Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Putting the Meaning Into Meaningful Change Research

Jessica Braid
Hoffmann-La Roche Ltd
Susanne Clinch
Hoffmann-La Roche Ltd
Hannah M Staunton
Hoffmann-La Roche Ltd
Patricia K Corey-Lisle (coreylip@gene.com)
Genentech Inc https://orcid.org/0000-0003-2148-6209
Bruno Kovic
Roche Canada: Hoffmann-La Roche Limited
Siobhan Connor
Hoffmann-La Roche Ltd
Teofil Ciobanu
Hoffmann-La Roche Ltd
Thomas Willgoss
Hoffmann-La Roche Ltd

Research Article

Keywords: patient-friendly, Industry perspective, interpreting group-level MCIDs

Posted Date: May 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-386476/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Purpose

Methods for deriving clinically meaningful change thresholds have advanced considerably in recent years, however, key questions remain about what the identified change score actually means for an individual patient or group of patient. This is particularly important in the case of ClinROs where the translation from clinically meaningful change to patient-relevance in daily living is not clear. This paper provides case studies from an Industry perspective, where we have addressed this challenge using varied approaches. We have explored meaningful change at both the group and individual level.

Methods

We provide several case studies to illustrate different approaches to understanding and communicating a meaningful outcome on a ClinRO. These include alternative methods for interpreting group-level MCIDs, and several examples of linking ClinRO items to patient-relevant real-world concepts e.g. through exit interviews, translation of ClinRO items into patient-friendly concepts, and use of the Rasch model to equate ClinRO items to real-world functional measures.

Results

Each case study provides unique learning opportunities. For example, contextualising group-level differences, converting MCIDs into other metrics like numbers needed to treat and responder deltas supports interpretation of clinical meaning, especially for clinicians. For interpreting individual-level meaningful change, exit interviews and the development of patient-friendly versions of ClinROs provide a means of linking clinician-focused content to real-world functional outcomes in a meaningful way for patients. Finally, the Rasch model can help predict probable item scores on a ClinRO associated with the threshold at which a function is gained or lost.

Conclusion

While methods for deriving meaningful change thresholds have evolved, there remains a significant challenge in communicating what observed changes mean to the patient, a challenge which is further complicated in ClinROs. These case studies showcase novel approaches to addressing this challenge and may provide a useful addition to the COA scientist's toolbox.

Introduction / Purpose (Stating The Main Purposes And Research Question)

Clinical Outcomes Assessments (COAs), defined by the U.S. Food and Drug Administration (FDA), should assess a concept(s) that has clinical and patient relevance, and affects how a patient feels, functions, or survives (FDA PFDD 4 2019). There is increasing focus from regulators, payers, clinicians, and patients on establishing the meaning behind changes in individual patient scores - beyond statistical significance

(FDA PFDD 4 2019; IQWiG 2020; Guralnik et al. 2020; Chatham et al. 2018). Demonstrating clinical and patient relevance to support a clinically meaningful change is necessary for all COAs, particularly for clinical trial endpoints. This represents an industry-level challenge in the establishment and communication to key stakeholders.

Methods for establishing clinically meaningful change thresholds have developed during the past decade, with an increasing focus on the need for anchor-based methods, supported by distribution-based evidence (FDA PFDD 3 2018). A key question remains: what does the change score/threshold mean for an individual patient or group of patients? (Coon & Cook 2018). This is particularly pertinent for Clinician-Reported Outcome (ClinRO) measures, used when patients are unable to self-report on their own health status or when clinical judgement is required to make an assessment (Powers et al. 2017). For ClinROs, the translation from a clinical trial score change into real-world impact is not always apparent. Multiple stakeholders require a clear link between a trial-based clinical change and real-life outcomes (e.g., Lievens et al. 2019).

The objective of this paper is to provide four illustrative cases where seeking to establish the clinical meaningfulness of the concept measured and the change scores of ClinROs through a variety of approaches. We explore examples at the group and individual level, with a focus on the variety of methods used and not the data examples presented here. Each case study provides examples of how we methodologically addressed the questions, and clearly identifies whether the methods within the case study support an evaluation of ***group*** or ***individual-level*** change. The results focus on how we propose interpreting the outcomes. Three of the case studies focus on establishing patient-relevance of the meaning of score changes on ClinROs, with the last case study focusing on clinical approaches.

