-stage development – particularly in challenging areas such as neurodegenerative

diseases- it is imperative to enhance early-stage research activities to increase probability of success in later stages.

1.2 Model-informed drug development

Ultimately the regulatory decision to approve a drug is based on many factors and is informed by the diverse data collected during the drug development process. Given the complexities of drug development, there is a need to enhance early decision-making - this is where model-informed drug development (MIDD) can play a key role.

MIDD is the process that integrates this data from multiple relevant studies or sources in a "*quantitative framework for prediction and extrapolation, centered on knowledge and inference generated from integrated models of compound, mechanism and disease level data and aimed at improving the quality, efficiency and cost effectiveness of decision making*"[4].

The Food and drug Administration (FDA) codified MIDD in the Prescription Drug User Fee act (PDUFA) VI in 2017 and continues to show commitment to these goals in subsequent reauthorizations with various initiatives such as MIDD paired meeting program[5].The European medicines agency (EMA) has also highlighted the value of MIDD as demonstrated with the EMA modeling and simulation working group publications[6, 7].

These modeling and simulation-based approaches serve as powerful tools to enable more efficient drug development, support dose optimization and potentially reduce drug attrition[8]. In this thesis we evaluate the statistical method, Item Response Theory, both cross-sectionally and within a pharmacometrics disease progression framework to enhance the utilization of data from clinical outcome assessments (COAs), the primary measurement tool used in neurodegenerative disorders in order to facilitate MIDD. This approach is also compared to traditional methods, highlighting its potential advantages.

1.2.1 Phamacometrics

At the forefront of MIDD is the quantitative discipline of pharmacometrics. Pharmacometrics quantifies drug pharmacokinetic (PK) /pharmacodynamic (PD), disease process and their relationships while leveraging disease and trial information through in silico mathematical and statistical models. Pharmacokinetics is the study of the time course of a drug through the body and factors that influence kinetics such as absorption, distribution, metabolism and elimination. Pharmacodynamics is the effect of the drug on the body. This effect

can be a measured response or signal indicating engagement with target, efficacious benefit or an adverse effect (safety).

There are several types of pharmacometrics models such as; population PK/or PD (PopPK/PD), to semi-mechanistic and mechanistic models such as but not limited to, physiological based PK (PBPK), and quantitative systems pharmacology (QSP) models. The focus of this thesis are population models. PopPK/PD models are modeled using non-linear mixed effect modeling (NLME). PopPK/PD models describe the typical individual (mean population effect, fixed effect) and also identify and quantify sources of variability (random effect) in PK and PD. Random effects describe variability on several levels: between-subject (inter-individual variability, IIV), within-subject (inter-occasion variability, IOV), and residual variability. Residual unexplained variability (RUV) is used to explain the remaining unexplained variability that is not captured by subject-specific level prediction. Often IOV is assumed to be captured within RUV and not further estimated. Intrinsic (e.g., age, gender) and extrinsic factors (e.g., drug-drug interaction) are attributes that contribute to the differences observed in outcomes among subgroups of the assessed patient population, these are referred to as covariates in pharmacometrics.

In this thesis NLME models were applied to both continuous and discrete data. The general equation for continuous data is described as

$$Y_{ij} = f\left(t_{ij,}g(\theta,\eta_i,X_i)\right) + h\left(t_{ij,}g(\theta,\eta_i,X_i),\mathcal{E}_{ij}\right) \quad (1.1)$$

$$\eta_i \sim \mathrm{N}(0,\omega^2) \quad (1.2)$$
$$\mathcal{E}_{i,j} \sim \mathrm{N}(0,\sigma^2) \quad (1.3)$$

Where *Yij i*s the observed dependent variable of an individual *i* at the time *j*, and the function *f* is the individual prediction given the structural model based on independent variables, t_{ij}, (e.g., dose) and vector *g* (.) which describes the individual parameters as a function of typical values (θ), random effects(η_i), and covariates (X_i). Function h (.) describes the residual variability. The random effect parameters ($\eta_i, \mathcal{E}_{i,j}$) are typically assumed to follow a normal distribution with a mean of zero and some variance (covariance matrices Ω, Σ).For discrete data the probability (P) of observing the "event" or "outcome" (k) is modeled using probability density function (pdf) *l*(.) as follows:

$$P(Yij = k) = l\left(t_{ij,}g(\theta,\eta_i,X_i)\right) \quad (1.4)$$

There are several analysis tools available for NLME modeling, but for the past 30 years NONMEM® (NONlinear mixed effects) has served as the primary

software platform for population analysis in pharmacometrics. NONMEM is also the primary software used in this thesis[9].

1.2.2 Maximum likelihood Estimation

Parameter estimation for NLME and many machine learning models relies on maximum likelihood methods, where the estimated model parameters (Θ) are those that maximize the likelihood we would observe the data as a function of the model.

The individual likelihoods (L_i) are derived from the conditional distribution of data given the random effects. The integral is then taken across all possible values of η as shown below.

$$\mathcal{L}_i(Y_i, \Theta) = P(Y_i, \Theta) = P(Y_i|\theta, \Omega, \Sigma) = \int P(Y_i\,|\theta, \Sigma, \eta_i) * P(\eta_i\,|\Omega) d\eta \quad (1.5)$$

Where Y_i are the observations for individual i and $\mathrm{P}(Y_i|\theta, \Sigma, \eta_i)$ is the conditional probability density of the observations and $\mathrm{P}(\eta_i|\Omega)$is conditional probability density of η.

The overall or population likelihood is the product of all individual likelihoods as shown below.

$$\mathcal{L}(Y_i, \Theta) = \prod_i \mathcal{L}_i(Y_i, \Theta) = \prod_i \int P(Y_i\,|\theta, \Sigma, \eta_i) * P(\eta_i\,|\Omega) d\eta \quad (1.6)$$

In NONMEM this is achieved by minimizing the objective function value (OFV) which is equivalent to the -2log (L).

In general, there are two approaches to calculate the likelihood; "approximation" and more "exact" methods. The choice of method depends on factors such as data type, properties for usefulness (e.g., large number of random effects (η)), and computation time. The approaches primarily differ in the integration methods and thus present both strengths and limitations. Details on the estimation methods used in this thesis are presented in methods.

The final models are those that sufficiently characterize the data, including aspects of variability and uncertainty. These models can then be used for prediction, through simulation approaches (i.e., clinical trial simulation) to answer key questions to inform future trials and decision-making.

1.3 Score Based Endpoints

1.3.1 Data types

In statistics data can be categorized in two high level groups: categorical and numerical. Categorical data describes data that can only belong to one of the distinct categories defined; these data can be nominal (no defined order) or ordinal (ordered in some way e.g, degrees of impairment). Numerical data, can be discrete(whole values) or continuous (any value within a range).

In this thesis, categorical, discrete and continuous data are leveraged; either real, simulated or derived data. In Papers I-V, item level data was leveraged as ordered categorical input for the item response models or artificial neural network. In addition, Paper III also leveraged derived continuous data for implementation of the sequential method for item response model development. Paper IV and V used total score data from a composite assessment, which is a discrete value on a bounded scale but often treated as continuous in disease progression modelling.

1.3.2 Clinical outcome assessments

Clinical outcome assessments (COA) are used to diagnose and guide treatment of diseases in a variety of therapeutic areas, such as chronic pain, oncology and neurodegenerative disorders. In the context of clinical trials, especially in neuroscience these assessments also serve as clinical endpoint measures or measures to define inclusion criteria. These data are often collected through questionnaires or more recently through the use of digital health technologies through tools like computing platforms and wearable sensors, primarily in clinical trial settings. The four major types of clinical outcome assessments are i) patient-reported outcomes (PRO) ii) observer-reported outcomes (ObsRO), iii) clinician-reported outcomes (ClinRO), and iv) performance-based Outcomes (PerfO)[10].

A COA is designed to measure a specific aspect of the disease construct and can be intentionally developed to assess multiple concepts to quantify a patient's level of functioning, well-being and survival based on symptoms and signs across multiple dimensions of the disease. COAs consist of single or multiple items (i.e., questions or subscores). Each item provides options associated with response measure, which may be numerical value or range of values or some other measurable criteria. When these items or subscores are combined, they form a composite assessment. The composite scale represents an endpoint specifically designed to reflect the clinical outcome of interest.

Rating scale-based COAs measure patient experience using a numerical rating scale. These questions can take various forms including binary ("yes"

or “no”) assigned 0 or 1, ordinal responses for example, ranging from mild impairment to severe impairment as well as count data. The scores from each of these questions when summed are called a composite or total score (TS). The most commonly used rating-scale based COAs seen in clinical practice are pain scales (e.g., Likert, Visual analogue scale), quality of life (QoL) questionnaires and functional questionnaires [activities of daily living (ADL)].

There are different methods to approach scoring of the responses when developing a COA to assess the concept of interest. When combining multiple items into a single score for a COA this is typically informed by using a measurement model. Two common models are reflective indicator and composite indictor models[10]. In a reflective model, all items are considered manifestations of a single underlying concept of interest (unidimensionality). In contrast, the composite indicator model combines items that when taken together, define the overall concept. These items are separate and may not be correlated (e.g., educational ranking, activities of daily living). For the reflective indicator model, item response theory (IRT) is one approach to design, evaluation, and the scoring of COAs, when adequate data is available[10].

The FDA and EMA, as well as international council of harmonization (ICH) have issued a series of guidance documents outlining technical requirements related to COAs when used as a clinical trial endpoint to ensure the data collected in the clinical trial is meaningful and can support regulatory decisions[10, 11]. These documents also provide guidance in the development of new COAs. Qualified COAs are those that have been reviewed by regulatory agencies, meeting specific standards for measuring disease and treatment outcomes in clinical trials[12]. Although using a qualified COA has its advantages, it is not required; non-qualified COAs can be used if they are validated for their intended purpose (i.e., fit for purpose based on the context of use). Developing a well-designed COA is challenging. Despite the rigorous process involved, issues such as inconsistent methods across diseases, lack of established measures, subjective patient reporting, cultural perceptions, inter-rater variability, can all affect consistency and reliability[13].

Key aspects of a high-quality COA include test validity (truly measures construct of interest), test-retest reliability and sensitivity to meaningful clinical changes. The outcomes should be unaffected by factors unrelated to disease or ailment (e.g., how test is administered, responder fatigue). Scoring methods should also be appropriate for the population of interest and be evidence based. Lastly, the outcome from the assessment should be interpretable and consistently understandable across populations[10].

While COAs are designed to be change-sensitive instruments with strong construct validity, many COA-based endpoints are still considered to be

highly variable. Additionally, when used at different stages of a disease, may not be sensitive to changes at the lower and higher ends of disease severity, leading to floor and ceiling effects

1.3.3 Rating-scale based clinical endpoints

In this thesis, cases where real data was leveraged, data from the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating scale (MDS-UPDRS) assessment and the Alzheimer's Disease Assessment Scale-Cognitive Subscale 11 (ADAS-Cog11) were used.

Parkinson's disease and MDS-UPDRS rating scale

Parkinson's disease (PD) is a chronic progressive neurodegenerative disease with a rapid increase in prevalence. Its clinical manifestations consist of both motor and non-motor symptoms leading to disability and reduced quality of life. PD's hallmark is the loss of dopamine producing neurons and Lewy body formations in the midbrain in early disease but as the disease progresses can spread through multiple brain regions[14]. Currently the treatment landscape is primarily symptomatic treatments with levodopa, dopamine agonists and monoamine oxidase B (MAO-B) inhibitors being the standard of care with a research focus to develop disease modifying therapies[14].

The MDS-UPDRS scale is a commonly used tool to assess the severity, progression and treatment of PD, since its introduction in 2008 [15]. This revision, initiated by the Movement Disorder Society, aimed to address weaknesses in the original UPDRS scale, and update to reflect recent scientific developments and clinical outlook. The full MDS-UPDRS is comprised of 4 Parts; I) non-motor score (13 items maximum total score of 52), II) motor aspects of daily living (13 items maximum total score of 52), III) motor examination (18 items maximum total score of 132), and IV) motor complications (6 items maximum total score of 24). Each item is rated on a scale from 0-4 (except Hoehn Yahr stage question which is 0 [asymptomatic] -5 [wheelchair bound/bedridden]). Historically MDS-UPDRS motor scale components (Parts II/III)) have been identified as the PD endpoint with optimal signal to noise ratio for detecting treatment effects, therefore most frequently used as trial endpoint in current day[16-18]. However, this is dependent on stage of disease and may change as the field advances.

Alzheimer's disease and ADAS-Cog11 rating scale

Alzheimer's disease (AD) is a slow-progressive neurodegenerative disease that causes dementia, primarily affecting older adults. It manifests as a decline of cognitive function, language, memory and executive functions.

The proposed disease mechanism involves amyloid beta plaques and neurofibrillary tangles (NFT, tau protein) in the medial temporal lobe and cerebral cortex. These accumulations disrupt neuronal function and cause synaptic loss, neuroinflammation, and cortical atrophy[19-21]. Cholinesterase inhibitors and N-methyl-D-aspartate (NMDA) receptor antagonists have been the standard of care for many years; these are only symptomatic treatments. However, recent developments in disease-modifying therapies have emerged with Leqembi's FDA approval in 2023 to treat early-stage AD[22].

The Alzheimer's disease assessment scale (ADAS) is a rating scale used to assess cognitive and non-cognitive dysfunction in patients with mild to severe AD[23]. The ADAS-Cog portion of the scale is more widely used and is considered a gold standard for diagnosis in AD since the 1980s serving as the endpoint in clinical trials seeking to demonstrate cognitive benefit. The ADAS-Cog subscale has two primary variants; ADAS-Cog11 with 11 items developed to measure mild to moderate AD and ADAS-Cog13 which has two additional items that are more sensitive to changes especially in mild cognitive impairment and early-stage AD[23]. In Paper V ADAS-Cog11 is used. ADAS-Cog11 has 11 tasks (i.e., items) that are both subject and observer-based assessments. The cognitive domains that are captured by these tasks are memory, language and praxis. Each item has a different categorical or count scoring range; six items have a range of 0 to 5, and one item a piece for a range of 0 to 4, 0 to 8, 0 to 10, 0 to 12 and 1 to 5. Maximum total score for the assessment is 70.

1.4 Data Analysis

To date it is still common practice, to assess data from a COA serving as a clinical endpoint as a composite single measure at end of trial, focusing on change from baseline compared to control group (placebo/standard of care). These endpoints are analyzed through pairwise comparisons using statistical tests such as t-test or analysis of variance (ANOVA). The heterogeneity of the disease, and inter-rater variability associated with the COAs often results in highly variable endpoints making it challenging to evaluate treatment effect. This underscores a need for more novel methods or innovative application of existing methods to better extract meaningful information from the COAs.

1.4.1 Item response theory

Item response theory (IRT) is a statistical methodology which originated in the psychometrics field in the 1950s and later was widely applied in

educational standard testing and military examinations[24-26]. In recent years, its use has expanded to evaluate health outcomes[27] and presents an elegant way to use the item-level data from the COAs, providing a more informed analysis.

An IRT model consists of both assessment-specific and subject-specific characteristics, which are independent of each other. Item characteristic functions (ICFs) are non-linear functions that describe the relationship between an individuals unobserved trait, such as ability, disease severity, or health status, and the probability of a response to each item [24, 28]. This unobserved trait is also referred to as the latent variable (Ψ). This relationship enables estimation of a disease score that is based on a weighted average of the individual item data, effectively placing subjects along a disease severity continuum. Items with a stronger relationship to the latent variable (i.e., ability or disability) implicitly receive more weight and have a stronger influence on the disease status [29]

Within a NLMEM framework this subject specific latent variable (Ψ) is modeled through random effects and the item specific parameters are fixed effect parameters that are constant within each item.

ICFs are visualized through item characteristic curves (ICCs) for dichotomous items or dichotomized polytomous item, indicating the probability of endorsing a correct answer (e.g., yes, true) as a function of the latent variable. For polytomous items, ICFs can also be visualized through option or category characteristic curves (OCC) which visualize the probability of each response option across the latent variable range. These terms are often used interchangeably in the literature, however the interpretation depending on the data presented will vary. The appearance and shape of the ICCs or OCCs as well as the number of item specific parameters used to describe the latent variable and probability of response relationship are dependent on the mathematical model employed. Data can be described by a family of logit models (i.e., 1 – parameter logit [PL], 2-PL, 3-PL etc.) [30]

A 1-parameter logit model contains one item specific parameter, discrimination (*a*). This parameter demonstrates the item's ability to discriminate between high and low abilities and defines the steepness of the slope of the ICF. The higher the value, the stronger discrimination between levels of ability. In a 2-PL model a second item parameter, difficulty (*b*) is added. The difficulty parameter defines the location on the latent variable scale where there is a 50% probability of responding correctly or 1 to a binary item. A 3-PL model includes a guessing parameter (*c*), which includes the probability of obtaining the answer with no prior knowledge or level of ability.

Figure 1 presents an example of the ICCs from a 2-PL model for 5 binary items with a subject response pattern of 0,1,1,0, and 1. The discrimination and difficulty parameters for each item are as follows: Item 1 (a=3, b=1.5), Item 2 (a=2.5, b=1), Item 3 (a=1.5, b= 0), Item 4 (a=2, b=2), and Item 5 (a=1, b=1.5).

Item 1 has a steeper slope and is more discriminatory across ranges of ability while Item 5 is the least discriminative item. Assuming scale directionality of greater than zero being higher disability, Item 4 is more difficult or less likely to be true than Item 5 until a higher degree of disability is met. This allows insight into items that may contribute more to the understanding of the disease at specific disease stages of the patient population.

