Patient Reported Outcomes in Multiple Sclerosis Clinical Trials: Measurements Lessons From the EXPAND Study

EP1234

Jeremy Hobart[1], Pamela Vo[2], Suzannah Ryan[2], Sophie Arnould[2], Laurie Burke[3]

[1]Peninsula Schools of Medicine and Dentistry, University of Plymouth, Plymouth, UK, [2]Novartis Pharma AG, Basel, Switzerland, [3]LORA Group, LLC, Normal, IL, US

Introduction

- Patient-reported outcomes (PROs) are widely used in multiple sclerosis (MS) clinical trials to evaluate treatment efficacy and measure health utilities for use in cost effectiveness analysis.
- However, detailed examinations of PRO performance within study data are rare, especially in secondary progressive MS (SPMS).
- Such examinations are valuable as PROs have unique properties, context-of-use dependent characteristics, and performance requirements.
- Knowledge of these features facilitates accurate interpretation of clinical trials results.

Objective

- To examine EXPAND study PRO properties, and their implications for clinical trial results.
- To highlight some important lessons learned.

Methods

EXPAND study and PROs examined

- The EXPAND study was an event-driven and exposure-driven, double-blind, phase 3 trial assessing the effect of siponimod versus placebo in patients with SPMS[1].
- EXPAND consisted of a randomized, double-blind, placebo-controlled 'core' part in patients with SPMS, followed by an open-label 'extension' part that will continue for up to a total of 7 year[2].
- Of the 1651 patients randomized, 1327 completed the core part (median duration 21 months, range <1 month to 37 months).
- Here, we examined data from the 12-item MS walking scale version 2 (MSWS-12v2), and the 3-level version of the EuroQol 5-dimension health status measure (EQ-5D-3L).

Analyses

The impact of baseline PRO score distributions on change measurement

- PROs have restricted measurement ranges that could constrain change measurement.
- As expected based on enrolment criteria, distribution of baseline MSWS-12v2 scores was skewed, with more people in the upper (more disabled) half of the scale range (**Figure 1**). Since SPMS is characterised by a progressive accumulation of disability over time, in EXPAND, skewness is expected to increase over time.
- We examined MSWS-12v2 change magnitude in participants grouped by their baseline MSWS-12v2 score.

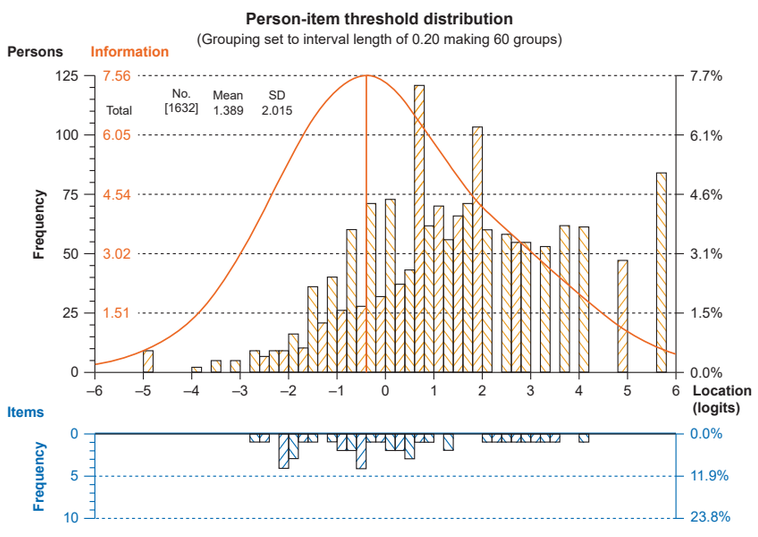
Response dependence

- In the measurement of change when the same instrument is used on more than one occasion, concern has been raised that responses on the second occasion will be similar to responses on the first occasion simply because the same items are responded to even if the property being assessed has changed.
- This form of dependence has been referred to as response dependence[3]. RD has never been examined in PRO clinical trial data.
- We used Rasch Measurement Theory (RMT) to identify, quantify and resolve RD in MSWS-12v2 data.

EQ-5D-3L

- There are several performance requirements of the EQ-5D-3L for valid health utility measurement resulting in useful conclusions about cost-effectiveness measurement in MS studies. For example, empirical evidence should support the single score derived from its 5 items (Health Utility Index, HUI); the HUI should reflect health status as perceived by PLwMS; the EQ-5D-3L should detect change when change occurs. We examined these measurement requirements

Figure 1. EXPAND study: baseline distribution of MSWS-12v2 RMT scores (n=1632)



Results

The impact of baseline PRO score distributions on change measurement

- We examined placebo arm participants with MSWS-12v2 data at baseline and 18 months as this sample had both a reasonable size and time interval.
- **Figure 1** shows the baseline MSWS-12v2 skew was more evident than the numerical indicators implied (mean=1.39 logits; floor effect=5.1%; skewness =0.1466).
- **Table 1** shows the MSWS-12v2 RMT Time 1, Time 2 and change scores for participants grouped by their baseline MSWS-12v2 raw score. The group change score differences are substantial (Low=1.048 logits; H-2=0.081 logits, a 12.9-fold difference).

