Pharmaceutical Research (2022) 39:1779–1787
https://doi.org/10.1007/s11095-022-03194-1

RESEARCH PAPER

Improved Confidence in a Confirmatory Stage by Application of Item-Based Pharmacometrics Model: Illustration with a Phase III Active Comparator-Controlled Trial in COPD Patients

Carolina Llanos-Paez[1] · Claire Ambery[2] · Shuying Yang[2] · Misba Beerahee[2] · Elodie L. Plan[1] · Mats O. Karlsson[1]

Received: 23 September 2021 / Accepted: 9 February 2022 / Published online: 1 March 2022

Abstract

Purpose The current study aimed to illustrate how a non-linear mixed effect (NLME) model-based analysis may improve confidence in a Phase III trial through more precise estimates of the drug effect.

Methods The FULFIL clinical trial was a Phase III study that compared 24 weeks of once daily inhaled triple therapy with twice daily inhaled dual therapy in patients with chronic obstructive pulmonary disease (COPD). Patient reported outcome data, obtained by using The Evaluating Respiratory Symptoms in COPD (E-RS:COPD) questionnaire, from the FULFIL study were analyzed using an NLME item-based response theory model (IRT). The change from baseline (CFB) in E-RS:COPD total score over 4-week intervals for each treatment arm was obtained using the IRT and compared with published results obtained with a mixed model repeated measures (MMRM) analysis.

Results The IRT included a graded response model characterizing item parameters and a Weibull function combined with an offset function to describe the COPD symptoms-time course in patients receiving either triple therapy (n = 907) or dual therapy (n = 894). The IRT improved precision of the estimated drug effect compared to MMRM, resulting in a sample size of at least 3.64 times larger for the MMRM analysis to achieve the IRT precision in the CFB estimate.

Conclusion This study shows the advantage of IRT over MMRM with a direct comparison of the same primary endpoint for the two analyses using the same observed clinical trial data, resulting in an increased confidence in Phase III.

Key Words chronic obstructive pulmonary disease · item response theory · mixed model repeated measures · non-linear mixed effect model · patient-reported outcomes

Introduction

The primary reason for Phase III clinical drug development failure is insufficient drug efficacy (55%), followed by safety (14%) and strategic reasons (14%) (1). Although the development of "ineffective" or "unsafe" compounds should stop during early stages of drug development, the proportion of failure owing to insufficient efficacy is currently larger in Phase III compared to Phase II (55% vs. 48%) (1). The causes of such failures may include not only insufficient drug efficacy but also insufficient knowledge of the treatment effect in the target population (2) leading to an insufficient sample size (e.g. sample size calculated from a commonly overestimated true treatment effect in Phase II (3)), and therefore high uncertainty in the efficacy estimate. A model-informed drug development (MIDD) decision-making framework has been proposed to increase Phase III trials probability of success through more precise estimates of the drug effect. This may result in an increased proportion of successful Phase III and IV trials (4).

Chronic Obstructive Pulmonary disease (COPD) is an inflammatory disease of the respiratory system, accounting for 54.9% of chronic respiratory diseases in 2017 (5). It was the third leading cause of death in 2016 and is projected to remain among the five leading causes of death by 2030 (6). The Global Initiative for Chronic Obstructive Lung Disease (GOLD) strategy (7) states that commonly used maintenance medications in COPD are short or long acting β2 agonists

✉ Mats O. Karlsson
mats.karlsson@farmaci.uu.se

[1] Department of Pharmacy, Uppsala University, BMC, Box 580, 751 23 Uppsala, Sweden

[2] Clinical Pharmacology Modelling and Simulation, GlaxoSmithKline plc, London, UK

(SABA and LABA, respectively), short or long acting anticholinergic (SAMA and LAMA, respectively) or combination therapies with/without addition of inhaled corticosteroids (ICS). The GOLD strategy recommends that patients with advanced COPD and persistent symptoms who are at risk of exacerbations use an inhaled triple therapy (ICS/LABA/LAMA) (7). The Lung Function and Quality of life Assessment in COPD with Closed Triple Therapy (FULFIL) is the first Phase III study that compared a once-daily single-inhaler triple therapy with a twice-daily inhaled dual therapy (8). Patient-Reported Outcomes (PROs) from the FULFIL trial (8, 9) were obtained from different tools, such as the Evaluating Respiratory Symptoms in COPD (E-RS:COPD) (10), to compare the effect of fluticasone furoate/umeclidinium/vilanterol (FF/UMEC/VI) and budesonide/formoterol (BUD/FOR) on patient's respiratory symptoms. PROs are an important aspect of drug development as they contribute directly with information on the patient's own perceptions of their disease. The E-RS:COPD consists of 11 items from the 14-item Exacerbations of COPD Tool (EXACT), intended to capture information specifically related to respiratory symptoms. This tool has been derived through the application of item and Rasch model analysis (11).