Methods And Results

Case study 1: Patient-centered perspective on individual-level meaningful change estimates for a COA, what do changes mean within a target population?

Context

Translating a ClinRo to patient-relevance is a common challenge. We describe a qualitative approach translating ClinRO items into patient-friendly concepts, to allow in-depth exploration of how functional abilities relate to ability to perform activities of daily living and how changes in functioning can impact Health-related Quality of Life.

Methods

The Motor Function Measure 32 (MFM32) is a ClinRO measure, developed based on clinical expert input, assessing the motor function abilities of individuals with neuromuscular disease. It has been validated for use in individuals with Type 2 and Type 3 SMA aged 2–60 years (Berard et al. 2005; Vuillerot et al. 2013; Trundell et al. 2020). The MFM32 includes 32 items scored on a 0-3 point Likert scale [0 (unable to

complete the ability) to 3 (able to complete the ability)] which are summed and transformed to a 0-100% total score with higher scores associated with greater functioning. In-depth, semi-structured, qualitative interviews were conducted with individuals with SMA and caregivers from the US. In order to facilitate discussion with patients, a patient-friendly version of the MFM32 was created using the MFM User Manual (Berard et al. 2021) and input from clinical experts and patient advocacy groups. Clinical terminology used in the items of the validated clinician-reported version of the MFM32 were reworded into patient-friendly language, maintaining focus on the specific ability assessed by items. Participants were asked to describe the Activities of Daily Living (ADLs) considered to be related to the functional abilities assessed in the MFM32 (Berard et al. 2021), the relevance of score changes (item and total score level) the impact these changes might have on their ability to perform activities of daily living using a patient-friendly version of the MFM32.

Results

Participants were able to relate one or more ADLs to each of the functional abilities assessed in the MFM32 and provide the perspective that maintaining functional ability as assessed by a patient-friendly version of the MFM32 was considered a meaningful outcome. This demonstrates the effectiveness in the approach of translating a complex ClinRO into a patient-friendly alternative, to aid discussion on meaningful change with patients. The results obtained in this study are particularly important in a progressive disease, where the patient's perspective on how functioning - as assessed by a ClinRO - could be related to real-world ADLs. A limitation of this approach is that the clinician-reported version of the MFM32 provides a more detailed assessment of motor function, and assesses intermediate functions that were difficult to explain in the context of the interviews, which instead focused on describing the functional ability associated with achieving a maximum score on each item. The changes in scores discussed are theoretical and the fact that a patient-friendly MFM32 item can be associated with an everyday activity does not mean that change on an item would necessarily lead to changes in that specific daily activity in the real-world. A limitation is that other clinical (e.g. presence of contractures or scoliosis) and situational (e.g. use of assistive devices) factors may influence actual functioning. This work has been accepted for publication in a peer reviewed journal (Duong et al. 2021).

Case study 2: Rasch measurement model for approximation of score changes: Can clinical changes translate into meaningful changes in functioning at the individual patient-level?

Context

A challenge in rare diseases is the heterogeneity of patients' symptoms, especially prominent when communicating the meaning of changes on a ClinRO across the spectrum of functional ability. Qualitative data supported the understanding of the relationship of MFM32 to daily life, however, regulatory and payer questions focused on the relevance of point changes in the context of a total score which ranged from 0-100.

Methods

Trundell et al. (2019) used a Rasch measurement model to predict probable item scores on the MFM32 based on an individual's level of underlying motor function. As described by Hobart and Cano (2009) the Rasch measurement model provides a method to score patients according to their abilities, presenting items according to their difficulties on the same linear interval scale. In this example, probable item scores on the MFM32 were predicted based on an individual's level of function on the underlying spectrum (i.e., from a logit [log-odds-unit] which is the unit of measurement in Rasch) (Andrich, 2011). Trundell et al. (2019), described a provisional set of analyses using the following steps to establish the relationship between MFM32 item scores and daily functions. In italics, an example is provided for each step:

