28 February 2020
EMA/CHMP/SAWP/74371/2020
Committee for Medicinal Products for Human Use (CHMP)

Qualification opinion on Multiple sclerosis clinical outcome assessment (MSCOA)

Agreed by Scientific Advice Working Party	17 January 2019
Adopted by CHMP for release for consultation	31 January 2019[1]
Start of public consultation	18 June 2019[2]
End of consultation (deadline for comments)	20 September 2019[3]
Adoption by CHMP	30 January 2020

Keywords	Multiple sclerosis clinical outcome assessment, performance tests, voice of the patient study[4]

[1] Last day of relevant Committee meeting.
[2] Date of publication on the EMA public website.
[3] Last day of the month concerned.
[4] To be identified here during preparation of the concept paper - keywords represent an internet search tool - Rapporteurs to propose and Working Party/Committee to adopt.

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000 An agency of the European Union

Executive summary

The Multiple Sclerosis Outcome Assessments Consortium (MSOAC) seeks qualification of a Clinical Outcome Assessment (COA) instrument. This COA is comprised of a battery of four performance outcome measures assessing important dimensions of multiple sclerosis (MS) - walking (Timed 25-foot Walk, T25FW); hand dexterity (9 Hole Peg Test, 9HPT); vision (Low Contrast Letter Acuity, LCLA), and mental processing speed (Symbol Digit Modalities Test, SDMT)) to assess treatment benefit in clinical trials of therapies for MS. The Concept of Interest (COI) for meaningful treatment benefit is "disability in multiple sclerosis", characterized as neurological or neuropsychological impairments that result in limitation in activities, participation, or roles, which are understood to be important by persons with MS (PwMS). The Context of Use (COU) focuses on the target population of adults with a diagnosis of MS and a relapsing-remitting (RRMS), secondary progressive (SPMS), or primary progressive (PPMS) clinical course. People with RRMS experience worsening disability despite the use of available disease-modifying drugs. SPMS and PPMS present an even greater unmet medical need, as presently only one disease-modifying treatment is available for progressive forms of MS, and the personal, family and societal costs associated with this disease are substantial. An improved COA instrument would enable a more accurate assessment of the therapeutic benefits in people with MS. Another reason for developing a COA instrument for relapsing and progressive MS target populations is that the distinction between them can be unclear, as a continuum exists from early stage and later stage MS. There are many people with MS who could be categorized as having either form of the disease for long periods of time. In addition, irrespective of a person's MS subtype, he/she may accumulate disabilities throughout the disease continuum that this COA instrument will measure. Finally, future therapies for all forms of MS will likely involve non-immune pathways, such as neural repair, and will require more reliable detection of more subtle clinical changes than the currently used Expanded Disability Status Scale (EDSS)-based and other physical endpoints are equipped to measure.

The intent is for this COA instrument to serve as a primary, co-primary, or secondary endpoint to assess efficacy in clinical trials at various stages of drug development, including proof of concept, dose-ranging, confirmatory and registration trials. The four performance measures are considered as a battery of tests, some or all of which could be used as a dysconjugate composite endpoint by sponsors in a clinical trial. For example, the T25W measure would not be used in PPMS and SPMS trials in which participants are non-ambulatory. If used in registration trials, the ultimate language included in product labeling will reflect which measures were used in the trials and would describe the effect of treatment on each measure.

MSOAC acknowledges that measures of ambulation, motor dexterity, and vision have been used in conjunction with EDSS by sponsors in different MS drug development programs, but none have been qualified by a regulatory authority, which could promote their widespread use in the MS community. Moreover, presently, measures to assess cognitive function in MS are not widely used as primary endpoints in clinical trials of MS drugs, so effects of therapy on aspects of cognition are not reflected in drug labelling. Consequently, the qualification of an instrument that includes SDMT would fill an unmet need; since detrimental effects on cognition accounts for much of the socioeconomic impact of MS and this dimension of MS is extremely important to PwMS. Importantly, worsening cognitive function, as measured by SDMT, occurs independently from worsening physical function, as captured by the EDSS or performance measures such as the T25FW, 9HPT and LCLA. Therefore, an instrument that measures a critical aspect of cognition with SDMT, in combination with important physical measures of ambulation, dexterity and vision, fills a measurement gap and provides a much more complete assessment of MS-related disability.