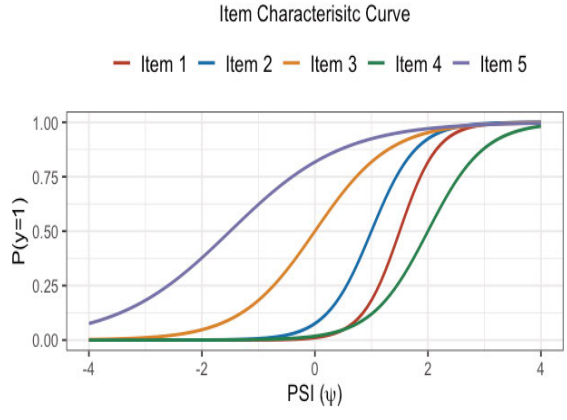

Figure 1. Example ICCs to demonstrate core IRT concept

In IRT modeling, items with polytomous or ordered categorical sets of responses are often described by graded response model (GRM) or generalized partial credit (GPC) models[31]. Graded response models require that each task, level of impairment, or criteria must be reached before moving to the next level of response. Generalized partial credit is used where a task can be divided in independent subtask and therefore can partially receive credit implying a continuum of performance, rather than ordinal ranking[31].

In this thesis, analysis of ordered categorical data was performed using GRM (Equation 1). The discrimination parameter is constant across all response categories, within an item but not constant across items. In a GRM the difficulty parameter is estimated for each response category, and are referred to as the difficulty thresholds. The difficulty threshold is the location on the latent variable scale where there is 50% probability to respond at or above a particular response category[32]. GRM treats the category responses as a series of dichotomous options. For example, a 4-point scale with response

options 0-3 will be dichotomized as follows: 0 v s. 1,2,3 ; 0,1 vs 2,3; and 0,1,2 vs 3. These estimated probabilities for each category are represented in equation 1.8 with the final cumulative probability equal to 1. The parameterization of a graded response model is as follows:

For subject i and item j, the graded response model describes the probability of achieving a response of at least s at time point t_k (if longitudinal data) as

$$P\left(Y_{ijk} \geq s|\Psi_i(t_k)\right) = \frac{e^{a_j(\Psi_i(t_k)-b_{j,s})}}{1+e^{a_j(\Psi_i(t_k)-b_{j,s})}} \quad (1.7)$$

and, the actual probability to achieve a score of exactly s as

$$P\left(Y_{ijk} = s|\Psi_i(t_k)\right) = P\left(Y_{ijk} \geq s|\Psi_i(t_k)\right) - P\left(Y_{ijk} \geq s+1|\Psi_i(t_k)\right) \quad (1.8)$$

where $\Psi_i(t_k)$ is the subject's time-dependent latent variable value, a_j is the item-specific discrimination parameter and $b_{j,s}$ is the threshold parameter for a specific item and score[33].

The response data on the ordinal scale is transformed to the interval scale, through the logit function, as shown above. Therefore, parameters are measured on the logit scale, which spans from negative to positive infinity. This means also that the item parameters are interpreted in relation to the latent variable or unobserved trait being measured. To ensure identifiability when estimating item parameters and the latent variable, a reference point for the population evaluated is set to a mean of zero with a standard deviation of 1 on the latent variable scale. The directionality of the scale is informed by this reference, and in the current context a latent variable >0 indicates more impairment or higher disability. Throughout the papers associated with this thesis the latent variable may be also referred to as disease severity scale.

Core assumptions of item response modeling

There are several key assumptions in terms of item response modeling: i) local independence between items, that is, a response to one item does not influence the response to another item, ii) the monotonicity- probability of endorsing a response will increase as latent variable increases, and iii) ICFs do not change in different populations[24, 28]. While these assumptions are foundational, there are times where there are exceptions and the rationale for this should be understood and addressed where applicable. Another key assumption is unidimensionality. In COAs, for example, it is equally important to acknowledge the concept of dimensionality. The number of latent variables or dimensions

that describes a disease construct can be many and are dependent on both the assessment and the subjects who have taken the assessment[34]

Unidimensionality implies that a single latent variable is the only shared factor between items influencing the subject's response probability[30, 34]. Factor analysis is the most common method used to assess dimensionality (i.e., number of latent variables) for the items in a scale and well-designed assessments are considered unidimensional if one dimension is the primary driver of a subject's probability of response. If other dimensions exist and are not related to the dominant construct then the unidimensionality assumption is violated [35]. Therefore, when a unidimensional model is applied to multi-dimensional data one must be careful as the meaning of the average construct will deviate from the original main construct. Under these circumstances use of a multidimensional IRT should be leveraged.

In pharmacometric applications, the number of latent variables is often defined based on clinical arguments (e.g., motor score or sidedness in Parkinson's disease) rather than a more formal factor analysis. However, one can imagine that realistically, most assessments are measuring more than one dimension, with potential for one or two primary dimensions driving the main construct. This should be considered during analysis and interpretation of results.

Fisher Information

Fisher information quantifies the information content of data used for parameter estimation, which is inherently linked to the weighting of data during maximum likelihood estimation. This weighting represents how much the data in our sample informs us about the population from which the sample is drawn.

The Cramer-Rao bound of inequality sets a lower bound for the variance of the estimator of the parameter based on the fisher information, to describe the achievable degree of precision given the data[36, 37]. This lower bound provides confidence that the estimate is closer to the true value (i.e., more precise). Therefore, a parameter that is estimated precisely provides more value than a parameter that is estimated with less precision.

In IRT, Fisher information can serve as a metric to understand the relative informativeness of different items. In this way, Fisher information quantitates how much data for each item is contributing to a particular ability level.

The amount of item information depends on the model structure of the ICF applied. An item is the most informative when the item's discrimination parameter is high[38]. Due to this relationship between ICF and item information, a change in Fisher information would indicate a change in the underlying item parameters with regards to the relative measure of ability.

IRT application in drug development

IRT in drug development offers a powerful and adaptable modeling framework to analyze and optimize COA-related endpoints, providing additional insights to facilitate the drug development process. At the time of this work there were several published examples of pharmacometric application of IRT, including but not limited to MDS-UPDRS in PD[39-41], Positive and Negative Syndrome Scale (PANSS) in schizophrenia[42], ADAS-COG score in AD[43],FACT-B in breast cancer[44], and Chronic Obstructive Pulmonary Disease (COPD)[45, 46]. The findings indicate promising application of IRT to characterize aspects of disease progression and treatment outcomes, with increased measurement precision demonstrating efficient utilization of data, lessened impact of missing data and higher power to detect drug effect than traditional methods[30, 39-46].

1.4.2 Artificial Neural Network Classification

In Paper IV we aimed to use item-level data to classify patients as either having PD or PD-Like symptoms without diagnostic evidence of disease. Item response theory served as the primary analysis method. This research also explored other emerging technologies, such as machine learning, expanding the scope of the research.

Machine learning is in the early stages of adoption in the pharmacometrics field, however its ability to learn patterns and relationships from high-dimensional input and output data to produce accurate predictions brings great value. Several machine learning approaches are commonly used for classification models, including logistic regression -which is commonly used in pharmacometrics- as well as decision trees, random forest, K-nearest neighbors (KNN) etc. The approaches differ based on data type and how data is processed. This thesis focuses on the implementation of artificial neural network, chosen for their ability to model complex patterns and relationships in data.

An artificial neural network is a computing algorithm that is based on the concept of the structure and function of the nervous system and biological neurons comprising interconnected nodes (neurons) organized in layers [47]. Similar to biological neurons, nodes receive and send information. The input layer receives the input data and the number of nodes required are equal to the number of features (covariates or predictors) in the data. Data is then passed through the hidden layers where it is processed to extract different features and patterns. Typically, if a feedforward neural network is used each layer passes its output to the next layer. In a fully connected (dense) network every node in one layer is connected to every node in the next layer. However,

choice of a dense or sparse layer is dependent on the analyst needs. There are no strict limits on the number of hidden layers, but guidelines for the number of hidden nodes are relative to the number of nodes in the input and output layer[48, 49]. Lastly, the output layer contains nodes that produce the predictions based on the processed data. For a binary classification model, the output layer typically has 1 node which will output a probability value between 0 and 1. Neural networks can be used for both supervised and unsupervised learning applications. Training of an ANN is done by iteratively adjusting the weights of the connections between the nodes using gradient-based algorithms, then the output is passed through an activation function generating an output for that node. The loss function then quantifies the difference between the predicted outputs and the true value and serves as a minimization function to indicate model performance- similar to the objective function value in NONMEM.

1.5 Clinical Trial Simulation

A key benefit of pharmacometrics models is their capacity to perform simulations to evaluate various scenarios for methodological investigations or clinical trial simulations to evaluate potential outcomes of a future trial to inform clinical trial designs[50, 51]. These simulations can be used to optimize trial design, quantify uncertainty around trial outcomes to inform decision making, assess alternative scenarios for comparing treatments, and to provide additional insights when clinical data is limited or even eliminate the need for a clinical trial by answering the question in silico. This tool enables risk-based decision making, supports cost-effective trial design and can help reduce the risk of trial failure. A single trial is a random event that is drawn from a distribution of possible trial outcomes. Therefore, Monte Carlo simulation is often used, as this technique performs repeated random sampling from a known distribution to account for uncertainty in the modeling system. In NLMEM this typically involves sampling from the ETA distribution based on the final model to generate unique data. This process is repeated a large number of times (e.g., >200), to simulate a range of possible *trials*. The results from each replicate are then analysed to obtain more precise estimates or to support hypothesis testing.

1.6 Hypothesis Testing

Sample size calculations are performed to determine the sufficient number of subjects to include in a clinical trial in order to detect a statistically significant effect (usually treatment effect), based on the primary endpoint.

Traditional approaches to determine sample size are either precision based (estimation) or power based (hypothesis testing). The sample size for any trial depends on aspects of the trial design, including acceptable level of significance, specified power for the trial, expected effect size, underlying event rate in the population and variability of the data in the population. Additional considerations can also include factors such as dropout rate and unequal allocation sizes[52]. To mitigate some uncertainty, trials may also over enroll subjects to ensure sufficient power in the case of potential dropout.

In hypothesis testing, a hypothesis is defined based on the primary endpoint of the trial. The null hypothesis (H_0) is the default assumption that there is no effect, or no difference between treatment arms. The alternative hypothesis (H_1) is that there is an effect (preferably a positive effect for investigational drug compared to control) or significant difference.

$$null\ hypothesis \qquad H_0: Effect\ \mathrm{TRT_A} = Effect\ TRT_B \tag{1.9}$$

$$alternative\ hypothesis \qquad H_1: Effect\ TRT_A \neq Effect\ TRT_B \tag{1.10}$$

The trial is designed to provide evidence for this and needs to be sufficiently powered to confidently determine whether to reject or fail to reject H_0. The significance level, α, is the probability of rejecting the null hypothesis when it is true, which is known as Type I error. This corresponds to false identification of drug effect, when no true effect exists. A Type II error (β) is not identifying an existing effect or failing to reject the H_0 when it is false. The probability of not making a type II error is defined as power (1- β)[53] [32]. At least 80% power is considered to be acceptable for a potential trial, but often drug developers aim for higher than this value. The ideal sample size should have large power to detect drug effect, and small probability for a Type I error.

Building on this, in model-based analyses, statistical criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC) and Objective function value (OFV) are commonly used as model selection criteria to compare goodness of fit for regression models. The likelihood ratio-test statistic (or log likelihood ratio) is a hypothesis test used to determine if the addition of a specific parameter(s) - such as drug effect - significantly improves the fit compared to the simpler model. This test compares the goodness of fit between two nested models; typically, a restricted (null) model and an unrestricted (alternative) model.

The likelihood ratio test (LRT)

$$tLLR(\Theta) = \mathcal{L}\left(\hat{\theta}_{unrestricted,full}, y\right) - \mathcal{L}\left(\hat{\theta}_{restricted,reduced}, y\right) \quad (1.11)$$

Or in the case of using OFV from a model

$$\Delta OFV = OFV\ full\ model - OFV\ reduced\ model \quad (1.12)$$

Follow chi square distribution with k degrees of freedom to determine the cut off

$$e.g., \Delta OFV > 3.84\ for\ p = 0.05\ and\ 1\ degree\ of\ freedom$$

If the ratio or difference is considered as a significant reduction in OFV compared to reduced model (>3.84) then it suggests the full model is significantly better at explaining the data. This assessment can also be employed to calculate power and type I error in a simulation and estimation paradigm., where the LLR test statistic is used to test for frequency of rejecting or failure to reject the null hypothesis (H_0) for each replicate trials.

2 Aims

The application of IRT in pharmacometrics to characterize disease is in its early stages. This thesis aims to advance the understanding and provide a methodological foundation for successful implementation of IRT in pharmacometrics, to analyze data from rating-scale based clinical outcome assessments to facilitate model-informed drug development.

The aims can be divided into two focus areas I) Foundational Methodology II) Clinical Application

I. Foundational Methodology – Assess the robustness of IRT models across varying sample sizes, assessment lengths (including shortened assessments), progression rates, drug effects and estimation strategies.
    - Accuracy and precision of item and latent variable parameters
    - Compare the performance of IRT models using Laplace and Gaussian-Hermite Quadrature (GHQ-EM) estimation algorithms

II. Application to clinical questions

Evaluate the utility of pharmacometric IRT for addressing clinical questions, including trial design considerations, its sensitivity as an analytical method for clinical endpoints and potential as a classification tool for patient diagnosis, specifically, Parkinson's disease.

3 Methods

The methods used in this thesis are generally represented in the order of model-building sequence, reflecting the foundational methodology (papers I-III), followed by clinical applications and assessment of treatment effect (papers IV-V).

3.1 Data

3.1.1 Clinical Data

The primary source of clinical data used in this thesis came from the open-access database of the observational Parkinson's Progression Markers Initiative study (PPMI, www.ppmi-info.org/data). The aim of PPMI is to identify biomarkers that can predict the risk, onset and the progression of Parkinson's disease, through a natural history study. The initiative, which is driven through collaborative partnerships provides a comprehensive and standardized open-access dataset that includes longitudinal data to accelerate breakthroughs for more effective treatments.

In paper I and IV MDS-UPDRS item level and total score data was leveraged. In paper I longitudinal data (motor sub-scale only) from 423 de novo PD (DeNoPD) patients (not used PD medication for more than 60 days prior to baseline) was used. More details in protocol[54]. In paper IV, item level and total score MDS-UPDRS (all parts) were analyzed at screening, baseline and post-baseline for both de novo PD patients (N=452,430 respectively) and 66 subject with scans without evidence of dopaminergic deficit (SWEDD). SWEDD patients present with PD-like symptoms but do not have PD. In both papers, observations from assessments completed pre-dose, up to 48 months were included, based on the available data at the time of extraction from database. Model parameters that were estimated based on this data were the basis for simulations performed in paper I and III.

In paper V, the clinical data was sourced from the Merck Sharp &Dohme (MSD) sponsored EPOCH trial (NCT01739348) a randomized, double-blind, placebo-controlled trial evaluating the efficacy of verubecestat in participants

with mild-to-moderate AD[55]. Trial data included 4 treatment arms: placebo, 20 mg, 40 mg, and 60 mg. In an earlier statistical analysis at a total score level, unrelated to this thesis -no dose response was found. ADAS-Cog11 data from a total of 2,210 patients from both placebo and active arms, across ~8 visits up to 78 weeks were analyzed. At the time the work was performed, access to data was granted as an employee of the company.

3.1.2 Simulated data

Simulated data was used in several projects. The model developed in paper I was adapted and used for a simulation study in paper III. The final estimates of the item parameters from the first 20 items of the MDS-UPDRS motor scale, along with the latent variable, progression rate and interindividual variability served as the foundation for simulation parameters. This new scale included seven 4-category items (s=0-3), ten 3-category items (s=0-2) and three 2-category items (s=0/1).

In paper II, generic item level data at a single timepoint (baseline) was simulated from item parameters that were randomly sampled from pre-defined distributions, using the R package *mirt* (version 1.31). The discrimination (*a*) item parameters were randomly sampled from a log-normal distribution (meanlog=0.05, sdlog=0.5). The threshold (*b*) parameters were randomly sampled from a uniform distribution increasing from -2.5 (b1) at the lowest threshold to 2.5 at the highest threshold (b4). Simulations of response categories (0-4) were repeated to ensure that all possible categories were represented for each item. Replicate datasets were simulated for each sample size and item scenario from the unique set of simulated item parameters.

Table 1 Description of simulation studies

Paper	Data type	Time	Clinical Endpoint	Items	Max total score	Number of Simulations	Number of Scenarios
II	Item	baseline	Generic	5,20	20 or 100	1000	6 scenarios 2 methods
III	Item	longitudinal	Generic	20	44	500	8 scenarios, 2 methods
V	Item/total	longitudinal	ADAS-Cog11	11	70	200	1 scenario, 3 methods

3.2 Models

In general, exploratory data analysis (EDA) was first performed to determine potential trends in response scores. Modeling and simulation was performed using NONMEM (ICON plc,v 7.4.2 or higher) facilitated by PsN (v4.7.15 or higher) or with the R (v3.5.2 or higher) or python-3.10.12. Data handling, visualization and summarization was completed in R. To assess the robustness of IRT models and their application to clinical questions, we conducted a comprehensive methodological evaluation. The various scenarios are summarized in the table below.