1. Qualitative data sources reviewed to identify daily activities impacted by SMA, as reported by patients and their families.
    - *Turning in bed was raised as important to patients and their families*
2. Physicians (N=2) and physiotherapists (N=6) experienced in administering and interpreting the MFM32 in patients with Type 2 and 3 SMA identified key MFM32 items that were most related to each daily activity.
    - *Item 7 on the MFM32 (ability to roll from supine to prone position) was identified as the item most related to the ability to turn in bed*
3. For each function, an MFM32 item (or items), and the threshold between response options most likely to predict ability to perform the daily activity, were identified by the expert panel.
    - *Experts identified a change in score from 1 (individual can roll partially) to 2 (individual can turn over into prone with difficulty and compensatory movements and/or cannot free the upper limbs from under the trunk) as the threshold for gaining/losing the ability to roll in bed*
4. Rasch analysis was conducted using two independent data sources (one from a clinical trial and the other from a real-world data source) to identify the logit associated with each response threshold. For abilities with multiple items, the mean of the logits was used. Where items had disordered thresholds, categories were combined until thresholds were ordered.
    - *The item characteristic curves (curves showing the ordering of response options for each item; for an example see Petrillo et al. 2015) for item 7 was used to identify the threshold of being equally likely to score 1 or 2 on the logit scale.*
5. For each daily activity, the specified logit was used to identify the most probable item response (0–3) for each of the other MFM32 items, from which a total score (0–100) was calculated.
    - *This logit value was then used to calculate the most probable item score for all other items using the item threshold map (allows prediction of the most likely responses for each item*

based on a person's location on the logit scale (i.e., on the underlying construct of motor function ability). For the ability to turn in bed, the resulting score was 52 on the MFM32 total score, which is in the middle of the scale.

Results

The result of this approach was a provisional figure depicting estimated MFM32 score thresholds associated with the gain or loss of meaningful daily activities (Figure 1).

The value of this approach is the ability to link the MFM32 items to real-world daily activities and to quantitatively approximate gain or loss of activities in the context of an MFM32 total score. These estimates were not used to form endpoints but rather to provide a schematic to demonstrate the relevance and meaning of score changes across the MFM32 score spectrum. While there are limitations to this method (e.g., imperfect fit to the rasch model due to the multidimensional nature of the scale), and it is not recommended as the sole means for determining score thresholds, the approach provides a unique way to visualise that patients gain or lose meaningful daily activities, regardless of their starting position on the scale. This analysis moves beyond a *one-size-fits-all* approach to establishing what a change score on a ClinRO may be in relation to a patient's starting functioning. This work has been expanded upon using an additional dataset and this research has been submitted for publication in a peer reviewed journal (Trundell et al. 2021).

Case study 3: Can exit interviews be used to provide context for the meaning of a change in quality of life (QoL)/functioning?

Context

The heterogeneity of autism spectrum disorder (ASD) in symptoms, symptom changes and impact is a well-established challenge. Additionally, available core symptom measures were not developed specifically for ASD. Additional work to understand what constitutes a meaningful change for available measures is required to validate the assessment and optimally interpret clinical trial data and real-world outcomes.

Methods

The Vineland^{TM}-II Adaptive Behavior Scales – Second Edition (Vineland^{TM}-II; Sparrow et al. 2005) is one of six measures of social communication recommended by an expert panel to assess measures for their utility in ASD clinical trials (Lord & Jones 2012). Although it was acknowledged that this complex ClinRO has limitations, it has been used in multiple clinical trials, such as those evaluating balovaptan (VANILLA [NCT01793441] (Bolognani et al. 2019), V1aduct [NCT03504917], and aV1ation [NCT02901431]). To characterize changes experienced from the trials, and support and contextualize changes reported from the Vineland^{TM}-II domains, exit interviews with study partners were conducted as part of aV1ation, involving children (age 5-12) and adolescents (13-17) with ASD, and V1aduct, involving adults (age 18+)

with ASD. Interviews consisted of in-depth, 60-minute, telephone-based conversation with study partners of trial participants who recently completed the final, Week 24 visit. Open-ended questions were asked to gain insights into the impact of ASD on the lives of individuals and their families, followed by a focused discussion on the meaning of any changes experienced over the course of the trial. In particular, questions focused on the domains captured by Vineland-II™: socialization, communication, and daily living skills, with discussion on changes in health-related QoL. Interview data was audio-recorded, transcribed and analyzed using thematic analysis to understand meaning and real-world impact of changes. Quantitative ratings of change from baseline, overall and per domain (based on 7-point scales), were completed by study partners during the interview, with the intent to categorize exit interview participants for quantitative evaluation in the psychometric analyses. Experienced, independent clinician reviewers also rated changes, using the same rating scales as the study partners, for a number of transcripts (n=20). This approach required clinicians to review selected transcripts and rate the perception of change from their independent clinical perspective. These data, alongside blinded descriptive data from selected outcome variables in the clinical trials, was used to inform anchor-based analyses exploring the interpretation of meaningful change on Vineland™-II.