The results of the statistical analysis of E-RS:COPD total score (RS-Total) from the FULFIL trial (8, 9) showed that FF/UMEC/VI significantly improved patient respiratory symptoms (assessed via E-RS:COPD tool) in comparison to BUD/FOR (9). This statistical analysis was performed using a mixed model repeated measures (MMRM) approach, a standard statistical method used in longitudinal clinical trials (12). This type of analysis may result in a loss of information and therefore a loss of precision in the efficacy estimate since MMRM analyzes the total score only, ignoring the contribution of each item in the questionnaire to the disease state. Moreover, in MMRM, time is handled as a discrete value rather than a continuous variable resulting in a further loss of information. To include all information from the data and thus increase power and precision to detect a drug effect, a MIDD approach can be applied to analyze PRO data such as RS-Total using a longitudinal non-linear mixed effect (NLME) analysis based on item-level data (i.e. item response theory based model - IRT) (13). In contrast to MMRM, IRT utilizes all components of the composite observations by relating its items to an underlying disease state that varies among individuals and changes with time. IRT thus utilizes all information captured by the questionnaire instead of only the total score as MMRM does.

In the planning of Phase III, different metrics are of interest for the design and analysis of trials such as the probability of study success or the probability of making a correct decision (14). An MIDD framework can provide a more informative approach to explore these metrics and therefore to improve late-stage clinical development productivity (4). This analysis aims to illustrate how a new methodology to analyze PRO data using an item-based pharmacometrics model would improve confidence in a confirmatory stage by comparing the precision around the primary efficacy endpoint obtained with IRT and MMRM.

Materials and Methods

Data and Patients

The FULFIL clinical trial (NCT02345161) was a Phase III, randomized, double-blind, double dummy, parallel-group, multicenter study that compared 24 weeks of once daily FF/UMEC/VI (100 µg/62.5 µg/25 µg) inhalation powder with twice daily BUD/FOR (400 µg/12 µg) in patients with COPD. Patients were randomized in a 1:1 ratio to FF/UMEC/VI or BUD/FOR. The study inclusion and exclusion criteria were reported elsewhere (8). In this analysis, nine patients were excluded from the intention to treat population for the following reasons: absence of E-RS:COPD score data for the whole study period (4 patients), dispensing errors (4 patients), and missing recorded time (1 patient). Daily data records were obtained by the completion of the E-RS:COPD tool using an electronic diary, which did not allow patients to skip individual items, although missing days (where patients did not provide answer for any of the items) were possible (10). The E-RS:COPD tool contains 11 items (a subset of the EXACT) that capture information related to respiratory symptoms of COPD (breathlessness, cough, sputum production, chest congestion and chest tightness), with seven items including five ordered categorical options and four items with four categories (Table I).

IRT Development

The IRT was developed in two steps: 1) characterization of the item characteristic functions (ICFs) and 2) the development of the longitudinal model (a flow chart of the analysis is shown Supplementary Fig. 1).

Item Characteristics Functions

The properties of each item were determined by non-linear functions (the ICFs) linking the unobserved patient's disease status, the latent variable θ, to the probability of giving a particular response for an item. Item parameters such as discrimination and difficulty parameters were estimated using an "independent occasion" approach (15). Under this approach, observations from each measurement occasion were treated as belonging to separate individuals, assuming a normal distribution with fixed mean and variance $N(0,1)$

Table I Content of the E-RS:COPD Tool

Item number	Item-level Construct	Score	Symptom Construct
7	Breathless today	0–4	Breathlessness
8	Breathless with activity	0–3	
9	Short of breath – personal care	0–4	
10	Short of breath – indoor activities	0–3	
11	Short of breath – outdoor activities	0–3	
2	Cough frequency	0–4	Cough and Sputum
3	Mucus quantity	0–3	
4	Difficulty with mucus	0–4	
1	Congestion	0–4	Chest Symptoms
5	Discomfort	0–4	
6	Tightness	0–4	

RS-Total is based on summation to yield ordinal-level scales with a range of 0–40

of θ at baseline, and estimated mean and variance $N(\mu,\omega^2)$ at later observations.

A logistic transformation was used to model each item with the probability of rating a score of at least k [$P(y_{ij} \geq k)$] and the probability of rating exactly score k [$P(y_{ij} = k)$] as shown in Eq.1 and Eq. 2, respectively. These probabilities are functions of θ_i, the latent variable of patient i, and the estimated fixed effect item parameters such as discrimination (a_j) and difficulty ($b_{j,k}$) parameters for item j; more specifically $b_{j,k}$ is the difficulty parameter for the item-score k, and y_{ij} corresponds to the observed data (E-RS:COPD scores) for an individual i and an item j.

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

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

Longitudinal Model

In this step, ICF parameters were fixed to the values obtained in step 1, and a data reconciliation was needed to include each individual's time-course data (original IDs). This was done since in step 1 each time point was treated as belonging to separate individuals. For the longitudinal model, the estimates of θ_i from the IRT, using ICF fixed parameters from step 1 were considered as the dependent variable (15) taking into account their uncertainty. These estimates of θ_i were obtained from an intermediate step where a large uncertainty for θ was considered to ensure that the total scores translated from the estimates of θ_i described the raw data well (as shown in Supplementary Fig. 2). The NONMEM control stream for this intermediate step to obtain θ_i for step 2 is available in the Supplementary material. A pre-specified Weibull function (Eq. 3) was used to describe changes in individual symptoms-time course (θ_i) for both treatment arms. Parameters such as disease progression time (T_{progi}), maximum response (R_{maxi}) and baseline latent variable ($\theta_{i,t=0}$) are subject-specific parameters with inter-individual variability (IIV), including a random variable with a mean of 0 and variance of ω^2. T_{progi} was assumed to be log-normally distributed with IIV modeled using an exponential function, whereas all other parameters were assumed to be normally distributed with an additive IIV model. Time t is in years and γ is the gamma value that governs the steepness of the curve. Different parameters per treatment arm were considered. Additionally, effects of smoking status and geographical regions on $\theta_{i,t=0}$ were included in the model as these covariates were considered in the MMRM analysis (9).