Table 2 Methodology Summary

Paper	Aim	Methodology	Clinical Endpoint	Patient Population	Real/Simulated	Items	Key Assessments
I	Shortened assessment	Simultaneous IRT	MDS-UPDRS motor	Early PD	Y/Y	34	Information, Drug effect, Progression rate
II	Estimation of ICFs	Laplace and GHQ-EM	Generic	N/A	N/Y	5,20	Item parameters
III	Estimation of ICFs and latent variable	Simultaneous and sequential IRT	Generic	N/A	N/Y	20	Drug effect, Progression rate, Type I error/power
IV	Classification	IRT,ANN, total score	MDS-UPDRS	Early PD/SWEDD	Y/N	68	Classification
V	Trial design/Power	Pairwise comparison, IRT composite score disease progression modeling	ADAS-Cog11	Mild-moderate AD	Y/Y	11	CTS,power

3.2.1 Estimation methods

In this thesis three main estimation methods were used; first order conditional estimation (FOCE/I, papers IV-V), Laplace (paper I-V) and Gaussian Hermite quadrature- Expectation Maximization (GHQ-EM, paper II)[56, 57].

In all papers model evaluation included standard GOF measures such as AIC, OFV, GOF plots (i.e., DV vs IPRED) where applicable. Additional model evaluations such as simulation-based diagnostics like mirror plots, GOF of ICCs assessed through generalized additive model (GAM) cubic spline, and visual predictive checks (VPC) were also leveraged.

FOCE

Although the total score data is a discrete value that is bounded, it is often treated as continuous in disease modeling. Therefore, it was modeled using FOCE (paper IV) without and with interaction (paper V). FOCE is a gradient based method that approximates the likelihood with respect to the inter-individual random effect (η), using first-order taylor expansion around conditional mode of the random effects. This conditional mode maximizes the joint density (that is the combined likelihood of observed data and random effects) for an individual. This process iterates, until convergence is achieved.

Laplace

IRT model estimation for both cross-sectional data (i.e., baseline) and longitudinal data was performed using Laplace estimation in NONMEM. Laplace is similar to FOCE except that it uses second order Taylor expansion for the conditional linearization with respect to each random effect (η) during minimization. In NONMEM, Laplace is best suited for categorical data.

Gaussian-Hermite quadrature (GHQ)-Expectation maximization (EM)

Gaussian Hermite quadrature (GHQ) is a type of gaussian quadrature that uses a weight function based on the standard normal distribution to approximate the integral. GHQ uses pre-specified points (i.e., quadratures) and corresponding weights across the distribution of the latent variable. Rather than calculate the integral directly GHQ evaluates the likelihood at the points and takes a weighted sum to approximate the integral. [58, 59]. The grid points are typically centered around zero, when working with standard normal distribution; values closer to zero are given more weight[59]. In mirt GHQ is employed as part of an expectation maximization algorithm, the sum of the approximated individual marginal likelihoods is then optimized in the second step.

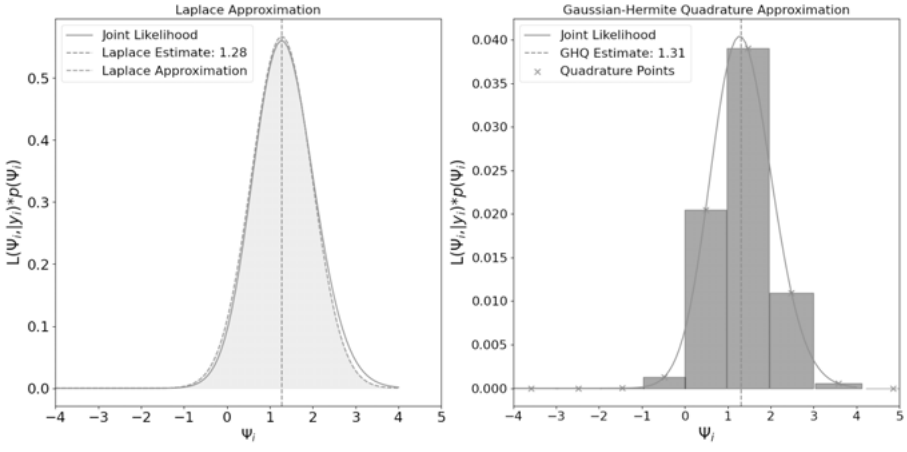

Figure 2 Illustration of the Laplace (left) and GHQ-EM joint likelihood approximation (right)

3.2.2 Longitudinal IRT models

The structures for item characteristic functions can be categorized under two groups; graded response (papers I-V) or 2-PL (Paper IV).

IRT implemented in a disease progression framework for pharmacometrics application, from here forward will be referred to as "longitudinal IRT". The longitudinal IRT model structure used in Paper I, III and IV is adapted and extends the IRT models for Parkinson's disease developed by Gotipatti et al [41] and Buatois et al[40]. In paper V the longitudinal IRT model for Alzheimer's disease was developed denovo.

Unidimensional IRT models were used throughout, except in paper IV a three latent variable model was used. Item parameters were modeled as fixed effects, while the latent variable was modeled through random effects. In all methods disease severity at baseline was assumed to follow a normal distribution on the latent variable (Ψ) scale with a mean of 0 and variance of 1. Linear disease progression rate was assumed and modeled through random effects. The linear model treats the change of disease status as a constant, as described in equation 3.1 below:

$$\Psi_i(t_k) = \Psi_{\mathrm{i}}^0 + \alpha_i t_k \tag{3.1}$$

Where, Ψ_i^0 is the individual (i) disease severity at baseline, α_{i} is the disease progression rate, and t_k is a specific time point.

In paper I which used data from Parkinson's patients, it is understood that the available PD therapies provide symptomatic treatment effects. Similarly, in paper V, the interventional therapy evaluated for treatment of mild-to-

moderate Alzheimer's was also symptomatic. Therefore, in papers I and V, an average symptomatic offset effect was estimated on the latent variable, assuming drug effect remained constant throughout the disease trajectory (equation 3.2)[60].

$$\Psi_i(t_k) = \Psi_{\mathrm{i}}^0 + \alpha_i t_k + E_i^0 \tag{3.2}$$

E is the drug effect that shifts the overall patient disease status and can take different structures as appropriate. In paper I, IIV for drug effect was also estimated. The data used in paper V included three dose groups, and therefore different structures of drug effect were evaluated to test for dose-response, including Emax and sigmoidal Emax.

In the simulation study in paper III, a disease modifying drug effect was evaluated. Specifically, a drug effect of 0.18 was applied proportionality to the disease progression rate as shown below:

$$\Psi_{i,k} = \Psi_i^0 + \alpha_i * (1 - DMEFF) * t_k \tag{3.3}$$

DMEFF (disease modifying drug effect) was assigned 0 at baseline for the treatment arm, placebo and fixed to zero in the model without drug effect. The disease modifying effect was assumed as a population effect only. Additional information on the methods used to define the ICFs for this simulations study can be found in the next section.

3.3 Item Information and Efficiency

In paper I, item information and efficiency were used to evaluate the performance of longitudinal IRT models in shortened or partial assessments, using both real and simulated data. In paper I, fisher information at the item level was first used to rank the items of the MDS-UPDRS motor subscale using the model with all-items included (100% scenario) estimated from real data. The rank order defined by the real data was assumed for the simulated data as well. The general workflow is shown in Figure 3. To assess the impact of the reduced assessment at the item level (i.e., item response model), item information represented by efficiency (i.e., the ratio of fisher information for reduced and full scenario) was used as the comparison metric for all scenarios. Change in mean disease progression and drug effect estimates were used to assess stability of the latent variable.

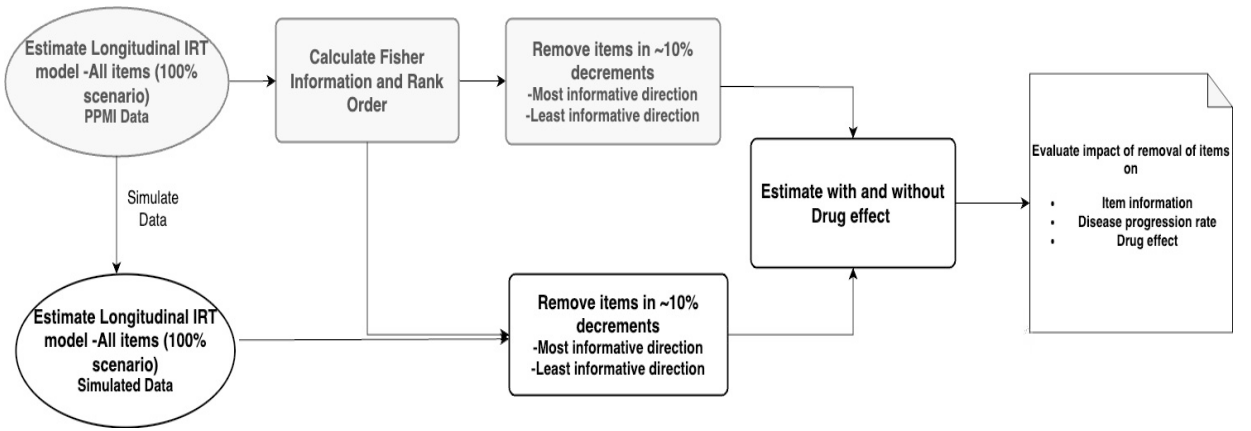

Figure 3 General workflow for Paper I

For each item *j* the item information function derived as a function of the latent variable, was calculated as minus the expectation of the second derivative (∂^2)of the log-likelihood[61], i.e.

Eq.1

$$I_j(D_i) = -\sum_{s=0}^{s_j} P\left(Y_{ij} = s|\Psi_i\right) \frac{\partial^2 \log P\left(Y_{ij} = s|\Psi_i\right)}{\partial {\Psi_i}^2} \tag{3.4}$$

where $P\left(Y_{ij} = s|\Psi_i\right)$ is the response (s) probability for the disability Ψ_i as defined above. Furthermore, the population information, $\mathcal{J}_j$ was defined as the item information integrated over the disability range, i.e.,

Eq.2

$$\mathcal{J}_j = \int_{-\infty}^{\infty} \mathrm{p}(\Psi_i)\mathrm{I_j}(\Psi_i)\, \partial\, \Psi_i \tag{3.5}$$

where $p(\Psi_i)$ is the probability density of the latent variable distribution in the population.

3.4 ICF model determination

In Paper V, item parameters were first estimated, then fixed to their final estimates. Next, the longitudinal component (i.e., disease progression rate, drug effect) was estimated. In this case the item parameters were informed by all-time points from start of the trial. In the rest of the thesis there were two methods used to estimate the ICFs and handle the longitudinal data: "simultaneous" method and "sequential" method. When item parameters and longitudinal components (i.e., disease progression rate and drug effect) are estimated simultaneously this method is referred to as the "simultaneous" method (paper I and III). In paper III a direct comparison of these methods was performed. A high-level overview of the sequential method is described below.

Sequential method (Step 1):

A graded response IRT model structure was used to estimate ICF parameters, treating each observation from a single subject as a separate individual. At baseline the disease severity was descried as a mean of zero and variance of 1. A post-baseline shift (separate mean and variance) was estimated for observations after time zero as shown in the NONMEM code below.

```
Baseline:   IF (TIME.EQ.0) PSI=THETA(X)+ETA(1)
Post baseline: IF (TIME.GT.1) PSI=THETA(Y)+ETA(2)
```

Sequential method (Step 2): Directly modeling the latent variable time-course; drug effect etc.

The subject and time specific $\Psi_{i,k}$ (latent variables) estimated in step 1 were modeled directly for each original subject ID. Therefore, in this step the dependent variable is the value of Ψ

The uncertainty in the latent variable was also estimated to account in differences in precision, leveraging an approach by Lacroix et al (2012);

$$Y_{i,k} = \Psi_i^0 + \alpha_i * (1 - DMEFF) * t_k + SE_{\Psi_{i,k}} * \varepsilon_{i,k} \qquad (3.6)$$

Ψ_i^0 is the baseline latent variable value fixed to zero with an estimated additive random effect and ε_{ik} is a zero mean random variable (RUV), other variables as defined above.

3.5 Estimation methods for Item parameters (Paper II and Paper III)

In paper II, a simulation study was performed to compare the performance of Laplace and GHQ-EM for estimation of item parameters, across different scenarios, using one observation per subject (assuming baseline). Four sample sizes (50,100,250,500) were evaluated and two assessment lengths. The general workflow is shown in figure 4.

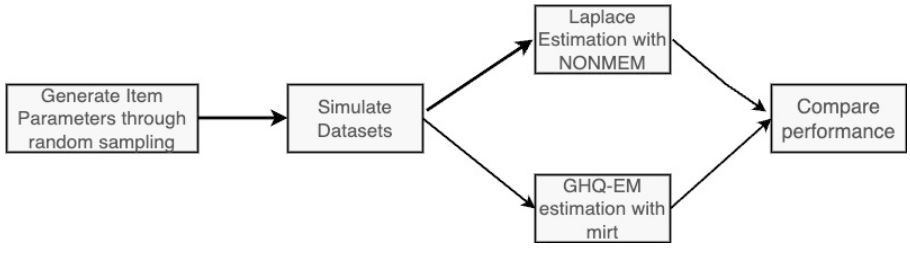

Figure 4 Workflow Paper II

NONMEM model control files were autogenerated using the R package-piraid (pharmacometrics item response theory aid, https://github.com/UUPharmacometrics/piraid, Uppsala University). This software tool bridges the gap between the current psychometric tools, which rarely accommodate longitudinal data and the pharmacometrics tools like NONMEM, where implementing IRT models can be challenging. The package is comprised of three main components: i) Scale, ii) the assembler which is responsible for creating NONMEM-IRT model and iii) the Inspector, responsible for the generation of diagnostic plots and other statistics[62].

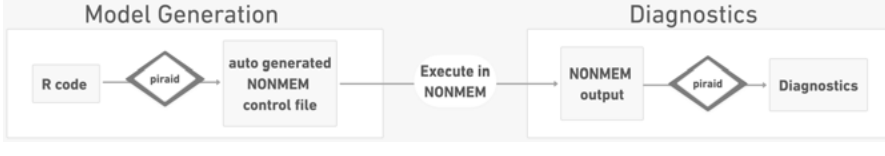

Figure 5 piraid Workflow

The *mirt* open-source R package(v1.3.1), which is used for performing multidimensional item response models, was utilized to implement the traditional fixed GHQ-EM method. Initial estimates in NONMEM and *mirt* were set to achieve similar starting conditions for both software. In NONMEM an upper bound for the threshold of fixed effects was set to 10 and the latent variable was modeled through subject specific random effect, assuming a normal

distribution. Model performance was assessed in terms of parameter estimation accuracy and bias, as outlined in the equations below.

$$E_{j,s,r} = \hat{\theta}_{j,s,r} - \theta_{j,s,r} \tag{3.7}$$

$$Bias_{j,s} = \frac{1}{N\ trials} \sum_{r=1}^{N\ trials} e_{j,s,r} \tag{3.8}$$

$$RMSE_{j,s} = \sqrt{\frac{1}{N\ trials} \sum_{r=1}^{N\ trials} e_{j,s,r}^2} \tag{3.9}$$

Where, *j* is item*, s* is scenario, *r* is replicate , *e* is estimation error, $\hat{\theta}_{j,s,r}$ is the estimated item parameter and $\theta_{j,s,r}$ is the true item parameter. A version of the RMSE (rRMSE) was also evaluated, which removed outliers. Additionally, log-likelihood, run time and completion rate were also evaluated.

To assess the resulting bias and precision impact on total score, the expected total score (TS) was calculated as the sum of the probabilities of selecting each response category (Y_{ijk}) as a function of ability (Ψ) for the test, i.e.,

$$TS_{s,r} = \sum_{j=1}^{M_s} P(Y_{ij} = 1) + P(Y_{ij} = 2) \cdot 2 + P(Y_{ijk} = 3) \cdot 3 + P(Y_{ijk} = 4) \cdot 4 \tag{3.10}$$

In paper III, estimation error, bias and precision (SD) for ICF parameters, progression rate and drug effect were evaluated as described above for each model scenario and replicate trial.

3.6 Classification Models and Machine Learning approaches

The methodology evaluated in the classification study (Paper IV) utilized a mixture model for total score, and for the analysis of item level data, IRT and artificial neural networks (ANN) were employed. Figure 6 describes the workflow in paper IV.