Results

This approach provided insight into the study partner perceived change in the trial participant. When study partners were asked to rank domains in order of importance, socialisation and communication were rated most highly for both aV1ation and V1aduct, thereby supporting the primary endpoint used in these studies. Upon review of the interviews and unblinding of treatment groups, the interviews provided additional context to changes (of which most were improvements) reported in the placebo and active drug arm, supporting the results from the COAs also collected in the trial, somewhat validating the trial findings. In aV1ation, the majority of study partners reported some form of improvement in the child they cared for, with few reporting no change and even fewer reporting worsening, whereas in V1aduct the majority reported no change. In aV1ation, when this anchor data and clinical trial data were triangulated to calculate meaningful change thresholds for the Vineland™ -II, one noticeable difference between the clinician and study partner feedback from the interviews was that clinicians were assessing change above a certain amount as being meaningful (based on the clinical anchors). However, from the study partner perspective any change, no matter how small, was usually considered meaningful. This suggests the meaningful change threshold that was being used as a benchmark, which had involved clinician input, was conservative, so potentially a lower value may still have been meaningful to study partners and people with ASD.

In aV1ation, the interviews also provided insights from study partners regarding changes which may have been a result of other life events, not attributed to the drug. When study partners were asked if they thought the trial participant was on placebo or active drug, of those who selected placebo, just under half reported a minimal improvement when asked to rate using the anchor. These same participants reported that such changes may have been a result of maturation, starting high school, moving or a personal desire to change. This important finding could be used to explain outliers and highlights the importance

of the mixed methods approach used in this study to better understand life changes throughout the duration of a clinical trial when interpreting results.

A limitation is that these are reliant on the willingness of study partners, which may result in a potential selection bias. As the exit interviews were consistent with the clinical trial data, it was likely a representative sample of trial study partners. An additional consideration is that only study partners were involved in these interviews, not ASD participants. Thus, it can't be assumed the changes were consistent with ASD trial participants' perception of change. The study was designed to account for potential difficulties ASD participants may have had communicating their experiences and therefore accurately elaborating on any changes. A general limitation applicable to trials in this population and applicable to the optional exit interviews is that the willingness of study partners to see an improvement could have led to changes being observed and reported more sensitively in the placebo and treatment arms. Overall, the exit interviews were received well and also allowed for general feedback on the positive impact of changes experienced from the study, with one study partner noting:

"If the people in the pharmaceutical company are going to read it, I would say that for families that are on the spectrum like us, it's life changing work what they're doing. It gives us the opportunity to, to fit and to hope for a fair future for our kid, that sometimes get shattered when we get the diagnosis. On a personal level I would like to say it feels like a miracle. So thank you."

Case study 4: Group level interpretation of ClinRO meaningful change thresholds within a Clinical Trial

Context

In designing pivotal trials to assess efficacy and safety of new therapeutics, drug developers should endeavor to power a study to detect a clinically meaningful treatment effect. Powering is done at a group level whereby confidence is needed to support that an observed difference between groups is clinically important. Though recent guidance has moved from group-level meaningful change towards the more patient-relevant concept of individual response, group-level evaluation continues to be important among clinical decision makers where group level inferences can inform comparisons between different treatments or decisions regarding public policy (Rai et al. 2015).

While anchor- and distribution-based methods (e.g. Copay et al. 2007) are proposed to support arguments, when there are limited data to support meaningful differences, or in indications where there is a high unmet need and limited treatment options available, clinical expert opinion is often sought to inform a relevant and meaningful treatment effect. The evaluation of relative clinical importance of an observed group level difference is context-dependent and is dynamic within an ever-evolving treatment landscape.