Table 1. Mean MSWS-12v2 RMT scores and score changes for groups defined by their MSWS-12v2 raw score at baseline

Group[1]	n	MSWS-12v2 Raw score range	MSWS-12v2 RMT scores Mean, SD (range) Time 1	Time 2	Mean change (T2 – T1)	Relative Change[2]
Low	55	0 to 21	–1.289, 0.896 (–3.856 to –0.121)	–0.241, 2.066 (–4.451 to 6.147)	1.048	**100%**
High	195	22 to 42	2.199, 1.711 (0.024 to 6.147)	2.605, 2.110 (–3.436 to 6.147)	0.406	**39%**
Subgroup						
H-1	94	22 to 31	0.812, 0.499 (0.024 to 1.640)	1.567, 1.929 (–3.436 to 6.147)	0.756	**72%**
H-2	101	32 to 42	3.490, 1.398 (1.881 to 6.147)	3.572, 1.795 (–0.977 to 6.147)	0.081	**7.7%**

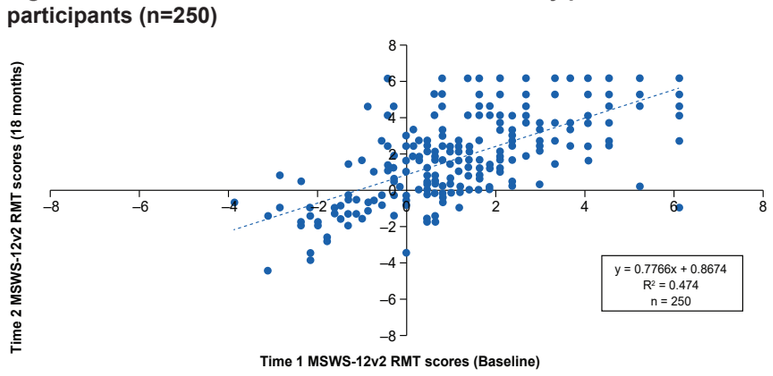
[1]First, we divided the sample into two groups at the scale midpoint. Next we divided the High group into two subgroups, H-1 and H-2 at the High group scale midpoint

[2]Computed as percent of the largest change group, e.g., High group = 0.406/1.048 = 39%

[4-6]RMT is a measurment theory that defined, conceptually and mathematically, the requirements for achieving measurements - as defined scientifically - from item response data. RMT scores are interval-level estimates, ranging from –3.856 to +6.147, derived from item responses using the Rasch model. Raw scores are ordinal level estimates, ranging from 0-42, derived by summing item scores.

- **Figure 2** shows that measurement appears constrained at the right-hand end of the diagram.
- Detailed simulation studies confirm that these group differences in change magnitude are primarily due to constrained measurement rather than regression to the mean, response dependence or latent correlations between baseline and Time 2 RMT scores.
- Although in subgroup with different walking disability level, progression rate over time could be different, a relative change of that magnitude is not expected.

Figure 2. MSWS-12v2 RMT scores for EXPAND study placebo participants (n=250)



Response dependence

- RD was studied in participants with MSWS-12v2 data at baseline and 6 months to provide a sample large enough to identify, quantify and resolve RD.
- Average RD across items was 0.7 logits. We estimated MSWS-12v2 change scores, with (raw) and without (resolved) RD, for groups defined by their baseline MSWS-12v2 scores.
- **Table 2** shows the MSWS-12v2 RMT Time 1, Time 2 (raw and resolved for RD) and change scores (raw and resolved) for the total, siponimod and placebo treated samples.
- In all groups, for all samples, RD resulted in measured change underestimating resolved change.
- MSWS-12v2 change with RD present was 35.4% to 91.6% of change magnitude with RD resolved.
- RD had the greatest impact in the subgroup most affected by constrained measurement (H-2), and with the largest subsample size.