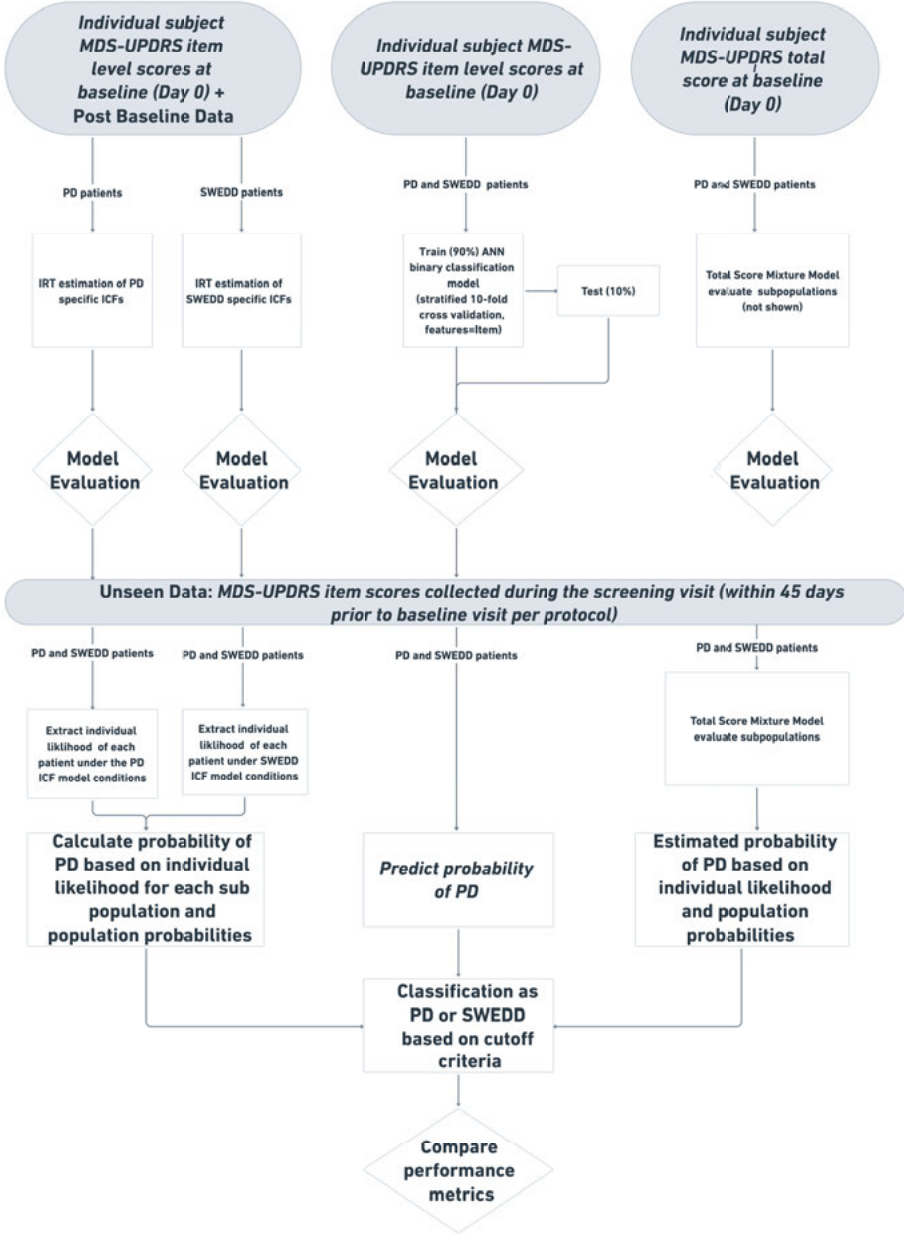

Figure 6 Classification model workflow (paper IV)

A longitudinal IRT model, assuming a linear disease progression with three latent variables (patient reported or sided and non-sided clinician-reported items) previously developed for Parkinson's disease patients was adapted for use for classification[41]. To expand this model to SWEDD subjects, item parameters were re-estimated in a SWEDD only model to obtain SWEDD specific ICFs. This resulted in the development of two separate models to classify DeNoPD and SWEDD patients.

In order to estimate or calculate the probability of DeNoPD, population probabilities (i.e., proportions) of 0.845 for DeNoPD and 0.155 for SWEDD were leveraged. This proportion reflects both the PPMI dataset and cases found in literature where subjects were misdiagnosed as PD, and were later determined to be SWEDD subjects after clinical trial enrollment.

A mixture model was applied to baseline and screening MDS-UPDRS total score data separately, excerpt from NONMEM code as shown below:

```
$PRED
  IF(MIXNUM.EQ.1)THEN
            Y=THETA(1)+ETA(1)+EPS(1)
            ELSE
  Y=THETA(2)+ETA(2)+EPS(2)
            ENDIF
EST=MIXEST

$MIX
  NSPOP=2
  P(1)=0.845      ;
  P(2)=1-P(1)       ;
```

The individual likelihoods (Li) were derived from the post-hoc individual -2 log-likelihoods (OFV). The probability of belonging to the DeNoPD cohort determined by both the mixture model for total score and the IRT model was defined as shown below:

$$\boldsymbol{P_{DeNoPD}} = \frac{\boldsymbol{L_{i,DeNoPD} * Prop_{DeNoPD}}}{\boldsymbol{L_{i,DeNoPD} * Prop_{DeNoPD} + L_{i,SWEDD} * Prop_{SWEDD}}} \quad (3.11)$$

In the IRT model, the likelihood of an individual to belong to a specific cohort was assessed by applying the model with the PD specific ICFs (LiDeNoPD) and SWEDD specific ICFS (LiSWEDD) to baseline and screening data separately. In the mixture model this was estimated directly during NONMEM execution using $MIXTURE subroutine which assigns a patient to the

subgroup with the highest probability. In the current versions of NONMEM this probability is output as PMIX variable in the psn.phm file.

An ANN for binary classification was also developed to differentiate between DeNoPD patients and SWEDDs leveraging each item as an input feature. A10-fold stratified cross validation using a 90/10 training and test split. Several architectures were explored to optimize model performance; number of hidden layers (1-4), nodes and learning rates. The output layer utilized a sigmoid activation function allowing for probability-based classification of individuals. In order to prevent overfitting a 20% dropout layer was used as a regularization technique. Comparison of training and test loss using the binary cross entropy loss function was used to monitor model performance. The final model also included a pre-training step.

3.6.1 Classifier performance

The likelihood of a subject belonging to the DeNoPD or SWEDD cohort was determined relative to probability cut-off thresholds. To assess the performance of the classifier the following statistical measures were used: sensitivity, specificity, precision, Matthews correlation coefficient (MCC) and receiver operating characteristic area under the curve (ROC AUC).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{True Positive Rate} \tag{3.12}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \text{True Negative Rate} \tag{3.13}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.14}$$

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP}) * (\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TN} + \text{FN})}} \tag{3.15}$$

Where, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

3.7 Covariate testing and Clinical Trial Simulations

The impact of covariates on the latent variable was assessed in paper V using a step-wise covariate modeling (SCM) approach. The covariates tested were age, gender, ApoE carrier status, MMSE baseline score and background use

of acetylcholinesterase inhibitor (AChEI) and memantine use. SCM criteria for selection were set to a p-value of 0.05 for forward inclusion and 0.005 for backwards elimination. In addition to the IRT model, a linear disease progression model for ADAS-Cog11 composite score was developed for comparison. This model was adapted from the model described by Conrado et al which incorporated a beta distributed residual variability[63]. The incorporation of the drug effect as well as SCM for covariate testing were performed similarly to the IRT model.

Monte Carlo simulations were performed in NONMEM, with item scores randomly sampled from a uniform distribution (Paper I, III and V). In paper I, a single simulated dataset was generated for evaluation using the estimates from the final model containing 100% of the information. Details of the CTS used in paper III and V are shown in the Figure 7.

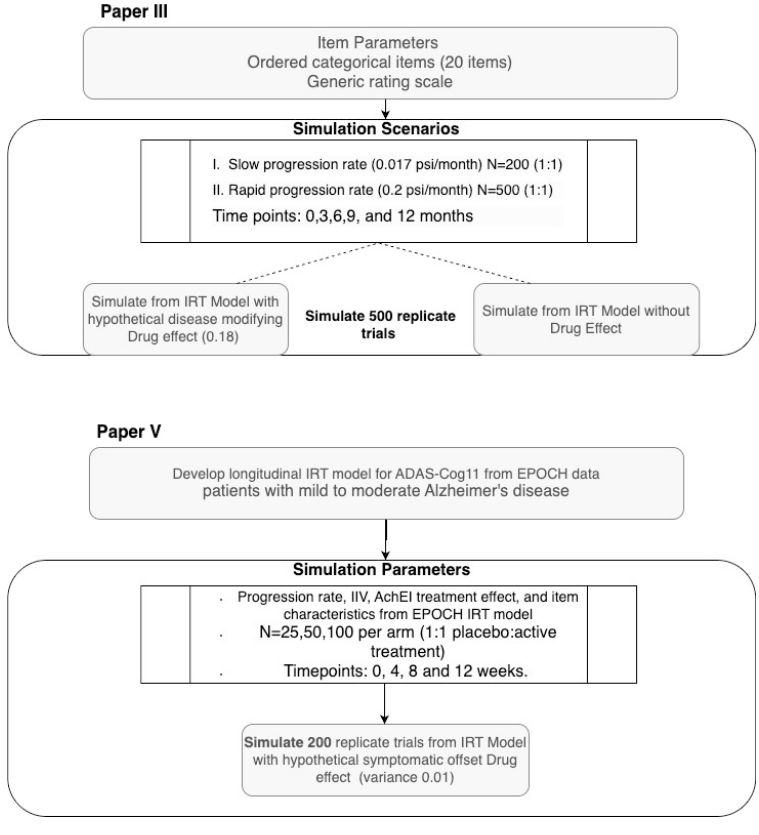

Figure 7 CTS workflow Paper III and V

In paper V, simulations were performed to illustrate the desired hypothetical drug effect of a different intervention than the drug originally modeled in the

IRT framework. Therefore, the drug effect value applied to the latent variable scale was determined by adjusting it to induce a 2-point change from baseline relative to placebo in ADAS-Cog11 composite score after 12 weeks of administration of the symptomatic treatment. The variability used in the simulations was based on the variability observed in the EPOCH trial and the typical variability in ADAS-Cog11 observed in a mild to moderate population [63]. For simulation, IIV was included on drug effect.

3.8 Hypothesis testing for Drug Effect

The ability of longitudinal-IRT models to accurately detect treatment effects while minimizing type I error under various model conditions was evaluated in paper III, with the aim of informing the comparison of two estimation strategies for determining item characteristic functions using a generic rating scale. Paper V evaluated the statistical power to detect symptomatic treatment effect utilizing the ADAS-Cog11 scale in a proof-of-concept, small sample size, short duration trial. Three methodologies for analyzing clinical endpoint data were evaluated i) pair-wise comparison for change from baseline at end of trial ii) longitudinal IRT model and ii) disease progression model for composite score.

For IRT and composite score disease progression models, the evaluation of type I error and/or power was performed by stochastic simulation estimation using maximum likelihood framework for hypothesis testing, facilitated using the “sse” command in PsN or R. Power was assessed using simulations from model with drug effect, and type I error was assessed using simulations from the reduced model. Each simulated dataset was estimated in both the full and reduced model. The significance level was set to 0.05. The type I error rate or power was determined by the proportion of replicate trials that met the significance threshold for the difference in objective function value (ΔOFV) using likelihood ratio test.

Table 3 Hypothesis testing for drug effect in paper III and V

Paper	Estimation approach	H0 (null hypothesis) Reduced model (no drug effect)	H1 (Alternative hypothesis) Full model (drug effect)
III	Simultaneous IRT	$\Psi_i(t_k) = BASE + SLOPE * TIME$	$\Psi_i(t_k) = BASE + SLOPE * (1 - DMEFF) * TIME$
III	Sequential IRT method with SE(step 2)	$Y_i = BASE + SLOPE * TIME + SE_{latent\_varaible} * \varepsilon_{i,k}$	$Y_i = BASE + SLOPE * (1 - DMEFF) * TIME + SE\_\Psi_{latent\_varaible} * \varepsilon_{i,k}$
V	Longitudinal IRT	$\Psi_i(t_k) = BASE * COV + (SLOPE * COV) * TIME$	$\Psi_i(t_k) = BASE * COV + (SLOPE * COV) * TIME + SYMEFF$
V	Composite score disease progression	$Y = BASE * COV + (SLOPE * COV) * TIME$	$Y = BASE * COV + (SLOPE * COV) * TIME + SYMEFF$

Slope=disease progression rate either on latent variable scale or total score scale

For pair-wise comparison of the composite score, a two-sample t-test was used to calculate power for detecting treatment effects across sample sizes to compare with the other methods in paper V. The t-test calculation was based on the target mean difference of 2-points in ADAS-Cog11 composite score and an assumed standard deviation of 5.8. The significance level was set to 0.05 (one-sided).

4 Results

The key results of this thesis are presented across the following themes: item characteristic function (Paper II, III), latent variable and latent variable time course (paper I, III), application to clinical questions (Paper IV, V). For more detailed results please see the perspective papers.

4.1 Item characteristic function (Paper II and Paper III)

A direct comparison of the Laplace method executed in NONMEM and GHQ-EM method executed in mirt, for the estimation of item parameters was performed. A single timepoint was assumed (i.e., hypothetical baseline) to inform the latent variable to simplify this evaluation. In general, low estimation error for the item parameters was observed for both methods, demonstrated by a median near zero. Laplace method trended towards increased precision for threshold parameters, but slightly larger mean estimation error for all parameters. As with other model-based analysis, a loss of precision was observed as less data, fewer subjects or items, was present to inform the model.

In the absence of longitudinal data, a minimum of 20 items and more than 100 subjects appeared to be necessary to yield more precise estimates of item parameters (Figure 8 and Figure 9). At smaller sample and item sizes bias increased and precision decreased both within and between algorithms. Higher accuracy in the estimation of threshold parameters was achieved using GHQ-EM, except when there were fewer subjects or items. GHQ-EM generally provided more accurate estimates for the discrimination parameters compared to Laplace. (Figure 8). The RMSE when removing outliers (1% extreme error values on either side of the distribution) is shown in Figure 9 as robust RMSE.

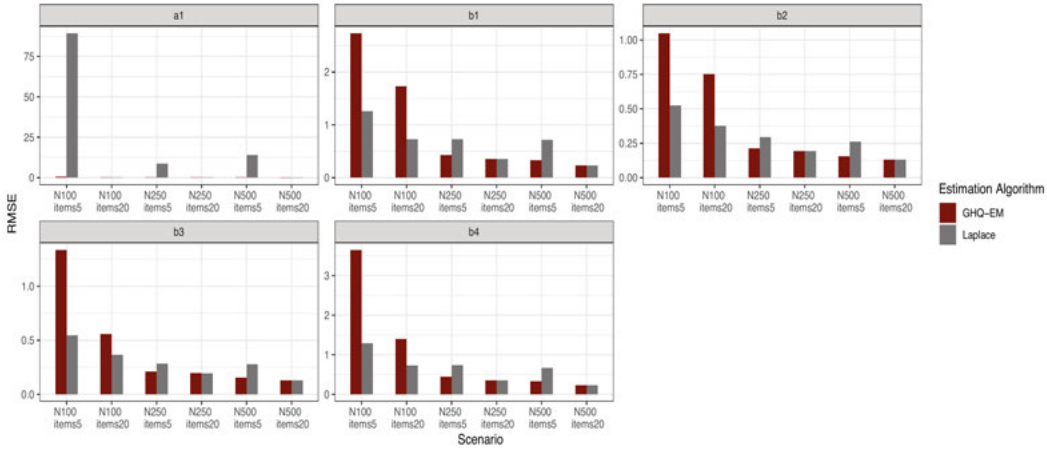

Figure 8 Item Parameter RMSE for each sample size and item scenario

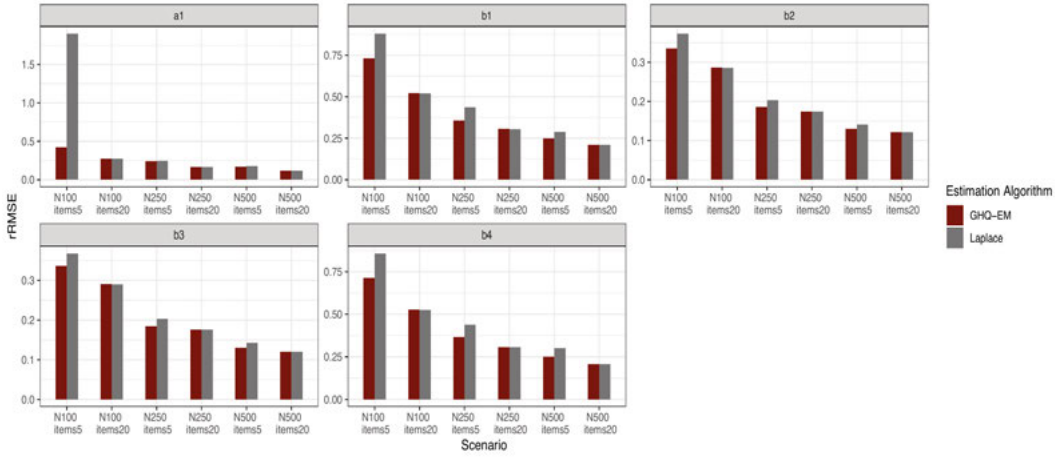

Figure 9 Item Parameter robust RMSE for each sample size and item scenario

In terms of estimated value for each parameter, a strong agreement between estimation algorithms was observed in the presence of at least 20 items. The estimation of the threshold parameter appears to be more influenced by the number of items than the discrimination parameter.