This is important when primary endpoints are derived from ClinROs where clinician opinion plays a prominent role. This study illustrates clinician-centric metrics support for discussions on a meaningful group-level difference on a complex multi-domain ClinRO.

Methods

This example is also based on research in ASD. We sought to define a clinically meaningful treatment effect on the Vineland™-II (Sparrow et al. 2005). The Vineland™-II is a complex ClinRO that relies on normative data to support a standardized scoring algorithm. This complexity, combined with a lack of data on meaningful change thresholds and a lack of existing treatments, meant that clinical experts struggled to identify a treatment difference that would be meaningful. To overcome this, we utilized three different clinically-relevant metrics to support discussions:

- Responder delta
- Numbers needed to treat (NNT)
- Standardized effect size

Results

Table 1 illustrates different metrics using a hypothetical ClinRO which has a standardized score ranging from 0-100 points (higher score indicating better functioning) and assumed standard deviation (SD) of 10.

Table 1: Illustrative data showcasing three alternative metrics for contextualizing a group-level difference

Group-level difference	% Responders Active (N=100)	% Responders Placebo (N=100)	Responder delta	Numbers Needed to Treat (NNT)	Effect size*
3 points	45	35	10%	9.60	0.3
4 points	50	35	15%	6.52	0.4
5 points	55	35	20%	4.95	0.5
6 points	60	35	25%	3.99	0.6
7 points	65	35	30%	3.33	0.7
8 points	70	35	35%	2.85	0.8

*assumes SD of 10

1. Difference in proportion of responders (responder delta)

The relevance of change is inherently more meaningful when individual-level changes are presented. Using either established or illustrative thresholds for clinically meaningful individual change, it is possible to model the proportion of responders in each treatment group by adjusting group-level treatment effect and keeping the placebo arm constant.

In considering the responder delta, clinical experts were better able to contextualize the group-level differences.

2. Numbers needed to treat

Building on the responder delta, another clinical concept is NNT. NNT can be defined as the average number of patients who need to have the treatment for one to have the positive outcome in the time specified (NICE Glossary 2021). The closer NNT is to 1, the more effective the treatment (CEBM 2021).

Citrome and Ketter (2013) posit NNT is one of the most clinically intuitive metrics which helps clinical decision makers evaluate the real-world effectiveness of a documented effect size. Even when no similar treatment is available, we have found that clinicians are able to evaluate the clinical relevance of any given NNT in the context of other treatments. Although numbers needed to harm (NNH) is required to obtain a full picture of the relative benefit-risk of a treatment, a hypothesized NNT value provides a unique and valuable lens for evaluating group-level treatment differences.

3. Standardized effect size

Finally, a standardized effect size is a metric that is well known to clinical experts. Effect size is traditionally used to support interpretation of clinical research (alongside statistical significance). Labels such as small (0.2), medium (0.5) or large (0.8) (Cohen, 1988) are often used by prescribing clinicians to consider the relative efficacy of a treatment and the resulting clinical importance.

Converting group-level Meaningful Change Scores into well-known metrics like NNT, standardised effect size and responder deltas can enable a panel of clinical experts to better articulate the clinical meaning of a specified change on a ClinRO. This can be valuable during the early development of new therapies where expert clinical opinion is sought to help powering for a clinically meaningful effect.

Discussion

These case studies highlight a series of novel approaches that have been employed to derive and demonstrate the meaningfulness of score changes on ClinROs to support stakeholder decision-making. Regardless of context (clinical trial or real-world evidence generation) and whether a ClinRO is being used to generate individual or group level data, the importance of deriving meaning from data by ‘going beyond the numbers’ to define relevance to patients and clinical experts stands true.

Important learnings include:

- Case study 1, the patient perspective is vital to determining the clinical relevance of items to their everyday lives, and providing patients with a patient-friendly version of the concepts measured in a ClinRO could be a useful approach to obtaining these insights.
- Case study 2, in heterogeneous conditions, different ADLs may be gained or lost depending on patients’ functional ability. Visual approaches using the Rasch measurement model can depict this

reality to external audiences such as regulators and payers.