$$\theta_i = \theta_{i,t=0} + R_{maxi} \cdot \left(1 - e\left(-\left(\frac{\ln(2)}{T_{progi}} \cdot t\right)^{\gamma}\right)\right) \quad (3)$$

Internal Model Evaluation

Non-parametric ICF smooth plots were developed to assess ICF fit (16), and the predictive performance of the model was assessed by using Visual Predictive Check plots (VPCs). The 2.5th, 50th and 97.5th percentile of the observed data were compared to the 95%CI for the 2.5th, 50th and 97.5th percentiles of the simulated (N = 500) data. At this stage, further changes to the Weibull model were considered if it was deemed necessary.

Clinical Endpoint

Clinical Endpoint Definition

The clinical endpoint was the change from baseline (CFB) in RS-Total over 4-week intervals for each treatment arm (FF/UMEC/VI and BUD/FOR). To provide context for the mean differences between groups and being able to compare

changes in PRO scores between the treatment arms, a clinically meaningful CFB was defined as a change equal to or greater than a minimal clinically important difference (MCID) of 2 units (9). This means that a decrease of at least 2 points from baseline in RS-Total was deemed clinically significant.

Precision in Clinical Endpoint Estimate

The point estimate and the precision in the estimated clinical endpoint was obtained through model simulations, with inclusion of the uncertainty in the estimated longitudinal IRT parameters. RS-Total, linked to the individual patient disease status (θi), were simulated using the final IRT parameter estimates. The derived relationship between θ and RS-Total (Fig. 1) was used as a basis in the simulations. These stochastic simulations included parameter uncertainty from the estimated asymptotic variance-covariance matrix of the estimates by using the $PRIOR functionality in NONMEM. Specifically, the NWPRI subroutine was used where prior of fixed and random effects were assumed to be normally and inverse Wishart distributed, respectively. Degrees of freedom for the inverse Wishart distribution were calculated based on standard error (SE) of the obtained parameter estimates (17). RS-Total were simulated (N = 2000) over a period of 24 weeks for each treatment arm, using a large number of virtual subjects (Nsubj = 15,000 per arm). The individual average RS-Total at 4-week intervals was calculated and subtracted from the individual RS-Total at baseline. The distribution of CFB in RS-Total for each simulation (including 15,000 subjects) are illustrated in Fig. 1. The median, 2.5th, and 97.5th percentiles of the average CFB from each distribution were used to represent mean (95%CI) CFB in RS-Total (Fig. 1). These IRT derived values were compared with those (published values) obtained using the MMRM analysis (9).

Sample Size

The relative sample size (N) of a study analyzed using MMRM versus IRT can be calculated based on the precision (CI) of estimates for MMRM (95%CI length - CIMMRM) and IRT (95%CI length - CIIRT) for the same sample size as shown in Eq. 4. An N larger than one indicates that a MMRM analysis requires a larger study size to achieve the same precision as an IRT analysis.

$$N = \left(\frac{CI_{MMRM}}{CI_{IRT}} \right)^2 \tag{4}$$

Software and Estimation Method

The software NONMEM (ICON Development Solutions, Ellicott City, Maryland) version 7.4.4 (18) was used for modeling (using the first-order conditional estimation method (step 2) plus Laplacian (step 1)) and simulation together with an Intel FORTRAN compiler and Perl-speaks-NONMEM (PsN, http://psn.sourceforge.net) version 5.1.0 (19). R software (The R Foundation for Statistical Computing) version 3.5.2 (20) and R packages, such as Xpose4 (http: //xpose.sourceforge.net, version 4.6.1) (21, 22) and Piraid (version 0.4) (23) were used for data management, graphical analysis, to produce summary statistics, and to examine the table outputs from NONMEM.

Results

Data and Patients

Data from 1801 patients (mean [standard deviation] age of 63.9 years [8.65], 43.8% smokers at study initiation) who received 24 weeks of either FF/UMEC/VI (n = 907) or BUD/