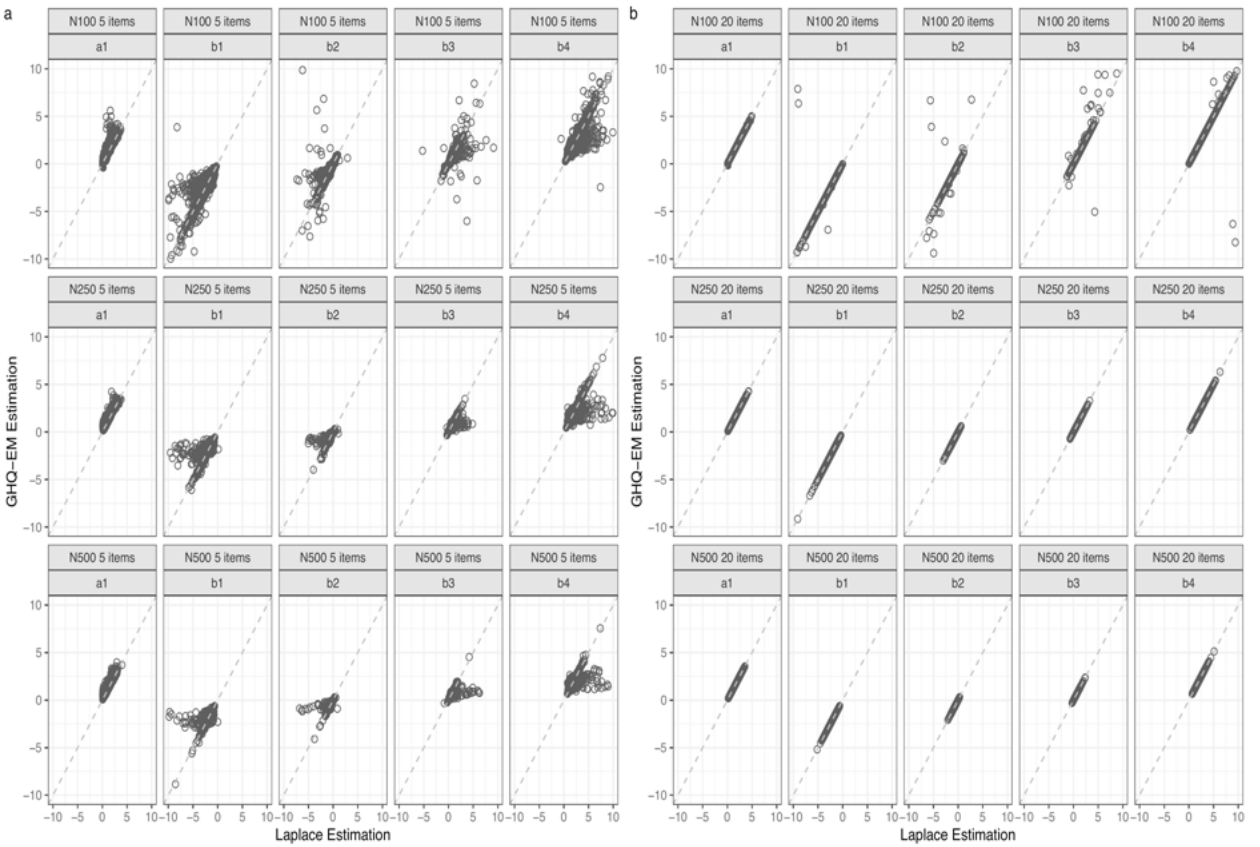

Figure 10 Observed Item parameter estimate comparison between Laplace and GHQ-EM a) 5 items b) 20 items

Despite the differences observed at the item parameter level, these differences were minimally observable on total score scale, indicated by the overlap of mean estimation error (2.5% and 97.5% percentile) for the expected total score. Small differences between methods could be observed at the tails of the latent variable. Comparison of the loglikelihood for the 5 and 20-items scenarios also indicated Laplace and GHQ-EM performed equally well for the scenarios with 20 items, demonstrating equivalent fit to the data regardless of sample size. Use of a fixed-grid of points for approximation resulted in GHQ-EM being computationally, many orders of magnitude faster than Laplace (19 s vs 5 minutes).

Building on the evaluation of estimation performance of item parameters using a single time point, the application of IRT within a disease progression framework introduces the complexity of longitudinal data. In paper III a generic 20- item assessment scale (max score 44), with 5 timepoints was used to evaluate this. Handling of post-baseline data was performed in two ways i) simultaneous estimation of the item parameters and longitudinal component ii) sequential as described in methods section. Minimal bias in the estimation of ICF parameters was observed and consistent magnitude of error was observed between simultaneous and sequential approaches under similar conditions (Table 4). As expected, an increase in sample size or progression rate increased estimation precision. The higher threshold parameters, where there was less data to inform the latent variable showed higher imprecision. In general, for both methods, the ICF parameters were well determined when the model was correctly specified. However, when using the simultaneous estimation approach, if there is misspecification in the longitudinal model there is a risk to introduce bias in the ICF parameters. An example of this is represented in the last column of Table 4.

Table 4 Across item aggregate summary statistics of item parameter estimation bias and precision

	Slow Progression Rate		Rapid Progression Rate		Rapid Progression Rate with Longitudinal Model Misspecification example
	Mean estimation error (SD)		Mean estimation error (SD)		Mean estimation error (SD)
Item Parameter	Simultaneous	Sequential	Simultaneous	Sequential	Simultaneous
DIS	-0.001(0.10)	-0.001 (0.12)	-0.003 (0.05)	-0.003 (0.06)	0.108 (0.06)
DIF1	0.004(0.17)	0.007(0.18)	-0.003 (0.09)	-0.004 (0.10)	-0.318 (0.17)
DIF2	0.033(0.25)	0.041 (0.29)	0.013 (0.12)	0.016 (0.14)	-0.283 (0.15)
DIF3	0.039(0.30)	0.041 (0.34)	0.015 (0.13)	0.007 (0.15)	-0.341 (0.15)

SD=standard deviation, DIS= discrimination, DIF1= fixed effect parameter for threshold parameter for category response 0 vs 1,2,3, DIF2= fixed effect parameter for threshold parameter for category response 0 ,1 vs 2, 3, DIF3= fixed effect parameter for threshold parameter for category response 0,1,2 vs 3

4.2 Latent variable construct (Paper I and III)

In paper III, the simultaneous IRT method consistently showed higher precision (Table 4) compared to the sequential method due to better-informed random effects (baseline and disease progression rate). This slight increase in precision for the simultaneous method was found to be attributed to the consistently lower uncertainty (SE at individual level) in the latent variable estimates compared to the sequential method. This was observed across all timepoints (Figure 11), with greater deviations at the extreme ends of the latent variable scale for the sequential method (Figure 12). This suggests that, while ICF parameters are estimated accurately with little bias and sufficient data is available to inform these measures, the latent variable may be less precisely determined for the sequential method compared to simultaneous.

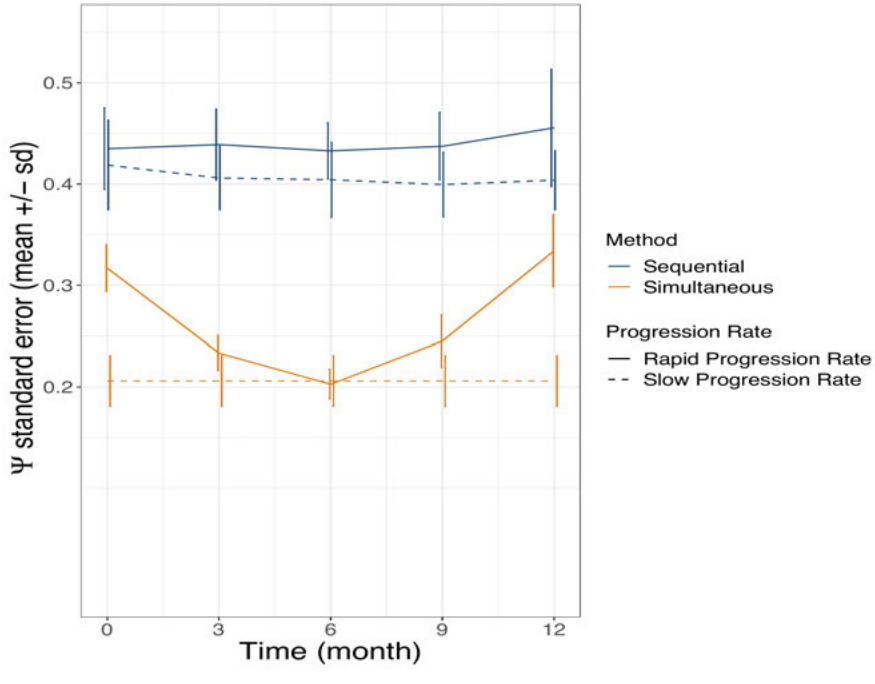

Figure 11 Representative trial Latent variable (Ψ) standard error (SE) time-course

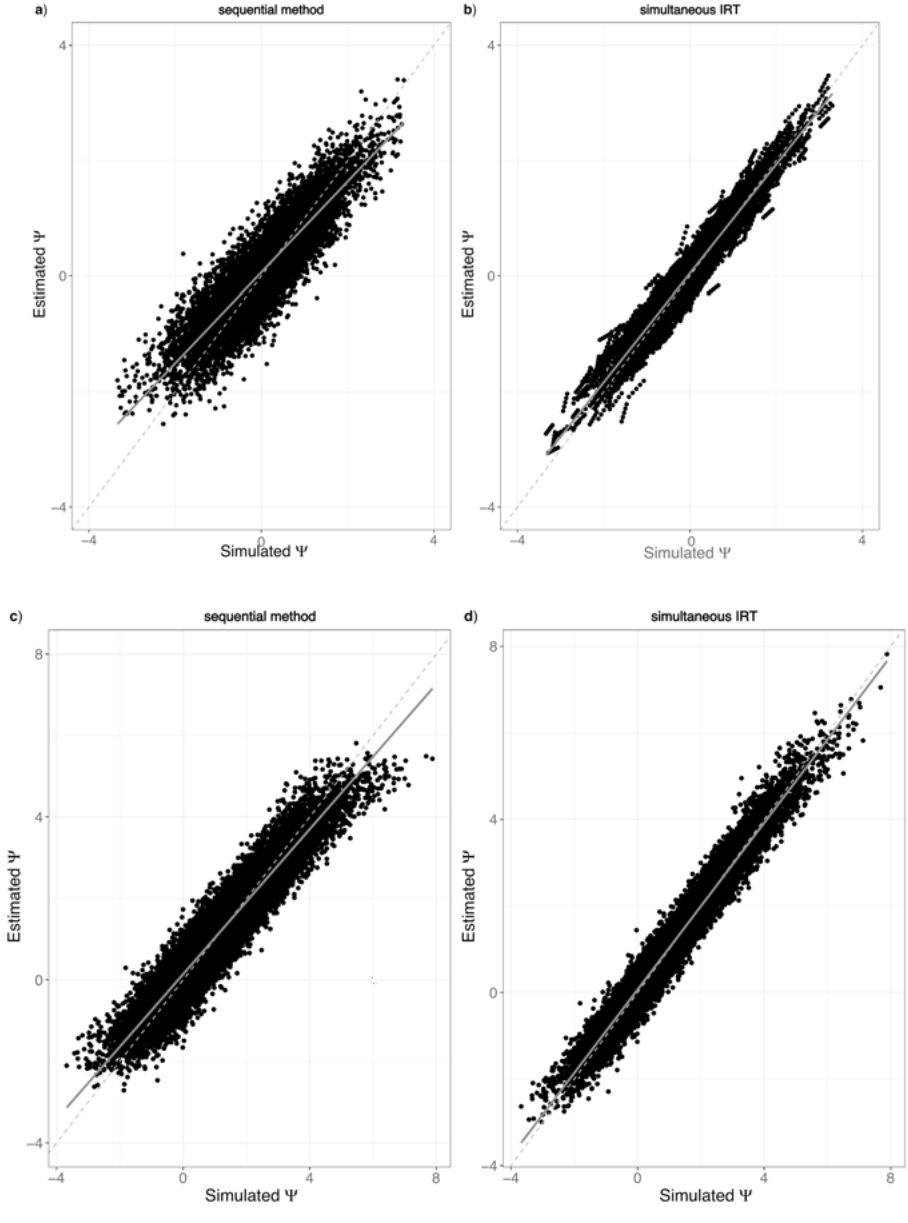

Figure 12 Representative trials (n=10) individual estimated latent variable (Ψ) vs simulated (true) latent variable (Ψ) correlation for the slow progression rate (a,b) and rapid progression rate (c,d) scenarios

Both methods performed similarly in terms of low estimation error in disease progression rate and drug effect (Figure 13) and type I error and power (table 5). Disease progression rate was estimated with minimal bias, with the sequential method indicating more bias under rapid progression rate, but still

negligible relative to the progression rate. Large bias was observed for drug effect in the slow progression rate, with a large range for SD indicating wide range of errors (both over or underpredicting).

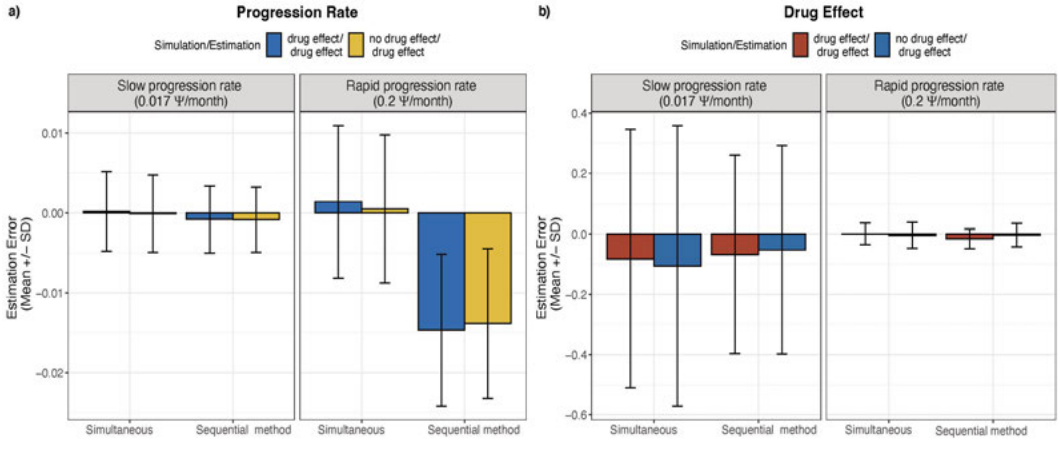

Figure 13 Mean Estimation error for (a) progression rate and (b) drug effect parameters for each progression rate scenario

Table 5 Simultaneous and sequential model power and type I error comparison

	Simultaneous		**Sequential**	
Progression Rate Scenarios	Power	Type I error rate (3.1%-6.9%*)	Power	Type I error rate (3.1%-6.9%*)
Slow progression rate (0.017 Y/month) N=200 (1:1)	7.4%	4.2%	5.8%	1.6%
Rapid progression rate (0.2 Y/month) N=500 (1:1)	99%	6.4%	99%	4.6%

*binomial distribution 95% confidence interval

Both methods achieved near 100% power in the rapid progression rate scenario with larger sample size. The type I error for each method was contained within the binomial confidence interval or below the significance level; therefore, indicating no evidence of type I error inflation for either method.

The simultaneous method was further utilized to assess IRT performance in shortened assessments (paper I). In reduced information and item scenarios, the longitudinal IRT model was able to successfully estimate parameters demonstrating relative stability in shortened assessments. Items were ranked by population information content from most informative to least informative.

The most informative item was Item 49 ("Global Spontaneity of Movement") and item 47 ("Postural Stability") was ranked the least informative in the early Parkinson's disease patient population from the PPMI study. More than 50% of the total information was contained in the top 10 most informative items.

Table 6 Item level Ranking of MDS-UPDRS Components by Information Content and Total Cumulative % Information Content for Items on Motor Subscale (34 Items)

			Cumulative % of Total Information Remaining	
Item	Test Name	Information at Baseline for Latent Variable	Removal From Most Informative Direction	Removal From Least Informative Direction
49	Global Spont. Of movement	0.58	100%	
35	Finger Tap-left hand	0.57		
37	Hand Move-left hand	0.50	81%	
18	Dressing	0.42		19%
39	Pronation-supine-left hand	0.42	71%	
31	Rigidity LUE	0.41		29%
28	Facial expression	0.37		
29	Rigidity Neck	0.33	59%	
36	Hand move-right hand	0.32		41%
34	Finger Tap-right hand	0.32	51%	
41	Toe tap-left foot	0.32		49%
24	Getting out of bed	0.30		
33	Rigidity LLE	0.29	41%	
43	Leg agility-Left leg	0.29		59%
42	Leg agility-Right leg	0.27		
40	Toe tap-right foot	0.24		
48	Posture	0.21	29%	
27	Speech 3.1	0.21		71%
32	Rigidity RLE	0.21		
17	Eating Tasks	0.20		
25	Walking and balancing	0.20	20%	
38	Pronation-supine-right hand	0.19		80%

30	Rigidity RUE	0.19		
21	Doing hobbies and other activities	0.19		
22	Turning in bed	0.16		
19	Hygiene	0.16	10%	
14	Speech	0.15		90%
45	Gait	0.14		
15	Saliva and Drooling	0.14		
44	Arising from chair	0.13		
20	Handwriting	0.10		
26	Freezing	0.09		
16	Chewing and swallowing	0.08		
47	Postural Stability	0.04		100%
Items on row with percentage are also included in the information content decrement step				

Fisher information curves for the 100% scenario for each item is represented in the publication. Item information curves illustrate how much the response from patients with a specific ability or disease severity level contribute to the estimation of the latent variable parameters-indicating the relative importance. The amplitude of the information curve demonstrates the item's overall contribution to the assessment, while the peaks and troughs represents its informativeness at a certain level of ability, providing deeper insight than the point estimate of population information alone. Some items provide more information for the center of the population (e.g., item 49 "Global Spontaneity of Movement"), while others are more informative at the extremes of the disability continuum. For example, item 26 "freezing", demonstrates higher information in patients with higher levels of disability (>2), which would imply this item is more informative for later stage disease.