- Case study 3, exit interviews could provide an opportunity to understand real-life examples of how study partners perceived change in study participants; identify the importance of additional considerations when thinking about disease improvement in supporting concepts covered by primary endpoints in clinical trials; highlight the importance of the attribution of treatment effect, and that this effect can be driven by multiple factors; that perceived change can be variable across different respondents.
- Case study 4 demonstrates the utility of clinician expertise is important supporting the interpretation of meaningful difference on a ClinRO at the group level using a variety of well-known metrics.

Although the innovation of the different approaches delivered value to improving interpretability of score changes on ClinROs, the case studies also had several important limitations worth considering. First, the approaches presented here should be considered as unique supplementary approaches to the traditional anchor-based approaches for the evaluation of meaningful change on complex ClinROs. The information obtained through these innovative methods cannot be relied upon as the sole source of an estimation. Second, ClinROs used in clinical trials are often measures taken from clinical practice which require validation (i.e., using classical test theory and Rasch measurement theory analytic methods) to ensure optimal performance in a clinical trial. The linearity of such scales is often not verified and as such single point changes may not mean the same across the measure causing challenges with describing and interpreting 'point changes'. It is important that appropriate validation studies detailing scale structure, reliability, validity and ability to detect change have been conducted prior to interpreting the meaning of changes on the scale.

Finally, in terms of future research in this field, it is important to note that in patient populations such as SMA and ASD where a wide degree of heterogeneity exists with regards to functional ability, a single estimate of meaningful change is often not appropriate. The authors encourage greater discussion and consideration of appropriate methods for factoring in baseline function in the calculation of meaningful change estimates.

Conclusion

While methods for deriving meaningful change thresholds have evolved considerably in recent years, there remains a challenge in communicating what observed changes mean to the patient, a challenge which is further complicated in ClinROs. These case studies showcase novel approaches to addressing this challenge and may provide a useful addition to the COA scientist’s toolbox. Our studies focused on neurological conditions and the applicability of our approaches should be verified across multiple therapeutic areas.

Declarations

Acknowledgements

For case study 1, the authors would like to thank collaborators including Tina Duong, Carole Vuillerot, Aurelie Barrier for clinical expert input to the patient-friendly MFM32 and overall study design and interpretation, Sharan Randhawa, Jessica Flynn, Rob Arbuckle for conducting the qualitative interviews and Fani Petridis, Johannes Reithinger, Rosangel Cruz, Jill Jarecki, Mencia De Lemus, Nicole Gusset, Ria Broekgaarden for expert input on study design.

For case study 2, the authors would like to thank Dylan Trundell who led this work and collaborators including Stephanie Le Scouiller, Laurent Servais, Ulla Werlauff, Stefan Cano, Louise Barrett, Marjorie Bernard, Anne Berruyer, Marta Gutiérrez, Dominique Vincent-Genod, Ksenija Gorni, and Carole Vuillerot.

For case study 3 and 4, the authors would like to thank Janice Smith for her input as well as Claire Burbridge, Elizabeth Gibbons and Michael Chladeck from Clinical Outcome Solutions who conducted the exit interviews.

References

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(5), 571-585.

Bérard, C., Payan, C., Hodgkinson, I., Fermanian, J., & MFM Collaborative Study Group. (2005). A motor function measure scale for neuromuscular diseases. Construction and validation study. *Neuromuscular disorders*, *15*(7), 463-470.

Bérard C, Girardot F, Payan C. *User's Manual: MFM-32 & MFM-20,* Accessible at https://mfm-nmd.org/get-a-user-manual/?lang=en (Accessed 29 March 2021).

Bolognani, F., del Valle Rubido, M., Squassante, L., Wandel, C., Derks, M., Murtagh, L., ... & Fontoura, P. (2019). A phase 2 clinical trial of a vasopressin V1a receptor antagonist shows improved adaptive behaviors in men with autism spectrum disorder. *Science Translational Medicine*, *11*(491).

Centre for Evidence-Based Medicine (CEBM), *Number Needed to Treat (NNT)*, Available at https://www.cebm.ox.ac.uk/resources/ebm-tools/number-needed-to-treat-nnt (Accessed 29 March 2021).

Chatham, C. H., Taylor, K. I., Charman, T., Liogier D'Ardhuy, X., Eule, E., Fedele, A., ... & Bolognani, F. (2018). Adaptive behavior in autism: Minimal clinically important differences on the Vineland-II. *Autism Research*, *11*(2), 270-283.