Table 2. Impact of response dependence on change scores

Group	N	Time 1[4] MSWS-12v2 scores Raw Score: range	RMT score: M, SD (range)	Time 2[3] MSWS-12v2 RMT scores Mean, SD Raw	Resolved[5]	Change in MSWS-12v2 RMT score Mean (T2-T1) Raw	Resolved	Relative change (raw/ resolved %)
Total sample n=1501								
Low	373	0 to 21	–1.105, 0.992	–0.544, 1.749	–0.434, 1.984	0.561	0.672	83.5%
High	1128	22 to 42	2.140, 1.539	2.088, 1.948	1.993, 2.143	–0.052	–0.147	35.4%
H-1	495	22 to 31	0.815, 0.442	1.141, 1.540	1.178, 1.888	0.326	0.362	90.1%
H-2	633	32 to 42	3.176, 1.273	2.829, 1.913	2.631, 2.115	–0.347	–0.546	63.6%
Siponimod subsample n=995								
Low	248	0 to 21	–1.145, 1.051	–0.642, 1.727	–0.539, 1.950	0.503	0.606	83.0%
High	747	22 to 42	2.192, 1.534	2.086, 1.955	1.978, 2.146	–0.106	–0.214	49.5%
H 1	317	22 to 31	0.839, 0.432	1.145, 1.578	1.172, 1.922	0.306	0.334	91.6%
H 2	430	32 to 42	3.190, 1.266	2.779, 1.916	2.573, 2.111	–0.410	–0.617	**66.5%**
Placebo subsample n=506								
Low	125	0 to 21	–1.026, 0.862	–0.352, 1.785	–0.255, 2.041	0.674	0.801	84.1%
High	381	22 to 42	2.038, 1.546	2.093, 1.936	2.022, 2.138	0.055	–0.017	76.4%[6]
H 1	178	22 to 31	0.773, 0.458	1.134, 1.472	1.187, 1.830	0.361	0.414	87.2%
H 2	203	32 to 42	3.147, 1.290	2.933, 1.907	2.753, 2.125	–0.214	–0.394	**54.3%**

[3]Time 2= 6 months post randomisation; [4]Time 1 = baseline; [5]Resolved for response dependence; [6]Computed as 0.055 / (0.055+0.017). Not change, when resolved for response dependence is in the other direction (improvement rather than deterioration)

RMT scores are interval-level estimates, ranging from –3.856 to +6.147, derived from item responses using the Rasch model.

Raw scores are ordinal level estimates, ranging from 0-42, derived by summing item scores.

EQ-5D-3L

- Neither weighted nor unweighted item scores satisfied statistical criteria for generating a single score according to clinical test theory (Cronbach's alpha=0.61), RMT (person separation reliability=0.56), or factor analytic methods (first component explains 40% of variance).
- **Figures 3** and **4** show health status measured by the EQ-5D-3L HUI was not normally distrubuted (**Fig 4a**) and poorly related to health status measured using the patient completed visual analogue scale (**Fig 3**), which was normally distriubted (**Fig 4b**).

Figure 3. Relationship between health status measured by the EQ-5D-3L HUI and health status measured by the visual analogue scale (VAS) at baseline (n=1623)

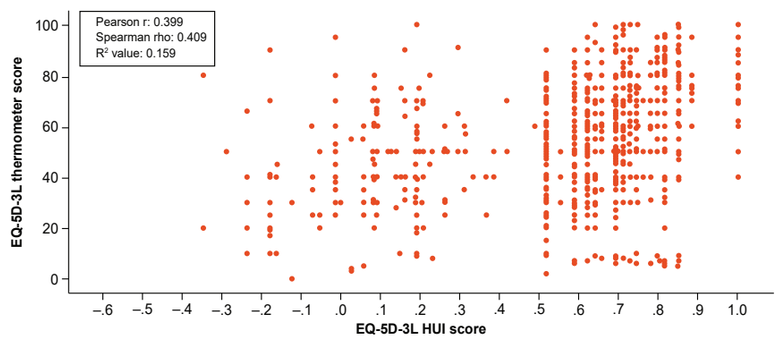
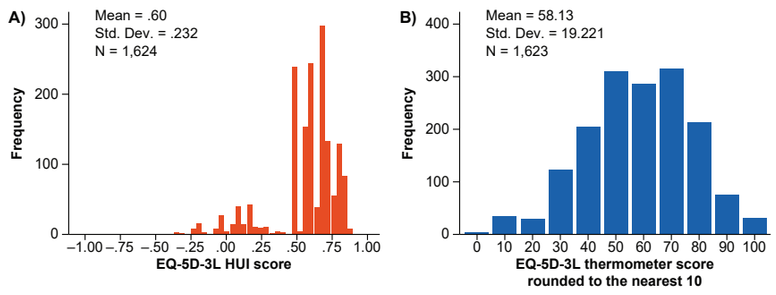