In the simulated data setting, the impact of reducing the assessment was minimal resulting in item characteristics that were similar between the full and shortened assessments with the lowest level of information remaining. This is demonstrated by the alignment in the drop in fisher information efficiency (defined as reduced scenario information/100% scenario information) and the anticipated loss of information (Figure 14). In the real data setting, the trends in information content showed substantial differences between the full and reduced assessments. This behavior illustrated that the relative weighting of the data readjusts when items were removed from the assessment when using real-world data (Figure 14-16). When items were removed from the least informative direction first, differences in efficiency between the simulated and

real data scenarios occur early at 90% information remaining. The difference in this and the observation from the most informative direction is likely driven by the number of items removed as the amount of information removed is the same in both scenarios.

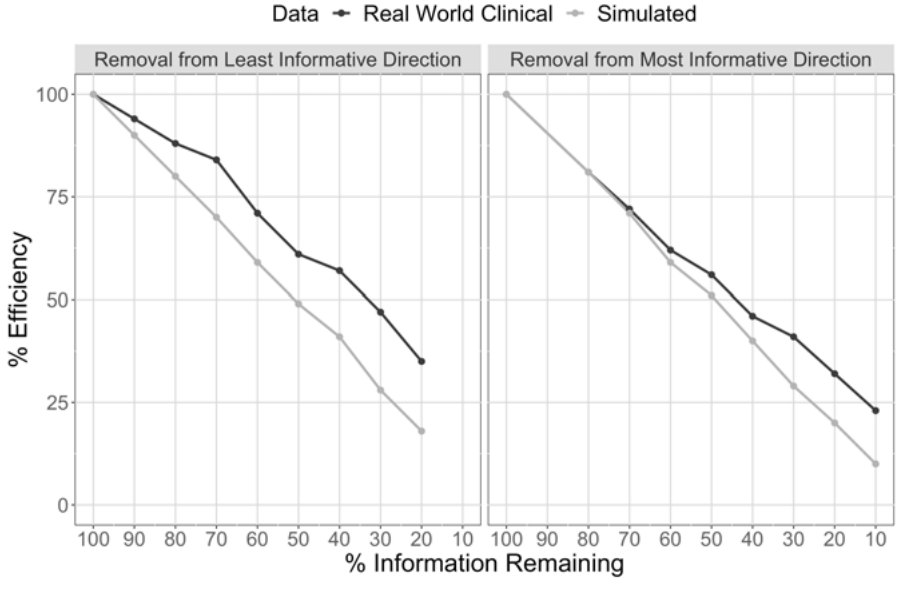

Figure 14 Efficiency on the population level for real and simulated data at each level of reduced information content scenarios

In real data setting, when the most informative items were removed first (Figure 16) there were larger differences in efficiency observed compared to the least informative direction (Figure 15). In addition, those items that were most informative for the more severe disease population seemed to be more affected than those that were more informative for the center of the population (the reference, -2 to 2); items that were more informative for the center of the population, such as “global spontaneity of movement” and “facial expression”, maintain near 100% efficiency, when removing items from the least informative direction (Figure 15).

When information was reduced, items that exhibited change often showed overall increased efficiency but this often came at the cost of loss of efficiency at tails of the disease severity spectrum. The increase in efficiency was more pronounced when less than 60% of the original information remained, as seen in “finger tap-left hand” item, showing increase in efficiency up to ~300%. Some items, such as item 37 “hand move-left hand” and item 14 “Speech” demonstrated noticeable changes in efficiency, with the former losing

efficiency in the higher and lower disease severity population as information decreased, when the least informative items were removed and the latter gaining information in the higher disease severity population, when the most informative items were removed.

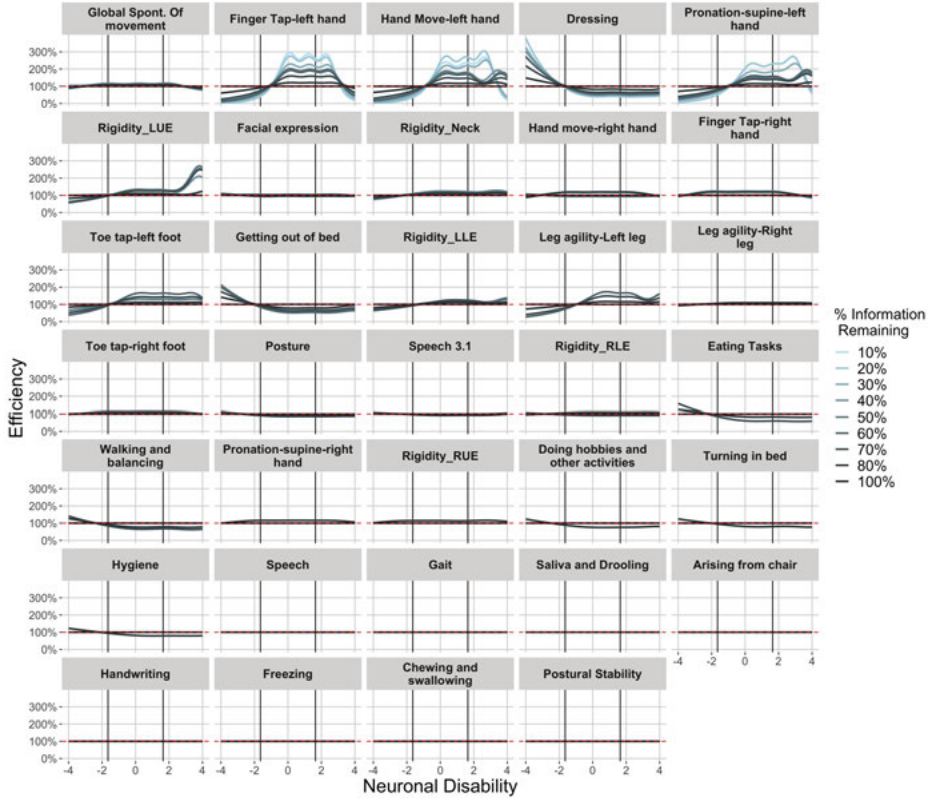

Figure 15 Real world clinical data item level efficiency for MDS-UPDRS motor items versus disability (Removal of Least informative Items First). Vertical Lines indicate the disability range for 95% of the reference population

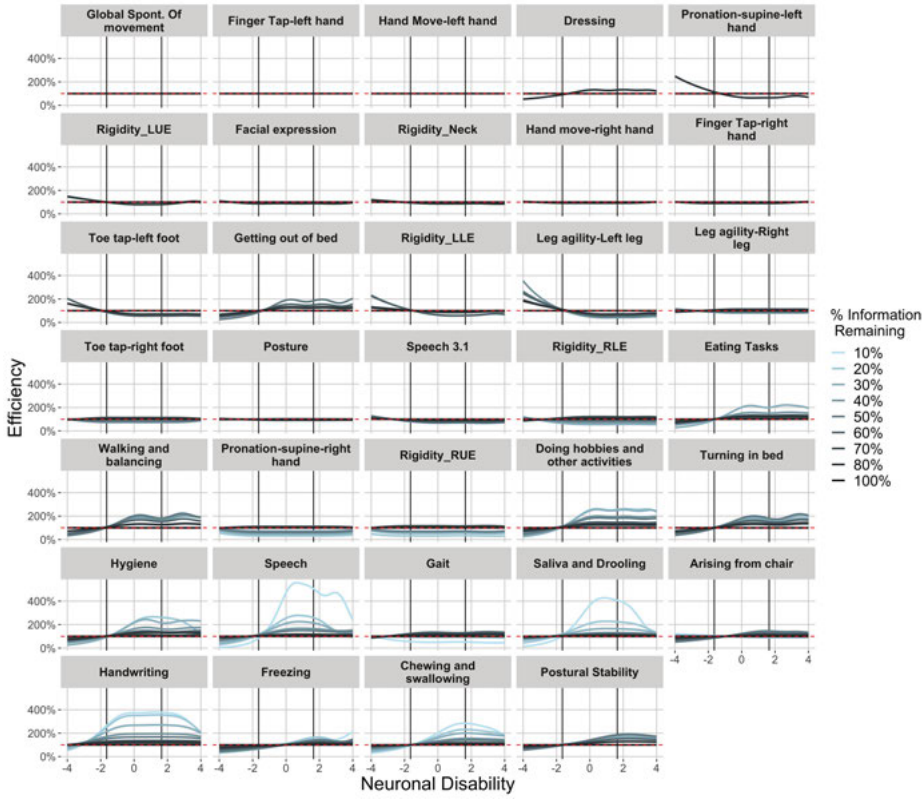

Figure 16 Real world clinical data item level efficiency for MDS-UPDRS motor items versus disability (Removal of Most informative Items First). Vertical Lines indicate the disability range for 95% of the reference population

Item removal from either direction led to a decrease in the estimated mean disease progression rate and symptomatic drug effect when the assessment was reduced from the original number of items (Figure 17). However, disease progression rate was still able to be accurately estimated up to ~40% of information remaining when items were removed from the most informative direction.

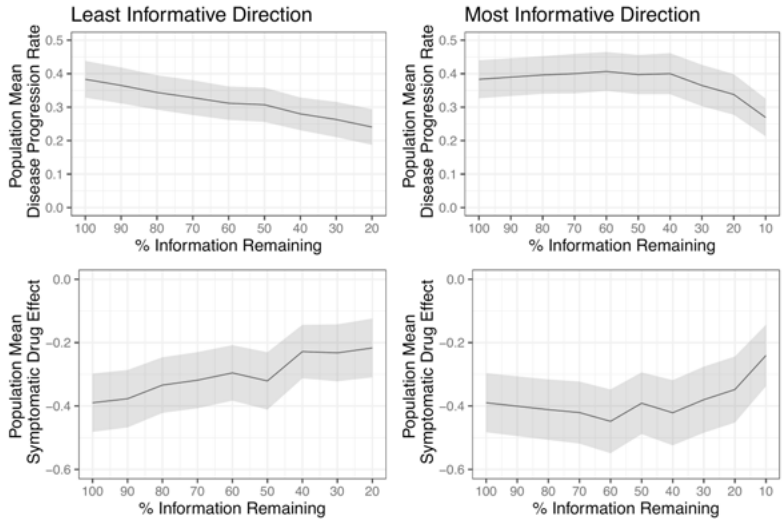

Figure 17 Estimated population mean disease progression rate (latent variable/year) and symptomatic drug effect estimates and 95% CI (shaded area) for observed motor subscale data at each level of reduced information content with least informative (left panel)

4.3 Clinical questions (Paper IV and Paper V)

The ADAS-Cog11 data from the verubecestat EPOCH trial was well described by the longitudinal IRT model and parameters were estimated with reasonable precision. (Paper V). The summary of parameter estimates and significant predictive covariates are shown in Table 7. VPC for composite score from the longitudinal IRT model and representative item level VPC for orientation are presented in the manuscript, demonstrating goodness of fit.

Table 7 Estimated Population Parameter values of the longitudinal-IRT model (logit scale)

Parameter	***Latent Variable-Disease severity (v)***			
	Fixed effect	RSE[a]	ω[b]	RSE[a]
Progression rate (v/day)	0.0014	3.5%	0.0225	5.90%
Symptomatic offset	-0.0801	21.8%		
Covariates on Progression rate				
Age	-0.0343	9.3%		
use of AChEI and memantine	0.4033	16.8%		
No use	-0.2035	53.6%		
use of memantine	0.1244	96.6%		

[a]Relative Standard Error is the SE/ parameter estimate *100; [b]Variance of Random Effect; Typical subject is median 73 years of age and use of AChEI

Option characteristic curves were used to visualize the item parameters, for each item (Figure 18). The individual item discrimination parameters presented in table 8 based on the EPOCH data indicate that immediate word recall, comprehension and naming objects and fingers are among the more discriminatory items, while word recognition and construction praxis appearing significantly less discriminate.

Table 8 Estimated item discrimination for ADAS-Cog11

Item	Discrimination (a)
Immediate Word Recall	1.80
Naming objects and fingers	1.47
Commands	1.42
Constructional praxis	0.878
Ideational praxis	1.36
Orientation	1.22
Word Recognition	0.824
Spoken Language Ability	1.41
Comprehension	1.46
Word Finding Difficulty in Speech	1.29
Remembering Test Instructions	1.35

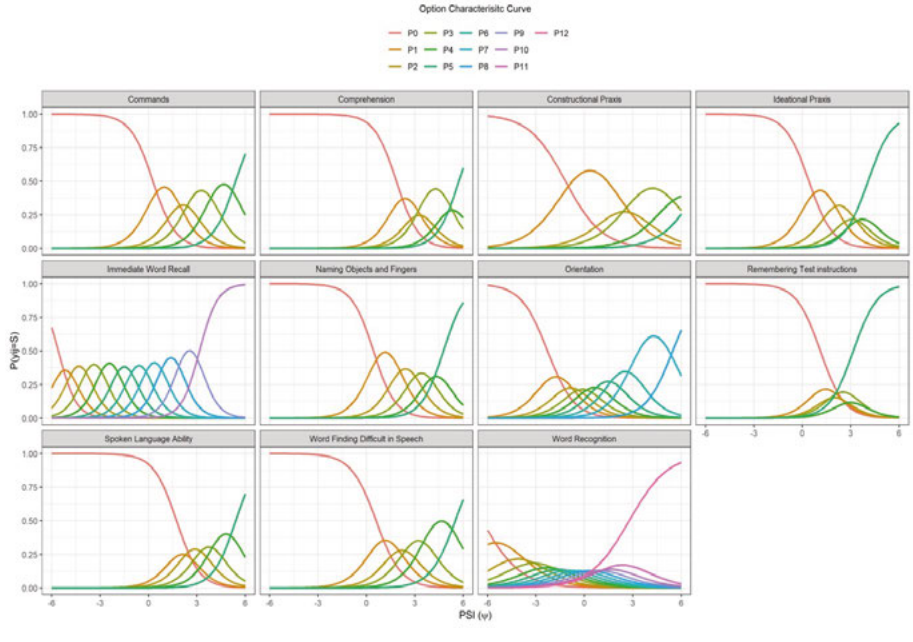

Figure 18 ADAS-Cog11 Estimated Option characteristic curves by item

The option characteristic curves suggest that for certain items, specifically remembering test instructions, the number of response categories may not be suitable for this patient population, as the categories do not progress monotonically along the disease severity continuum-showing substantial overlap for middle categories. Representative figure for the time course for the normalized item scores for each item from the simulated trials with a sample size of 50 participants per arm is shown in the manuscript.

Data from CTS (trials N=25,50,100 subjects per arm) was analyzed and hypothesis testing was performed using the following methodologies: i) longitudinal IRT model, ii) composite score disease progression modeling iii) and the most commonly used approach, pair-wise comparison for end of trial change from baseline. IRT method demonstrated higher power to detect a treatment effect associated with a 2-point change in ADAS-Cog11, compared not only to the traditional methodology using composite score at end of trial, but also the composite score disease progression model with 60% and 50% fewer subjects needed for IRT for at least 80% power.

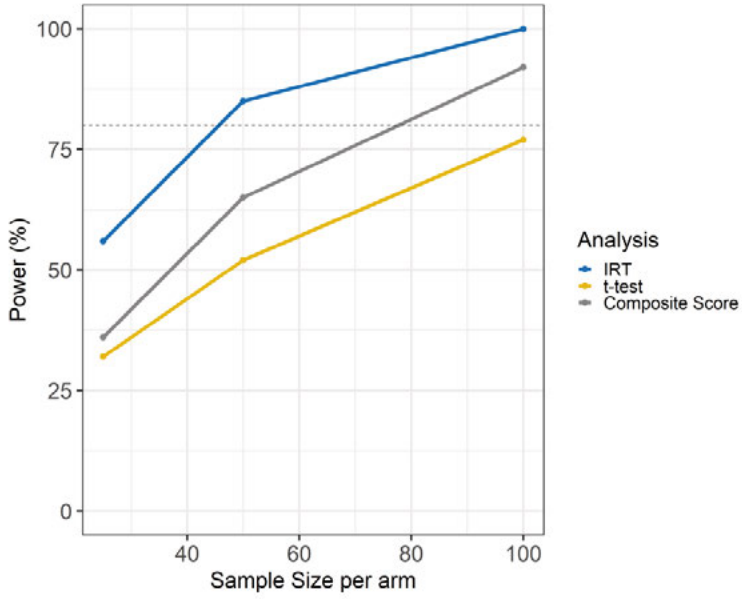

Figure 19 Power to detect a hypothetical symptomatic effect versus number of subjects per arm for IRT, t-test, and composite score disease progression analysis for ADAS-Cog11 score (dashed line is 80% power)

Expanding beyond the traditional applications of IRT in pharmacometrics, Paper IV assessed whether item response models and artificial neural network, utilizing item level data from the MDS-UPDRS could differentiate between PD patients and individuals with PD-like symptoms but without evidence of

dopaminergic deficit on a DaTscan (SWEDDs). The distribution of MDS-UPDRS total score for both patient populations is shown in figure 20.