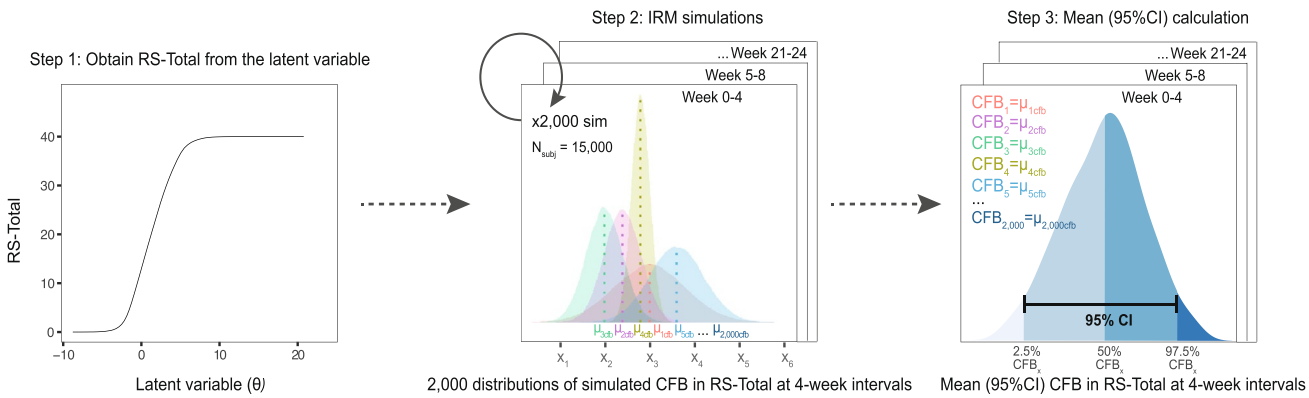

Fig. 1 Schematic representation of the workflow for the simulations including parameter uncertainty. Distribution of CFB in RS-Total for each simulation considering 15,000 subjects and the mean (95%CI) CBF in RS-Total obtained from the average CFB of each simulation

FOR (n = 894) were included in this analysis. Baseline characteristics are shown in Table II. One-thousand six-hundred forty-three (91%) patients provided data up to at least week 24 with a median (range) missing days of 5 (0–101), whereas 158 patients (9%) stopped filling out the questionnaire after 84 (1–160) days with 3 (0–86) missing days.

IRT and Simulations Including Parameter Uncertainty

ICF parameters were estimated with good precision (Supplementary Table I). Item characteristic curves that illustrate the relationship between disease status and probability of giving a certain score for all items are shown in Supplementary Fig. 3. Addition of an offset drug effect value and correlation between θi,t = 0, Rmax and offset improved the description of data. The offset drug effect (Eq.5) value, that is different between treatment arms, was assumed to be normally distributed with *IIV modeled using an additive function.*

$$\theta_i = \begin{cases} \theta_{i,t=0} & t = 0 \\ \theta_{i,t=0} + R_{maxi} \cdot \left(1 - e\left(-\left(\frac{\ln(2)}{T_{progi}} \cdot t\right)^{\gamma}\right)\right) + offset_i & t > 0 \end{cases} \quad (5)$$

Different parameters per arm were estimated with a median (range) relative standard error (RSE) of 0.13 (0.02–0.75) (Table III). This model also showed an acceptable fit to the total score data for both treatment arms. Specifically, the model predicts satisfactorily the median observed total score data, as seen with agreement between

Table III Parameter Estimates for the Longitudinal Model

Parameter estimates	FF/UMEC/VI Value (RSE)	BUD/FOR Value (RSE)
$\theta_{t=0}$ (unitless)	0.33 (0.23)	0.29 (0.28)
T_{prog} (year)	0.08 (0.03)	0.08 (0.03)
R_{max} (unitless)	−0.31 (0.11)	−0.16 (0.25)
γ (unitless)	9.27 (0.13)	16.9 (0.75)
Offset (unitless)	−0.27 (0.08)	−0.08 (0.30)
ω^2 $\theta_{t=0}$	1.09 (0.03)	1.47 (0.03)
ω^2 T_{prog}	0.45 (0.03)	0.46 (0.03)
ω^2 R_{max}	0.89 (0.04)	0.98 (0.06)
ω^2 Offset	0.37 (0.05)	0.42 (0.05)
ω^2 R_{max} ~ ω^2 Offset	11% (0.24)	–
ω^2 $\theta_{t=0}$ ~ ω^2 Offset	−12% (0.20)	−10% (0.27)
ω^2 R_{max} ~ ω^2 $\theta_{t=0}$	–	−13% (0.17)
Residual unexplained variability	0.32 (0.02)	
Smoking effect on $\theta_{t=0}$ (additive)	0.13 (0.62)	
Region 1 on $\theta_{t=0}$ (additive)	−0.64 (0.13)	
Region 3 on $\theta_{t=0}$ (additive)	−0.61 (0.13)	
Region 4 on $\theta_{t=0}$ (additive)	−0.20 (0.40)	
Region 5 on $\theta_{t=0}$ (additive)	−0.41 (0.27)	
Region 6 on $\theta_{t=0}$ additive)	−0.98 (0.12)	

Region 1 (21%): Germany, Greece, Italy; Region 2 (24%): Russian Federation, Ukraine; Region 3 (21%): Bulgaria, Hungary, Romania, Slovakia; Region 4 (18%): Czech Republic, Estonia, Poland; Region 5 (6%): China, Republic of Korea; Region 6 (10%): Mexico; $\theta_{t=0}$: baseline latent variable; T_{prog}: disease progression time; R_{max}: maximum response; RSE: relative standard error (for omega and sigma RSEs are reported on the approximate standard deviation scale); ω^2: variance describing inter-individual variability