Figure 4. Distribution of EQ-5D-3L HUI and VAS scale scores at baseline (n=1624, 1623)



- **Figure 5** shows the results of mapping the EQ-5D-3L mobility item scores to MSWS-12v2 item scores. The middle EQ-5D-3L score (red bar = some problems) can be associated with a very wide range of MSWS-12v2 item scores.
- **Table 3** shows that 12 months into EXPAND, ≈91% of participants had the same EQ-5D-3L mobility item score, when only ≈7% has the same MSWS-12v2 score. Similar results were found for the other four EQ-5D-3L items. This means that the EQ-5D-3L is not the appropriate utility instrument to use in order to detect change over time.
- Findings question the clinical meaningfulness of the HUI, the extent to which the HUI indicates health status perceived by PLwMS, the weighting process, and indicate a very limited ability to measure change compared with other PROs.

Figure 5. Results of calibrating the EQ-5D-3L mobility item with the MSWS-12v2 items

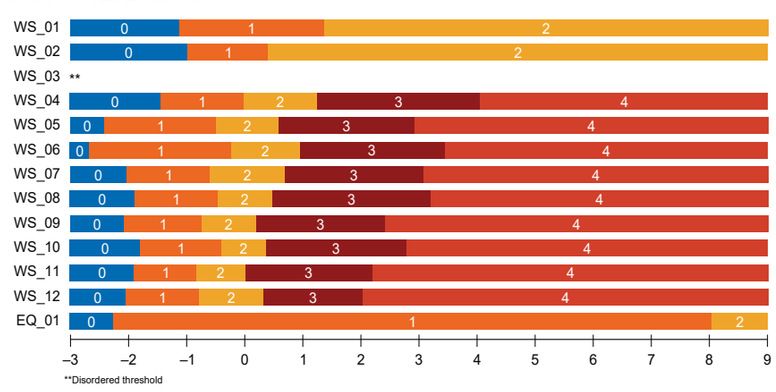


Table 3. Change in EQ-5D-3L mobility item

Score change at 12 months	EQ-5D-3L mobility item	MSWS-12v2
Worse	65 (4.8%)	721 (53.0%)
No change	**1235 (90.8%)**	**99 (7.3%)**
Better	60 (4.4%)	541 (39.8%)
Total	1360	1361

Conclusions

- All measurement issues identified cause Type II error. These can underestimate treatment effects, differences between groups and cost effectiveness.
  - For example, The EQ-5D mobility item was shown to have poor ability to detect changes and differences, and MSWS-12v2 was shown to have a skewed distribution that constrained ability to detect change in more disabled people, The PRO (and others) were shown to be impacted by response dependence.
- Lessons for future MS clinical trials include:
  - Carefully match PRO scale range and BL sample score distribution to enhance change measurement.
  - Use methods that identify, quantify and account for response dependence.
  - The EQ-5D is a suboptimal health utility measure due to the instrument's psychometric property's inability to detect change for the MS population.

References

1. Kappos L et al., *The Lancet* 2018; 391: 1263–1273
2. Soldan P et al., *Neurology* 2015; 84: 81–88
3. Andrich D et al., *Statistical Methods in Medical Research* 2017; 27:3709-3725
4. Hobart J et al., *Lancet Neurol* 2007; 6: 1094–105
5. Andrich D et al., *Psychometrika* 1978; 43:561-573
6. Rasch, G. (1960). Probabilistic model for some intelligence and achievement tests. Copenhagen: Danish Institute for Educational Research

Acknowledgements

The authors acknowledge the following Novartis employees: **Neha Kulkarni** and **David McMinn** for medical writing assistance and coordinating author reviews, **Daniela Piani-Meier** for data analysis and **Vajhula Sarma** for creative design assistance. The final responsibility for the content lies with the authors.

Disclosures

Jeremy Hobart has received consulting fees, honoraria, support to attend meetings or research support from Acorda, Asubio, Bayer Schering, Biogen Idec, F. Hoffmann-La Roche, Genzyme, Merck Serono, Novartis, Oxford PharmaGenesis and Teva; **Pamela Vo**: Novartis employee; **Suzannah Ryan**: Novartis employee at time of poster development; **Sophie Arnould:** Novartis employee; **Laurie Burke**: Has provided strategic advice to and participated in clinical outcome assessment development initiatives on behalf of multiple pharmaceutical industry companies

This study was funded by Novartis Pharma AG, Switzerland.

Poster presented at the 38th Congress of the European Committee for Treatment and Research in Multiple Sclerosis, 26–28 October 2022.

Visit the web at: https://bit.ly/ectrims2022

Presenter email address: jeremy.hobart@plymouth.ac.uk

Scan this QR code to download a copy Poster