Citrome, L., & Ketter, T. A. (2013). When does a difference make a difference? Interpretation of number needed to treat, number needed to harm, and likelihood to be helped or harmed. *International Journal of Clinical Practice*, *67*(5), 407-411.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillsdale, NJ: Laurence Erlbaum Associates.

Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research*, *27*(1), 33-40.

Copay, A. G., Subach, B. R., Glassman, S. D., Polly Jr, D. W., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal*, *7*(5), 541-546.

Duong, T., Braid, J., Staunton., H et al. (2021). Understanding the relationship between the 32-item motor function measure and daily activities from an individual with spinal muscular atrophy and their caregivers' perspective: a two-part study. In Press.

FDA, U.S. Food and Drug Administration. (2018, October 15-16). Patient-Focused Drug Development Guidance Public Workshop (PFDD3). *Methods to Identify What is Important to Patients & Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments*, Available at https://www.fda.gov/media/116277/download (Accessed 29 March 2021)

FDA, U.S. Food and Drug Administration. (2019, December 6). Patient-Focused Drug Development Guidance Public Workshop (PFDD4). *Incorporating Clinical Outcome Assessments into Endpoints for Regulatory Decision-Making*, Available at https://www.fda.gov/media/132505/download (Accessed 29 March 2021).

Guralnik, J., Bandeen-Roche, K., Bhasin, S. A., Eremenco, S., Landi, F., Muscedere, J., ... & Vellas, B. (2020). Clinically meaningful change for physical performance: perspectives of the ICFSR Task Force. *The Journal of Frailty & Aging*, *9*(1), 9-13.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess*,*13*(12).

IQWIG, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (2020, November 5), *Allgemeine Methoden*, Available at https://www.iqwig.de/methoden/allgemeine-methoden_version-6-0.pdf?rev=180500 (Accessed 29 March 2021).

Lievens, Y., Audisio, R., Banks, I., Collette, L., Grau, C., Oliver, K., ... & Aggarwal, A. (2019). Towards an evidence-informed value scale for surgical and radiation oncology: a multi-stakeholder perspective. *The Lancet Oncology*, *20*(2), e112-e123.

Lord, C., & Jones, R. M. (2012). Annual Research Review: Re-thinking the classification of autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *53*(5), 490-509.

NICE, National Institute for Health and Care Excellence, *Glossary - Number needed to treat*, Available at https://www.nice.org.uk/glossary?letter=n (Accessed 29 March 2021).

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of

worked examples. *Value in Health*, *18*(1), 25-34.

Powers III, J. H., Patrick, D. L., Walton, M. K., Marquis, P., Cano, S., Hobart, J., ... & Burke, L. B. (2017). Clinician-reported outcome assessments of treatment benefit: report of the ISPOR clinical outcome assessment emerging good practices task force. *Value in Health*, *20*(1), 2-14.

Rai, S. K., Yazdany, J., Fortin, P. R., & Aviña-Zubieta, J. A. (2015). Approaches for estimating minimal clinically important differences in systemic lupus erythematosus. *Arthritis Research & Therapy*, *17*(1), 1-8.

Sparrow, S. S., Cicchetti, D., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales-2nd Edition Manual*. Minneapolis, MN: NCS Pearson, Inc.

Trundell, D., Le Scouiller, S., Gorni, K., Seabrook, T., & Vuillerot, C. (2020). Validity and reliability of the 32-item motor function measure in 2-to 5-year-olds with neuromuscular disorders and 2-to 25-year-olds with spinal muscular atrophy. *Neurology and Therapy*, *9*(2), 575-584.

Trundell D, Le Scouiller S, Servais L et al. Using Rasch analysis to estimate thresholds associated with gain/loss of daily function on the Motor Function Measure (MFM). Poster presented at the Cure SMA Researcher Meeting, Poster at 23rd International SMA Research Meeting, Anaheim, CA, June 28-July 1, 2019.

Trundell, D., Scouiller, S., & Servais, L. (2021). *Using Rasch analysis to estimate thresholds associated with gain or loss of daily activities on the 32-item Motor Function Measure (MFM32).* Manuscript submitted for publication.