Table II Patient Characteristics at Baseline

Baseline characteristics	FF/UMEC/VI (n = 907)	BUD/FOR (n = 894)
Age (years)	64.2 (8.56)	63.6 (8.73)
Time with COPD (years)	< 1 y: 28 (3%)	< 1 y: 39 (4%)
	1–5 y: 325 (36%)	1–5 y: 334 (37%)
	5–10 y: 300 (33%)	5–10 y: 281 (31%)
	10–15 y: 165 (18%)	10–15 y: 147 (16%)
	≥ 15 y: 90 (10%)	≥ 15 y: 93 (10%)
FVC (L)	2.84 (0.80)	2.87 (0.79)
FEV_1 (L)	1.25 (0.46)	1.24 (0.45)
Male (n)	675 (74%)	658 (74%)
Smoker (n)	396 (44%)	392 (44%)
COPD GOLD disease status	Moderate: 298 (33%)	Mild: 1 (0.1%)
	Severe: 501 (55%)	Moderate: 290 (32%)
	Very severe: 107 (12%)	Severe: 477 (53%)
		Very severe: 124 (14%)
RS-Total[b]	12.2 (5.85)	12.9 (5.96)

In this analysis, nine patients were excluded from the intention to treat population of the FULFIL clinical trial (NCT02345161) for the following reasons: absence of E-RS:COPD score data for the whole study period (4 patients), dispensing errors (4 patients), and missing recorded time (1 patient); [b] RS-Total was calculated as the mean value during baseline period defined as from day −14 to day −1

the observed and simulated data in the VPC (Supplementary Fig. 4). Goodness of fit plots are shown in Supplementary Fig. 5, and longitudinal model parameter estimates are shown in Table III. While typical T_{prog} was similar between the two arms, more negative R_{max} (−0.31 vs. -0.16) and offset (−0.27 vs. -0.08) values in the FF/UMEC/VI arm, indicated higher benefit of the FF/UMEC/VI treatment to the patient compared to the BUD/FOR arm. Model predicted and raw data for CFB in RS-Total at 4-week intervals are shown in Fig. 2. According to the IRT, in the BUD/FOR arm, the CFB in RS-Total score did not achieve the MCID at any time point, whereas in the FF/UMEC/VI arm, the mean CFB in RS-Total achieved the MCID from week 9 onwards, which is in agreement with the observed data (Fig. 2). NONMEM control stream and snippet of data for both models ICFs and longitudinal are provided in the Supplementary material.

Based on simulations including parameter uncertainty, the IRT considerably improved the precision of the drug effect compared to the MMRM at every time point (Fig. 3 and Table IV). At the end of treatment (week 21–24), the mean (95%) CFB in RS-Total was −2.47 (−2.61, −2.30) with IRT compared to −2.31 (−2.62, −2.00) with MMRM in the FF/UMEC/VI arm, and the mean (95%) CFB in RS-Total was −0.97 (−1.10, −0.81) with IRT compared to −0.96 (−1.27, −0.65) with MMRM in the BUD/FOR arm. Furthermore, a relative sample size (N; obtained using Eq. 4) of 4.00 (FF/UMEC/VI) and 4.72 (BUD/FOR) times larger would be required in the MMRM analysis to achieve the precision obtained with the IRT analysis at the end of the study (week 21–24). Sample size requirements at each week interval are shown in Fig. 3.

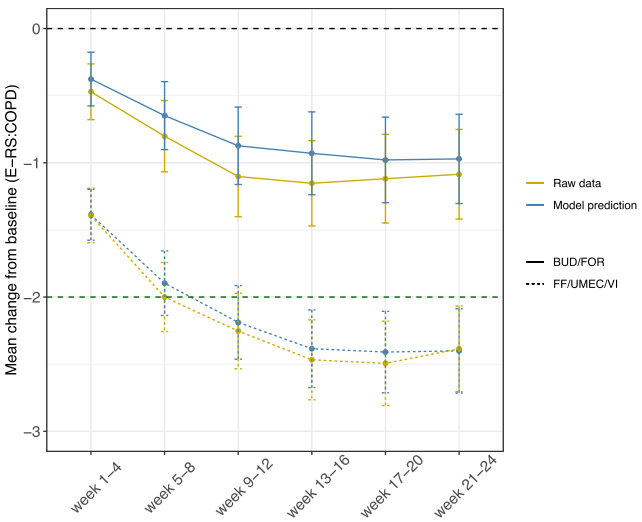

Fig. 2 Mean (95%CI) CFB in RS-Total at 4-week intervals for observed data (yellow) and predicted values from the IRT (blue). Green dashed line indicates the MCID target