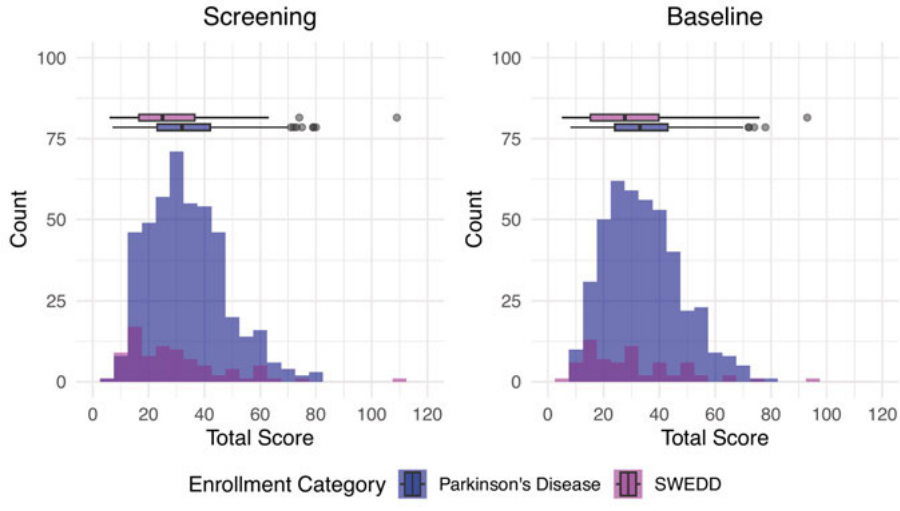

Figure 20 Distribution (Boxplot and histogram) of MDS-UPDRS total score for DeNoPD and SWEDDs. Boxplot represents inter-quartile range (IQR), median and whiskers to minimum, maximum value within 1.5 IQR. DeNoPD (Screening N=452, baseline N=430) and SWEDD (Screening)

For SWEDD subjects the earlier determined ICFs based on DeNoPD population were determined not sufficient to describe this non-PD population. Therefore, leveraging baseline and post-baseline data SWEDD-specific item parameters were estimated levering the simultaneous method, resulting in statistically significant reduction in OFV and determined to be the final model used for prediction. The development of the ANN involved hyperparameter tuning to optimize its performance and select the most appropriate model. Model performance was evaluated using performance classifiers as described in methods. MCC was the metric that was able to provide the most differentiation between models during the model development stage. This resulted in selection of an ANN with 1 hidden layer consisting of 60 nodes.

4.3.1 Classifier performance

The total score model was developed simply to provide context to the other two methods, and had low differentiation ability due to overlap in total score between the two populations. It classified all subjects as DeNoPD, regardless of the probability cut-off, showing no distinguishing power.

Table 9 Classifier method quality statistics (Mean and 95% CI) for screening data

	$P_{DeNoPD} \geq 50\%$ Cut-off			$P_{DeNoPD} \geq 90\%$ Cut-off		
	IRT	**ANN**	**Total Score**	**IRT**	**ANN**	**Total Score**
MCC	0.379 (0.259-0.498)	0.413 (0.279-0.535)	0.14 (0.104-0.221)	0.385 (0.283-0.483)	0.404 (0.301-0.498)	NR
Sensitivity	93% (91-96%)	96% (94-97.8%)	100% (NR)	82% (79-86%)	82% (78.9-85.9%)	0%
Specificity	43% (31-55%)	38% (26-50%)	1.5% (NR)	67% (54-78%)	68% (56.7-79.1%)	100% (NR)
Precision	91% (89-94%)	91% (88.1-93.4%)	88% (84-90%)	94% (92-96%)	94% (91.8-96.5%)	NR
Overall ROC AUC (95% CI)	82.70% (77.1-87.6%)	85.20% (80.5-89.3%)	65.90% (58.8-73.1%)			

AUC=area under the receiver operating characteristic curve; MCC= Matthews correlation coefficient; NR=not reportable

Both the IRT and ANN methods performed well as classifiers, when assuming a probability cut off of PDeNoPD≥ 50%. Specificity was ~40% for both IRT and ANN. MCC was evaluated as a metric to address potential bias of majority positive class. Although not shown in the table, using MCC as a performance metric, evaluating multiple cutoffs (50%-90%) revealed similar performance for the IRT and ANN models, with an MCC of ~0.4, while ANN consistently scored slightly above 0.4. An MCC of at least 0.3, demonstrates at least moderate positive relationship between predicted and true class. These results suggest that a cutoff of PDeNoPD ≥ 50% is necessary to identify patients more likely to have PD, while a threshold of PDeNoPD ≥ 90% allows for gain in specificity, without sacrificing much sensitivity. These proposed methodologies can supplement clinical examinations, helping neurologists distinguish between PD and SWEDD patients, and enabling more accurate identification of those who need a DaTscan, thereby reducing patient burden and diagnostic costs. The ROC curve in Figure 20 represents the sensitivity and specificity of each method for the classification of DeNoPD and SWEDD at varying cut-off values for the probability of DeNoPD. Training and testing the ANN for classification was much faster (minutes vs. days) than IRT item estimation, but both methods had similar, relatively fast classification prediction times.

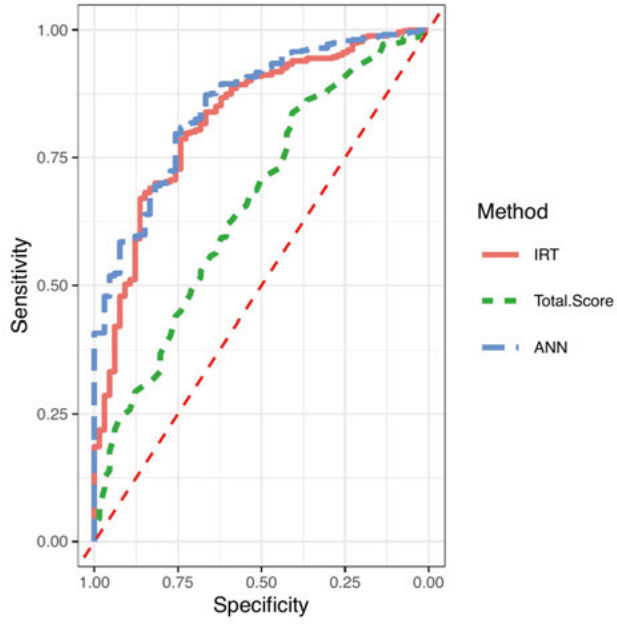

Figure 21 Receiver operating characteristic curves for ANN, IRT, and total score models

5 Discussion

This thesis evaluated the robustness of IRT models and their ability to quantitate drug effect under varying conditions, including different sample sizes, assessment lengths, and progression rates. In addition, to incorporate longitudinal data into the IRT models, two approaches for defining ICFs were explored, assessing their impact on the estimation properties of the item parameters as well as the latent variable. It also examined the impact of Laplace and GHQ-EM estimation algorithms on item parameter estimation. Lastly, we applied IRT methodology to clinical questions, comparing its performance to other methodologies not only to inform trial design but also explored its application for classification of diseases while evaluating its benefits, limitations and overall suitability for clinical decision-making.

As earlier described, COA-based endpoints can be highly variable, this highlights the need to develop more sensitive endpoints or analytical methods to leverage this data in a more optimized manner.

5.1.1 Performance in shortened or partial assessments (Papers I and II)

In reduced information and item scenarios, the IRT model was able to successfully estimate parameters demonstrating relative stability in shortened assessments. Item information was used as the metric for comparison, assessing how well each item measures along the symptomatic construct and was able to be calculated easily following estimation of the model parameters. While the item characteristics remained consistent when items were removed in the simulated data setting, the changes observed in the item characteristics in the real data setting indicate that there was a shift in the average measured construct when a unidimensional model was applied. Assuming a single dimension, in which all items are assumed to be measuring the dominant dimension, as items are removed secondary dimensions begin to receive more weight and the meaning of what is being measured changes. Using real-world data, we learned that while the model remains relatively stable, meaningful changes are occurring within the underlying construct. Specifically, removing items led to an increase in information for the remaining items. For some items, a loss of

information was observed at higher severity levels and as a result there was a reduction in the estimated mean disease progression rate and estimated drug effect. However, there was no meaningful difference in the mean population parameters until about 60% and 40% of information remaining when items were removed in the least informative and most informative direction respectively. This highlights the varying importance of different items across disease stages as well as the relevance or applicability of using a particular rating scale or subscales developed for one stage of disease in earlier or later stages.

It is important to note that the approach taken here used the ranking defined in the 100% scenario throughout each evaluation. This was done for consistency to make for a clearer comparison to the original assessment. However, reranking after each removal may also be an intuitive approach.

Due to the realignment of the average disease severity dimension caused by the removal of items, caution is warranted when comparing results utilizing different item subsets. An understanding of the benefits gained and information loss due to removal of items is required when interpreting results obtained from a shortened assessment compared to the full.

Another study focused on a comparing two estimation algorithms, Laplace and GHQ-EM as approaches for non-linear regression of item level data from rating-scale based COAs. The assessment also explored the impact of sample size and the number of items (i=5 or 20) on the ability of each method to estimate the item parameters with reasonable precision. It was found that at least 20 items and 100 subjects are needed for good accuracy and precision when performing a cross-sectional IRT analysis of a COA using both methods.

5.1.2 Estimation Strategies (Paper II and III)

Parameterization of IRT models involves a large number of parameters, and estimation using Laplace method (most common method) can be time-consuming, depending on the number of subjects, items and timepoints involved. Gaussian-Hermite quadrature uses a fixed grid of points across the data distribution which allows for faster estimation. Considering the beneficial features and limitations of each algorithm, this thesis compared the Laplace method implemented in the pharmacometrics software NONMEM and the GHQ-EM method executed the psychometric software *mirt*, for estimation of item parameters. The results show similar performance under varying conditions with increased bias and differences between methods observed when the number of subjects or items were reduced. This simulation study, with 1000 replicates, included outliers that represented only a small fraction of the data.

Defining the upper boundaries in NONMEM was of great importance, as one of the primary differences observed between methods was in the

estimation of discrimination versus threshold parameters, which the latter were influenced by upper bound parameter setting. Reducing the upper boundaries for the threshold parameters in NONMEM from 50 to 10 resulted in a more appropriate search space for the scale. In contrast, *mirt* used default bounds ranging from negative and positive infinity and demonstrated slightly better accuracy for the threshold parameters compared to NONMEM, indicating that *mirt* was less impacted by boundary setting.

Paper III offers insights for selection of the estimation strategy for determining the item characteristic functions and incorporating longitudinal data into the IRT framework by comparing two commonly used methods. In general, the results showed that the simultaneous and sequential methods performed similarly, low estimation error and reasonable precision for key parameters; type I error and power to detect drug effect. In general, the latent variable was well-determined across all scenarios. However, increased uncertainty was observed at the lower and upper ends of the latent variable scale for both methods, and greater uncertainty observed overall for the sequential method. The differences between methods were more pronounced in the rapid progression rate scenario, where larger estimation bias and reduced precision were observed for the sequential method compared to simultaneous. For the sequential method in the rapid progression rate scenario, the mean estimation error and 95% CI for disease progression rate across 500 trials, was different from zero. When individual progression rate estimates were assessed, the range of disease progression estimates for the simultaneous method seems to cover a slightly wider range of disease severity than the sequential method This may suggest that the sequential method may be more susceptible to shrinkage, which could be the cause of the slightly higher bias seen under this scenario.

The differences observed between the simultaneous and sequential method latent variable estimation were primarily due to the number of random effects. Due to treating each observation as a separate individual, the sequential method's random effects are informed by less data resulting in greater parameter uncertainty for the latent variable and increased estimation error in disease progression. To reduce some of this uncertainty, in the second step of the sequential method uncertainty of the latent variable (SE term) from the first step of the estimation was included, which resulted in higher precision and lower bias compared to without SE, as expected (Supplementary materials in publication). While the simultaneous method demonstrated slightly better precision, the sequential method offered additional advantages to the model building process. Advantages such as, enabling use of common pharmacometric diagnostics (e.g, residuals, IPRED/PRED vs observed) to allow for deeper

evaluation of parameters including the latent variable, mitigating bias in ICFs caused by longitudinal misspecification and accelerated development of the longitudinal model components.

5.1.3 Application to Clinical questions (Paper IV and V)

Pharmacometric IRT models have traditionally been utilized for characterizing disease progression. However, in this thesis we extended the application of IRT to explore its potential for disease diagnosis as well. Given the growing interest in machine learning techniques and their recent integration into clinical development, one study (Paper IV) expanded the evaluation to encompass not only disease progression modeling of total score and IRT modeling but also a comparative analysis with machine learning (ML). This provides a valuable benchmark for applying IRT methodology to classification tasks, extending its use beyond traditional psychometric analysis into pharmacometrics for clinical decision-making. Paper IV demonstrated IRT's usefulness in distinguishing between PD patients and SWEDDs using MDS-UPDRS item-level data, despite minimal differences observed in total score. Achieving this through model-based classification approaches without imaging data, but instead by deriving individual likelihoods of Parkinson's disease is a new approach.

One method employed total score, as this is the common way to analyze data from COAs, however this method failed to adequately differentiate between the DeNoPD and SWEDD populations, by all metrics, due to the similarity in total score distributions between the two populations. The total score model ultimately categorized all subjects as DeNoPD (i.e., majority positive class) demonstrating no ability to distinguish between cohorts regardless of cut-off (AUC 65%). The artificial neural network (ANN) was chosen as the ML method for comparison with IRT due to its ability to handle large, complex datasets, model non-linear relationships and perform classification tasks. An additional advantage of the ANN is its significantly faster computational time, with results obtained in minutes compared to the days required for large IRT models. Future work could explore the use of other machine learning techniques for classification tasks using item-level data. In fact, random forest was also evaluated for classification however, indicated weaker classifier performance than ANN and IRT under the final model (supplementary materials Paper IV), therefore was not further interrogated or discussed.

Both IRT and ANN methods demonstrated solid performance as classifiers with ROC AUC >80%, sensitivity and precision >90% at a cut off $P_{DeNoPD} \geq$ 50% (probability of PD). Increasing the probability cutoff enhances the specificity to ~70% for both methods, with only a slight reduction in sensitivity, which remains above 80% at a cut-off of $P_{DeNoPD} \geq$ 90%. These methodologies

prove valuable in assisting physicians to identify patients more likely to have PD with a cutoff of $P_{DeNoPD} \geq 50\%$ offering a balanced approach, while a higher threshold of $P_{DeNoPD} \geq 90\%$ can be applied for greater specificity. An important aspect explored in this study was the selection of appropriate quality statistics and its applicability to the data. Although ROC AUC is a widely used quality metric for binary classification, it may not always provide an accurate measure of performance, especially when there is class imbalance. The same can be said for precision and F1 scores. To account for this class imbalance, MCC was used as a more robust alternative. The MCC revealed a clearer distinction between total score method and other approaches, as the total score method had a MCC close to zero indicating weak classifier, and the other methods achieved MCC values greater than 0.3.

The application of IRT models using individual likelihood for classification represents a novel and promising approach and can be leveraged in clinical drug development to inform inclusion criteria and also in clinical practice for disease diagnosis. By leveraging item level data, both IRT and ANN methods effectively differentiated PD patients from SWEDD, offering the potential to avoid unnecessary, costly DaTscans or identify patients who may benefit from additional imaging. These proposed methodologies have the potential to reduce patient burden in clinical practice and offers opportunities to lower healthcare costs. Additionally, in Parkinson's disease or other heterogenous diseases where refining clinical trial inclusion criteria can increase the probability of success these methods could prove highly valuable and help reduce late-stage failures.

As mentioned in the introduction, using IRT to analyze data from rating scale-based COAs has proven valuable in detecting and quantifying treatment effects, with numerous examples in the literature showing that it provides greater power to detect drug effects compared to traditional approaches. However, these studies typically evaluated longer duration trials, large sample size and disease modifying treatment effects, with limited application in real-time drug development. One of the earliest noteable examples of pharmacometrics IRT in Alzheimer's was implemented using the Alzheimer's disease Neuroimaging initiative (ADNI) and Coalition against major diseases (CAMD) databases, much of which was derived from observational trials at the time of post-hoc analysis[43]. In early clinical development, proof of concept trials aim to identify treatment effects quickly with limited sample sizes, which can be further complicated in diseases with challenging enrollment and highly variable endpoints. Paper V assessed the power to detect drug effect in a short duration (12-week) proof of concept trial in a limited sample size using ADAS-Cog11

scale. The superior power demonstrated by the IRT model compared to traditional approaches is consistent with findings from simulations for longer trials.

In recent years, there has been a shift in drug development in neurological diseases, marked by increased data sharing across companies through consortiums. This collaborative approach enhances knowledge and understanding of the disease, enabling more informed decisions and efficient development of novel therapies. To build on this idea, its essential to emphasize that sharing successful application of analytical methods is just as important as sharing data. For example, finding more efficient ways to analyze commonly collected data and extend current methodologies like IRT to new applications and context, for both analysis and trial design can play a crucial role in advancing the field. As pharmacometrics-IRT becomes more established, its integration into clinical drug development plans has the potential to improve trial design and endpoint analysis, as demonstrated in this case study. This case study also highlights the importance of cross-functional collaboration between clinical development physicians, statisticians, and quantitative clinical pharmacologists in designing the trial protocol. Such collaboration is essential for integrating diverse expertise and ensuring a well-rounded and innovative approach to trial design and analysis.

Model-based tools that leverage longitudinal data, such as disease progression modeling, even at the composite score level, have proven valuable in detecting treatment effects. However, they are often underutilized, mainly appearing as post-hoc supplemental analyses rather than being formally incorporated as a trial endpoint even if in an exploratory manner. With limited public disclosure of such methods actually being applied to inform clinical drug development real time, it is difficult to challenge the status quo. Similarly, analyzing item-level data faces the same hurdles, with the composite score often regarded as the gold standard for many diseases. The limited use of item-level analysis or IRT methodology to support trial endpoints-even in an exploratory manner -may stem from lack of clear regulatory guidance on best practices for these innovative approaches. However, in 2022 and 2023 the FDA released a 4-part series of guidance to inform development, selection and use of COAs and COA-based endpoints. This provides a strong foundation for future. While the FDA doesn't typically endorse specific modeling approaches, it evaluates their use based on strength of supporting evidence on a case-by case basis. The lack of historical precedence for regulatory acceptance or inclusion of these methods (longitudinal, or item-level) in regulatory submission packages, may also contribute to their underutilization or, at the very least, limited disclosure due to the uncertainty regarding their perceived value and impact on drug approval.