Vuillerot, C., Payan, C., Iwaz, J., Ecochard, R., Bérard, C., & MFM Spinal Muscular Atrophy Study Group. (2013). Responsiveness of the motor function measure in patients with spinal muscular atrophy. *Archives of Physical Medicine and Rehabilitation*, *94*(8), 1555-1561.

Figures

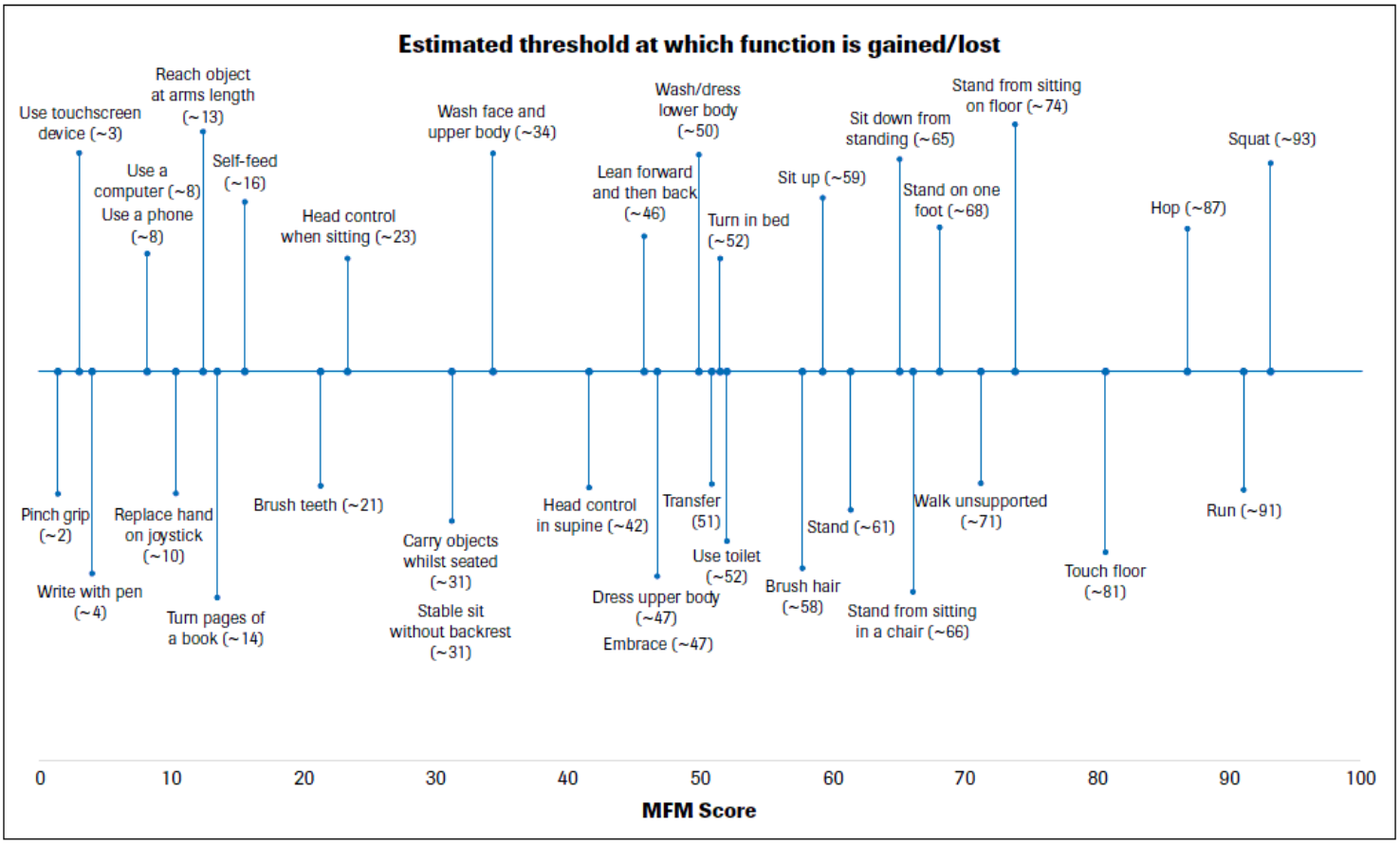

Figure 1

Estimated MFM32 score thresholds associated with the gain or loss of meaningful daily functions