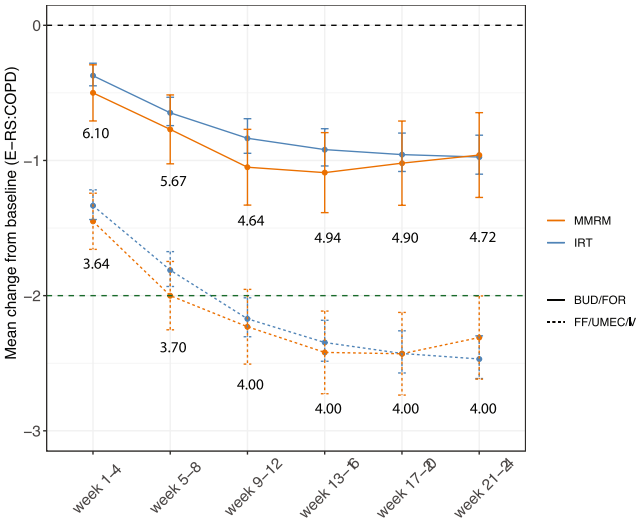

Fig. 3 Mean (95%CI) for CFB in RS-Total at 4-week intervals simulated with IRT (blue) and published MMRM values (red). Green dashed line indicates the MCID target. Values correspond to the relative sample size (Eq. 4)

Discussion

Using data from a Phase III study, a longitudinal IRT improved the precision in the efficacy endpoint compared to MMRM. The use of NLME analysis based on item-level data (IRT) was recently proposed as an alternative to MMRM for evaluation of efficacy in the analysis of data from a Phase II study, where IRT improved the precision of the estimated drug effect considerably in comparison to MMRM (24). The benefits of using an IRT over other statistical methods such as least-square mean analysis have already been demonstrated, highlighting its higher power to detect a drug effect (13, 25). This could be explained by the fact that IRT is a more informative approach that not only describe the longitudinal aspects of the data but also accounts for information from the scores given to the individual items of the questionnaire. In contrast, MMRM ignores this information in the data since i) it is an analysis of total score data only and ii) each visit is modeled independently as a factor with earlier time points contributing less information about the end-of-treatment response than what is the case for NLME model. MMRM analysis is designed for robustness rather than efficiency (12), hence it could be expected that IRT analysis is more efficient than MMRM. However, this study provides an insight into the magnitude of improvement in precision that an IRT model may provide and the consequences this can have on the required sample size.

The primary objective of the Phase III study analysis is the confirmation of efficacy (and safety) profiles of an investigational new drug. Overall, an increased precision around the efficacy estimate was observed with IRT compared to

Table IV Mean (95%CI) Change from Baseline (CFB) in RS-Total for IRM and MMRM at 4-week Intervals

	FF/UMEC/VI		BUD/FOR	
Week interval	CFB in RS-Total. Mean (95% CI) [SE]		CFB in RS-Total. Mean (95% CI) [SE]	
	MMRM	IRM	MMRM	IRM
1–4	−1.45 (−1.66, −1.24) [0.11]	−1.33 (−1.44, −1.22) [0.06]	−0.50 (−0.71, −0.29) [0.11]	−0.37 (−0.45, −0.28) [0.04]
5–8	−2.00 (−2.25, −1.75) [0.13]	−1.81 (−1.93, −1.67) [0.07]	−0.77 (−1.02, −0.52) [0.13]	−0.65 (−0.74, −0.53) [0.05]
9–12	−2.23 (−2.51, −1.95) [0.14]	−2.17 (−2.30, −2.02) [0.07]	−1.05 (−1.33, −0.77) [0.14]	−0.84 (−0.95, −0.69) [0.07]
13–16	−2.42 (−2.73, −2.11) [0.15]	−2.35 (−2.49, −2.18) [0.08]	−1.09 (−1.39, −0.79) [0.15]	−0.92 (−1.04, −0.77) [0.07]
17–20	−2.43 (−2.74, −2.12) [0.16]	−2.43 (−2.57, −2.26) [0.08]	−1.02 (−1.33, −0.71) [0.16]	−0.96 (−1.08, −0.80) [0.07]
21–24	−2.31 (−2.62, −2.00) [0.16]	−2.47 (−2.61, −2.30) [0.08]	−0.96 (−1.27, −0.64) [0.16]	−0.97 (−1.10, −0.81) [0.07]

MMRM. This analysis highlights that a smaller sample size would have been required to confirm the drug effect if a NLME model-based approach were to have been used instead of the MMRM approach. A median relative sample size over the 4-week intervals of 4.0 (FF/UMEC/VI) and 4.9-fold (BUD/FOR) larger would have been required for the MMRM analysis to achieve the IRT precision. These values are comparable with *a 3.5-fold smaller study size with IRT compared to MMRM analysis using RS-Total data from a Phase II trial* (24). This is particularly important in cases where a large number of patients cannot be recruited for pivotal trials (e.g. rare diseases) with presumably a clearer benefit of using the IRT approach due to the increased utilization of the information contained in the item-level data. Thus, the increased precision in the efficacy estimate with IRT can lead to a higher probability of making a correct decision which appears important in light of the many failures in Phase III attributed to underpowered trials.

This analysis illustrates how simulations using an IRT with parameter uncertainty included can be used to obtain the precision of the primary efficacy endpoint. A large number of virtual subjects (15,000) were used in the simulations to make sure that uncertainty comes mainly from the model parameter estimates rather than the sample size. The SEs associated with each parameter estimates were obtained from the variance-covariance matrix in NONMEM. To further investigate how SEs obtained from different techniques such as bootstrap or sampling importance resampling would have impacted the precision in the efficacy endpoint was not in the scope of this study; however, it would be of interest to investigate how this would affect the calculated sample size.