6 Conclusion and Future Perspectives

In clinical drug development, item response theory has the potential to guide internal decision making by informing trial-design and enhancing the analysis of rating-scale based clinical endpoints. This thesis provides a methodological foundation for applying item response theory in pharmacometrics to facilitate MIDD, providing a thorough evaluation of its robustness, estimation techniques and applicability to modeling rating scale-based clinical outcome assessments. The primary applications of analysis focused on Parkinson's and Alzheimer's disease, but the methodological challenges and results are broadly applicable and can be extended to other disease indications and rating scales.

Key Findings and Considerations:

- Quantitative analysis using item-level data allows for extraction of more information from a rating scale-based COA than a composite or total score, leading to a more detailed and informative analysis when supported by a correctly specified and adequate model.
- There may be a desire to use a shorten form of a COA or use a partial assessment to evaluate drug effects, identify more sensitive items for characterization of disease in particular disease stages or overall reduce noise generated from less informative items to create more sensitive endpoints. When removing items there are several factors that should be considered, especially when comparing across trials, other analyses or to the gold standard. In addition to understanding the trade-off between information lost and gained by changes to the underlying construct, it is crucial to evaluate concordance with the full scale, understand scale dimensionality and patient population differences. Decisions to shorten an assessment should involve expert input from clinicians, particularly if the shortened version is to be used as an endpoint, in which case discussions with regulators are essential.
- Our findings indicate that for a robust IRT analysis with acceptable precision a sample size of at least 100 subjects and 20 items is recommended, if utilizing a single time point. If fewer items, a larger sample size or multiple time points may be necessary.

- Laplace implemented in NONMEM and GHQ-EM implemented in mirt estimate item parameters with comparable accuracy and precision. The GHQ-EM method implemented in the *mirt* package is user-friendly and fast, making it a suitable option to consider in the pharmacometric application of IRT either as a stand-alone tool for exploratory analyses or providing initial estimates to a longitudinal IRT model developed in NONMEM.
- The simultaneous and sequential methods for estimating ICFs perform well with reasonable precision. The simultaneous method provided a slight advantage in precision, while the sequential method expedites model development, minimizes ICF estimation risks from longitudinal model misspecification, and enables use of common pharmacometric diagnostics.
- A well-defined IRT model with robust data is able to classify Parkinson's and SWEDD patients with acceptable level of sensitivity and specificity to compliment clinical examination, with similar classification performance as the artificial neural network. Matthews correlation coefficient was determined to be a more robust metric for comparison of methods for data with class imbalance. This evaluation highlights an opportunity to apply these methods not only for Parkinson's and SWEDDS classification to reduce uncertainty of diagnosis and guide physician in selection of DaTscan candidates, but also for potential expansion to other disease indications, with proper testing and development.
- IRT models demonstrate higher power to detect drug effect compared to more traditional methods that use composite scores or single time point analysis. This is due to the enhanced utilization of all available score data capturing both the test characteristics and patient characteristics, and utilization of longitudinal data. Despite this, IRT and composite score disease progression models remain underutilized in real-time drug development, typically appearing as post-hoc analyses. Application and socialization of these more innovative approaches in real-time drug development either through publications, inclusion as an exploratory endpoint or inclusion as supportive evidence in regulatory submissions, would add to shared knowledge in the field and lay the foundation for increased application in practice and broader acceptance.

7 Acknowledgements

The work in this thesis was carried out in the Pharmacometrics Research Group at the *Department of Pharmacy* (formerly at the Department of Pharmaceutical Biosciences), Faculty of Pharmacy at Uppsala University, Sweden. I am grateful for the opportunity to engage with such a warm, kind, and supportive community of brilliant students, researchers and faculty.

My main supervisor Professor Mats O. Karlsson, you have made this the most rewarding experience. Your scientific brilliance and leadership fostered an excellent environment for learning. Despite me being based primarily in the United States, I always felt your support- thank you for your invaluable guidance, kindness, laughter and your somehow being everywhere all at once.

To my deputy-supervisor Sebastian Ueckert your vibrant energy and excitement during our discussions on technical topics were truly inspiring. Your passion motivated me. You were happy to dig into my many questions- I thank you for your patience, kindness and for your teaching spirit. I really appreciated the moments of humor and the engaging conversations we shared—it made working together such a pleasure.

The opponents for my half-time review: Dr. Emilie Henin, Prof. Maria Kjellsson, and Dr. Emilie Schindler; The questions I received from you helped guide my thesis and organize key questions.

Thank you to Lena, Siv, and Rikard for helping me navigate the university's processes and computing platforms, and sharing tidbits about Sweden. Andy H. and his wife Kristin for your kindness during the holiday.

To all the members of the pharmacometrics research group, past and present and visiting over my frequent and not so frequent visits- thank you for the camaraderie, lunch time laughs, fika (tea for me), cake, Uppsala adventures and after works. Viktor for helping me get acclimated to Uppsala, helping decode my ICA shopping visits when I buy what I think is yogurt and is not yogurt. Chenyan, my office mate during my first visit who helped me feel very welcome. Yassine, my UPSS and IRT buddy, enjoyed our intense scientific discussions and our many laughs. Xiaomei for your assistance during my travels. Thank you also to Gustaf, Joao, Eva, Estelle, Iris, Carolina, Lenaïg, Budi, Lina, Alan, Rami, Shijun, Diego, Stella, Jenny, Elodie, Elin, Henrik, Ari, Sreenath, Moustafa, Anders, Mohammed, Alzahra, Raphael, Amaury,

Frida,Yuanxi, Maddelena, Eman, and Rory (& Cody), for sharing your insights both scientific and otherwise you all contributed greatly to my experience. Apologies to anyone I may have missed.

I would also like to thank my former Merck colleagues: Malidi Ahamadi for encouraging me, challenging me, preparing me, and opening doors to even begin this journey which was pivotal to my future. Wei Gao, thank you for your unwavering advocacy, encouragement, and leveraging my potential-early. Vikram S. for your support, guidance and planting the seed. Thank you also to my co-authors Raj M. and Nitin M., as well as Kelly Y and Bela P for your strong support.

Last but not least my family and friends. To my parents, grandma and my brother for your unwavering belief in me, your constant support, and your encouragement of my dreams—no matter how grand or humble. I am forever grateful. To all my friends, your thoughtful check-ins, the airport runs, encouragement and simply being my champions—my deepest gratitude for your support.

Leticia

8 References

1. Sheiner, L.B., *Learning versus confirming in clinical drug development.* Clin Pharmacol Ther, 1997. **61**(3): p. 275-91.
2. Sertkaya, A., et al., *Costs of Drug Development and Research and Development Intensity in the US, 2000-2018.* JAMA Netw Open, 2024. **7**(6): p. e2415445.
3. Hwang, T.J., et al., *Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results.* JAMA Intern Med, 2016. **176**(12): p. 1826-1833.
4. EFPIA MID3 Workgroup, S.M., R Burghaus, V Cosson, SYA Cheung et al, *Good Practices in Model-Informed Drug Discovery and Development: Practice, Application, and Documentation.* CPT Pharmacometrics Syst Pharmacol, 2015. **5**(3): p. 93-122.
5. *Model-Informed Drug Development Paired Meeting Program*. [cited 2019; Available from: https://www.fda.gov/drugs/development-resources/model-informed-drug-development-paired-meeting-program.
6. Lalonde, R.L., et al., *Model-based drug development.* Clin Pharmacol Ther, 2007. **82**(1): p. 21-32.
7. Workgroup, E.M., et al., *Good Practices in Model-Informed Drug Discovery and Development: Practice, Application, and Documentation.* CPT Pharmacometrics Syst Pharmacol, 2016. **5**(3): p. 93-122.
8. *Pharmacometrics*. Available from: https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/division-pharmacometrics.
9. Bauer, R.J., *NONMEM Tutorial Part I: Description of Commands and Options, With Simple Examples of Population Analysis.* CPT Pharmacometrics Syst Pharmacol, 2019. **8**(8): p. 525-537.
10. *Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments*. 2022, FDA Maryland.
11. *Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making*. 2023, FDA Maryland.
12. *CLINICAL OUTCOME ASSESSMENT (COA) COMPENDIUM*. 2021.
13. Walton, M.K., et al., *Clinical Outcome Assessments: Conceptual Foundation-Report of the ISPOR Clinical Outcomes Assessment - Emerging Good Practices for Outcomes Research Task Force.* Value Health, 2015. **18**(6): p. 741-52.
14. Kalia, L.V. and A.E. Lang, *Parkinson's disease.* Lancet, 2015. **386**(9996): p. 896-912.
15. Goetz, C.G., et al., *Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results.* Mov Disord, 2008. **23**(15): p. 2129-70.

16. Holden, S.K., et al., *Progression of MDS-UPDRS Scores Over Five Years in De Novo Parkinson Disease from the Parkinson's Progression Markers Initiative Cohort.* Mov Disord Clin Pract, 2018. **5**(1): p. 47-53.
17. Pagano, G., et al., *Trial of Prasinezumab in Early-Stage Parkinson's Disease.* N Engl J Med, 2022. **387**(5): p. 421-432.
18. Vu, T.C., J.G. Nutt, and N.H. Holford, *Disease progress and response to treatment as predictors of survival, disability, cognitive impairment and depression in Parkinson's disease.* Br J Clin Pharmacol, 2012. **74**(2): p. 284-95.
19. Breijyeh, Z. and R. Karaman, *Comprehensive Review on Alzheimer's Disease: Causes and Treatment.* Molecules, 2020. **25**(24).
20. Guarino, A., et al., *Executive Functions in Alzheimer Disease: A Systematic Review.* Front Aging Neurosci, 2018. **10**: p. 437.
21. De-Paula VJ, R.M., Diniz BS, Forlenza OV, *Alzheimer's disease*, in *Sub-cell Biochem.* . 2012. p. 329-352.
22. *LEQEMBI (lecanemab-irmb) injection, for intravenous use [package insert].* . 2023, Esai Inc.: Nutley, NJ.
23. Kueper, J.K., M. Speechley, and M. Montero-Odasso, *The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review.* J Alzheimers Dis, 2018. **63**(2): p. 423-444.
24. Hambleton RK, S.H., *Item response theory Principles and applications*. 1985, Boston, MA: Kluwer-Nijhoff Publishing.
25. Lord, F.M. and M.R. Novick, *Statistical theories of mental test scores*. 1968, Charlotte NC: Information Age Publishing.
26. Rasch, G., *Probabilistic models for some intelligence and achievement tests. Copenhagen: Danish Institute for Educational Research*. 1960, expanded edition 1983, Chicago, IL: Mesa Press.
27. Cai, L., et al., *Item Response Theory.* Annual review of statistics and its application, 2016. **3**.
28. Nguyen, T.H., et al., *An introduction to item response theory for patient-reported outcome measurement.* Patient, 2014. **7**(1): p. 23-35.
29. Balsis, S., et al., *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): p. 288-94.
30. Ueckert, S., *Modeling Composite Assessment Data Using Item Response Theory.* CPT Pharmacometrics Syst Pharmacol, 2018. **7**(4): p. 205-218.
31. Reckase, M.D., *Multidimensional Item response Theory* Statistics for Social and Behavioral Sciences. 2009, New York, NY: Springer.
32. Hambleton, R.K., H. Swaminathan, and J. Rogers, *Fundamentals of Item Response Theory*. 1991, Newbury park, CA: Sage
33. Samejima, F., *Handbook of Modern item response theory*. 1997, New York, NY: Springer.
34. Yang, F.M. and S.T. Kao, *Item response theory for measurement validity.* Shanghai Arch Psychiatry, 2014. **26**(3): p. 171-7.
35. Zhang, B., *Application of unidimensional models to tests with items sensitive to secondary dimensions.* Journal of experimental education, 2008. **77**(2).
36. Rao, C.R., *Information and the Accuracy Attainable*
in the Estimation of Statistical Parameters Bulletin of Calcutta Mathematical Society, 1945. **37**: p. 81-91.
37. Cramér, H., *Methods of Estimation: Mathematical Methods of Statistics*. 1946, Princeton, NJ: Princeton University Press.

38. Cappelleri, J.C., J. Jason Lundy, and R.D. Hays, *Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures.* Clin Ther, 2014. **36**(5): p. 648-62.
39. Gottipati, G., et al., *Item Response Model Adaptation for Analyzing Data from Different Versions of Parkinson's Disease Rating Scales.* Pharm Res, 2019. **36**(9): p. 135.
40. Buatois, S., et al., *Item Response Theory as an Efficient Tool to Describe a Heterogeneous Clinical Rating Scale in De Novo Idiopathic Parkinson's Disease Patients.* Pharm Res, 2017. **34**(10): p. 2109-2118.
41. Gottipati, G., M.O. Karlsson, and E.L. Plan, *Modeling a Composite Score in Parkinson's Disease Using Item Response Theory.* AAPS J, 2017. **19**(3): p. 837-845.
42. Krekels, E., et al., *Item Response Theory to Quantify Longitudinal Placebo and Paliperidone Effects on PANSS Scores in Schizophrenia.* CPT Pharmacometrics Syst Pharmacol, 2017. **6**(8): p. 543-551.
43. Ueckert, S., et al., *Improved utilization of ADAS-cog assessment data through item response theory based pharmacometric modeling.* Pharm Res, 2014. **31**(8): p. 2152-65.
44. Schindler, E., et al., *A Pharmacometric Analysis of Patient-Reported Outcomes in Breast Cancer Patients Through Item Response Theory.* Pharm Res, 2018. **35**(6): p. 122.
45. Germovsek, E., et al., *A Novel Method for Analysing Frequent Observations from Questionnaires in Order to Model Patient-Reported Outcomes: Application to EXACT(R) Daily Diary Data from COPD Patients.* AAPS J, 2019. **21**(4): p. 60.
46. Llanos-Paez, C., et al., *Improved Decision-Making Confidence Using Item-Based Pharmacometric Model: Illustration with a Phase II Placebo-Controlled Trial.* AAPS J, 2021. **23**(4): p. 79.
47. Han, S.H., et al., *Artificial Neural Network: Understanding the Basic Concepts without Mathematics.* Dement Neurocogn Disord, 2018. **17**(3): p. 83-89.
48. Heaton, J. *The Number of Hidden Layers*. 2017 [cited 2025; Available from: https://www.heatonresearch.com/2017/06/01/hidden-layers.html
49. Hinton, G.E., S. Osindero, and Y.-W. Teh, *A Fast learning algorithm for deep belief nets.* Neural Computation, 2006. **18**(7): p. 1527-1554.
50. Morris, T.P., I.R. White, and M.J. Crowther, *Using simulation studies to evaluate statistical methods.* Stat Med, 2019. **38**(11): p. 2074-2102.
51. Bonate, P., *Clinical Trial Simulation in Drug Development.* Pharm Res, 2000. **3**: p. 252-256.
52. Kirby, A., V. Gebski, and A.C. Keech, *Determining the sample size in a clinical trial.* Med J Aust, 2002. **177**(5): p. 256-7.
53. Chow, S.-C., J. Shao, and H. Wang, *Sample Size Calculations in Clinical Research*. 2 ed. Chapman & Hall/ CRC Biostatistics Series. 2008: Taylor & Francis Group, LLC.
54. Maerek, K. *Parkinson's Progression Markers Initiative: The PPMI 001 2014 study protocol* 2014 [cited 2024; Available from: https://www.ppmi-info.org/sites/default/files/docs/archives/Amendment-9.pdf.
55. Egan, M.F., et al., *Randomized Trial of Verubecestat for Mild-to-Moderate Alzheimer's Disease.* N Engl J Med, 2018. **378**(18): p. 1691-1703.
56. Bauer, R.J., *NONMEM Tutorial Part II: Estimation Methods and Advanced Examples.* CPT Pharmacometrics Syst Pharmacol, 2019. **8**(8): p. 538-556.

57. Beal, S.L. and L.B. Sheiner, *NONMEM Users Guide Part VII Conditional Estimation Methods*. 1992-1998, Globomax, LLC: Hanover, MD.
58. Bock, R.D. and M. Aitkin, *Marginal Maximum LIkelihood Estimation of Item Parameters: Application of an EM Algorithm.* Psychometrika, 1981. **46**(4): p. 443-459.
59. Pinheiro, J.C. and D.M. Bates, *Approximations to the log-likelihood function in the nonlinear mixed-effects model.* J. Comput. Graph. Stat. , 1995. **4**: p. 12-35.
60. Mould, D.R., *Models for disease progression: new approaches and uses.* Clin Pharmacol Ther, 2012. **92**(1): p. 125-31.
61. Samejima, F., *Homogenous case of continuous response model.* Psychometrika, 1973. **38**: p. 203-219.
62. Arrington, L., et al., *An R package for Automated Generation of Item Response Theory Model NONMEM Control File*, in *Population Approach Group 28*. 2019: Stockholm, Sweden.
63. Conrado, D.J., et al., *An updated Alzheimer's disease progression model: incorporating non-linearity, beta regression, and a third-level random effect in NONMEM.* J Pharmacokinet Pharmacodyn, 2014. **41**(6): p. 581-98.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 375

Editor: The Dean of the Faculty of Pharmacy

A doctoral dissertation from the Faculty of Pharmacy, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-552893

ACTA UNIVERSITATIS
UPSALIENSIS
2025