This analysis is not exempt of limitations, which are described as follow. A predetermined (decided before start analyzing the data) longitudinal model (Weibull function) was considered; however, the addition of an extra parameter (offset) was required to describe the rapid onset of bronchodilator effects in both treatment arms, and thus improve the predictive performance of the model (based on the VPC plot). The absence of this parameter in the model would have led to a poorer description of the data, which is not necessarily related to a worse or better precision and/or relative sample size (26). The authors acknowledge that, in order to use a pharmacometric model as primary analysis in confirmatory trials, this model need to be pre-specified to avoid the risk of type 1 error inflation due to multiple testing during model building. One benefit of MMRM in this aspect is its flexibility and the fact that can adequately be pre-specified, though some assumptions are still required (12). Furthermore, in this analysis, model uncertainty (e.g. uncertainty from the structural part of model or from the random effects) and its impact on the precision around the efficacy endpoint was also not investigated. To mitigate model uncertainty in a NLME model-based analysis, it has been proposed to conduct model averaging (27, 28). Model averaging would cover the model space by assigning a goodness-of-fit derived weight to the different proposed structural models, thereby including model uncertainty for a pre-specified analysis. Lastly, it can be argued that a simpler NLME model for the total score as one continuous endpoint could achieve similar results, in terms of precision and power, than an IRT-based model analysis when assessing treatment effect on the CFB of total scores. A formal comparison between these two types of analyses was not performed in this study; however, previous work has suggested that IRT-based models are more informative and require sample size that are approximately 20–40% smaller than analysis of total score data (13, 29–31). Despite these limitations, this analysis shows the advantage of NLME analysis with a direct comparison

of the same primary endpoint for the two methods (IRT and MMRM) using observed clinical trial data rather than focusing on drug effect parameters which are often unobserved. Based on simulations, it has been shown already that a NLME analysis can be more powerful than MMRM in some (albeit not all) scenarios (32).

Conclusions

The positive impact of using a NLME model-based approach in decision-making during drug development has already been shown (4, 33). This analysis shows the advantage of using a NLME model based on item level data over a standard approach used today in drug development (MMRM) for the same endpoint, increasing the precision in the efficacy estimate and thereby significantly reducing the required sample size to confirm drug effect.

ACKNOWLEDGMENTS AND DISCLOSURES.

The authors would like to acknowledge Maggie Tabberer for her contributions to the data analysis and interpretation. Trademarks are owned by or licensed to EVIDERA. CL-P, ELP and MOK declare that they have no conflict of interest. CA, SY and MB are GSK employees and hold GSK shares.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11095-022-03194-1.

Funding Statement GSK funded this research in the form of a Research payment to Uppsala University. Funding for this research was also provided by the Swedish Research Council Grant 2018-03317.

Author Contributions CL-P, ELP and MOK designed and carried out the analysis. SY, CA and MB were responsible for providing the data and review the analysis. CL-P drafted the manuscript, which were critically revised and received final approval by all authors.

Funding Open access funding provided by Uppsala University.

References

1. Harrison RK. Phase II and phase III failures: 2013–2015. Nat Rev Drug Discov. 2016;15:817–8. https://doi.org/10.1038/nrd.2016.184.
2. Lalonde RL, Kowalski KG, Hutmacher MM, Ewy W, Nichols DJ, Milligan PA, et al. Model-based drug development. Clin Pharmacol Ther. 2007;82:21–32.
3. Erdmann S, Kirchner M, Götte H, Kieser M. Optimal designs for phase II/III drug development programs including methods for discounting of phase II results. BMC Med Res Methodol. 2020;20:253.
4. Milligan PA, Brown MJ, Marchant B, Martin SW, van der Graaf PH, Benson N, et al. Model-based drug development: a rational approach to efficiently accelerate drug development. Clin Pharmacol Ther. 2013;93:502–14.
5. Xie M, Liu X, Cao X, Guo M, Li X. Trends in prevalence and incidence of chronic respiratory diseases from 1990 to 2017. Respir Res. 2020;21:49. https://doi.org/10.1186/s12931-020-1291-8.
6. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 2006;3:e442.
7. Global strategy for the diagnosis, management and prevention of COPD, global initiative for chronic obstructive lung disease (GOLD)]. 2020. https://goldcopd.org/wp-content/uploads/2019/12/GOLD-2020-FINAL-ver1.2-03Dec19_WMV.pdf. Accesed 24 Nov 2020.
8. Lipson DA, Barnacle H, Birk R, Brealey N, Locantore N, Lomas DA, et al. FULFIL trial: once-daily triple therapy for patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2017;196:438–46.
9. Tabberer M, Lomas DA, Birk R, Brealey N, Zhu C-Q, Pascoe S, et al. Once-daily triple therapy in patients with COPD: patient-reported symptoms and quality of life. Adv Ther. 2018;35:56–71.
10. Evidera. EXACT Program 2020. https://www.exactproinitiative.com/content/. Accessed 21 Apr 2021.
11. Jones PW, Chen W-H, Wilcox TK, Sethi S, Leidy NK. Characterizing and quantifying the symptomatic features of COPD exacerbations. Chest. 2011;139:1388–94.
12. Mallinckrod CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. Drug Inf J. 2008;42:303–19. https://doi.org/10.1177/009286150804200402.
13. Ueckert S, Plan EL, Ito K, Karlsson MO, Corrigan B, Hooker AC, et al. Improved utilization of ADAS-cog assessment data through item response theory based Pharmacometric modeling. Pharm Res. 2014;31:2152–65. https://doi.org/10.1007/s11095-014-1315-5.
14. Chuang-Stein C, Kirby S. Quantitative decisions in drug development. 1st ed: Springer International Publishing; 2017.
15. Schindler E, Friberg LE, Lum BL, Wang B, Quartino A, Li C, et al. A Pharmacometric analysis of patient-reported outcomes in breast Cancer patients through item response theory. Pharm Res. 2018;35:122.
16. Ueckert S. Modeling composite assessment data using item response theory. CPT Pharmacometrics Syst Pharmacol. 2018;7:205–18 https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/psp4.12280.
17. Chan Kwong AH-XP, Calvier EAM, Fabre D, Gattacceca F, Khier S. Prior information for population pharmacokinetic and pharmacokinetic/pharmacodynamic analysis: overview and guidance with a focus on the NONMEM PRIOR subroutine. J Pharmacokinet Pharmacodyn. 2020;47:431–46. https://doi.org/10.1007/s10928-020-09695-z.

18. Beal S, Sheiner L, Boeckmann A, Bauer RJ. NONMEM Users Guide (1989–2009). Icon Development Solutions, Ellicot City, Maryland, USA; 1989-2009.
19. Lindbom L, Ribbing J, Jonsson EN. Perl-speaks-NONMEM (PsN)--a Perl module for NONMEM related programming. Comput Methods Prog Biomed. 2004;75:85–94.
20. R core team. R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. 2020. https://www.r-project.org/index.html
21. Jonsson EN, Karlsson MO. Xpose--an S-PLUS based population pharmacokinetic/pharmacodynamic model building aid for NONMEM. Comput Methods Prog Biomed. 1999;58:51–64.
22. Keizer RJ, Karlsson MO, Hooker A. Modeling and simulation workbench for NONMEM: tutorial on Pirana, PsN, and Xpose. CPT Pharmacometrics Syst Pharmacol. 2013;2:e50.
23. Arrington L, Nordgren R, Ahamadi M, Ueckert S, Sreeraj M, Karlsson MO. An R package for Automated Generation of Item Response Theory Model NONMEM Control File. In: PAGE 28. 2019. http://www.page-meeting.org/?abstract=8869
24. Llanos-Paez C, Ambery C, Yang S, Tabberer M, Beerahee M, Plan EL, et al. Improved decision-making confidence using item-based Pharmacometric model: illustration with a phase II placebo-controlled trial. AAPS J. 2021;23:79. https://doi.org/10.1208/s12248-021-00600-1.
25. Ueckert S, Hooker AC, Karlsson MO, Plan EL. Item Response Theory Model as Support for Decision-Making: Simulation Example for Inclusion Criteria in Alzheimer's Trial. In: PAGE 23. 2014. http://www.page-meeting.org/?abstract=3267
26. Buatois S, Ueckert S, Frey N, Retout S, Mentré F. cLRT-mod: an efficient methodology for pharmacometric model-based analysis of longitudinal phase II dose finding studies under model uncertainty. Stat Med. 2021;40:2435–51.
27. Aoki Y, Röshammar D, Hamrén B, Hooker AC. Model selection and averaging of nonlinear mixed-effect models for robust phase III dose selection. J Pharmacokinet Pharmacodyn. 2017;44:581–97. https://doi.org/10.1007/s10928-017-9550-0.
28. Buatois S, Ueckert S, Frey N, Retout S, Mentré F. Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. AAPS J. 2018;20:56. https://doi.org/10.1208/s12248-018-0205-x.
29. Schindler E, Friberg LE, Karlsson MO. Comparison of item response theory and classical test theory for power/sample size calculation for questionnaire data with various degrees of variability in items' discrimination parameters. In: PAGE 24. 2015. http://www.page-meeting.org/?abstract=3468
30. Buatois S, Retout S, Frey N, Ueckert S. Item response theory as an efficient tool to describe a heterogeneous clinical rating scale in De novo idiopathic Parkinson's disease patients. Pharm Res. 2017;34:2109–18. https://doi.org/10.1007/s11095-017-2216-1.
31. Chen C, Jönsson S, Yang S, Plan EL, Karlsson MO. Detecting placebo and drug effects on Parkinson's disease symptoms by longitudinal item-score models. CPT Pharmacometrics Syst Pharmacol. 2021;10:309–17. https://doi.org/10.1002/psp4.12601.
32. Rigaux C, Sebastien B. Evaluation of non-linear-mixed-effect modeling to reduce the sample sizes of pediatric trials in type 2 diabetes mellitus. J Pharmacokinet Pharmacodyn. 2020;47:59–67. https://doi.org/10.1007/s10928-019-09668-x.
33. Kim TH, Shin S, Shin BS. Model-based drug development: application of modeling and simulation in drug development. J Pharm Investig. 2018;48:431–41. https://doi.org/10.1007/s40005-017-0371-3.