The individual performance measures that are included the COA instrument could be used in conjunction with a range of other performance measures and functional scales as well as other secondary outcome measures, such as imaging, relapse assessment, and Patient-Reported Outcome (PRO) measures.

Based on the coordinators' reports the CHMP gave the following answers:

Question 1

"With the submission of this briefing package, MSOAC is seeking qualification opinion for the COA instrument. The single question for SAWP is "Does EMA agree that the evidence presented in the briefing package is sufficient to support the qualification of the MSOAC instrument?" (p27/ 215 and page 194/215 of the Briefing Document)

CHMP answer

Previous history:

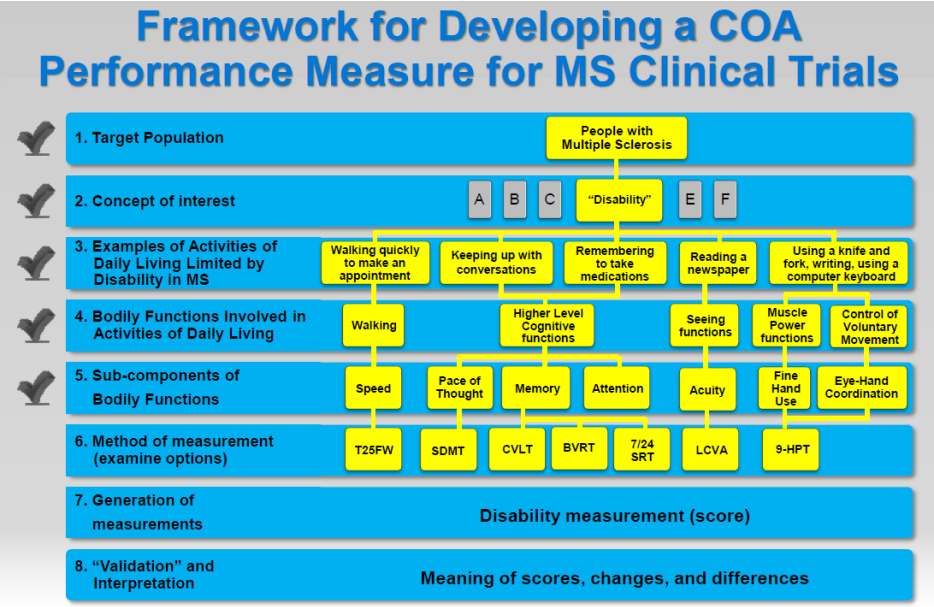
The COA is comprised of a battery of four performance outcome (PerfO) measures assessing important dimensions of MS:

- walking (Timed 25-foot walk, T25FW)
- hand dexterity (9 Hole peg Test, 9HPT)
- vision (Low contrast Letter acuity, LCLA)
- mental processing speed (Symbol Digit Modalities Test, SDMT)`

to assess treatment benefit in clinical trials for MS.

The MSOAC explained the concept of interest (i.e. disability); the context of use (relapsing-remitting or progressive MS); intent of use (primary / key secondary endpoint at various stages including proof of concept, dose-ranging, confirmatory and registration trials); criteria for selection of the 4 domains (relevant, objectively quantifiable, favourable psychometric properties, reflecting functional change, pragmatic); and gave a justification of the dimensions chosen (ambulation, upper arm functioning , visual function and information processing speed as measured by T25FW, 9HPT, LCLA and SDMT respectively).

The rationale behind the development of the COA instrument is clear and considered justified by the arguments given above. In this follow-up advice the concept of the early qualification procedure depicted in the figure below was further worked out.



Source EMA/CHMP/SAWP/232298/2014, Figure 3: Framework for developing a COA Performance Measure for MS clinical trials

The attractiveness of the performance tests chosen i.e. T25FW, HPT, LCLA and SDMT lies in their objectivity, reproducibility, reliability and sensitivity to change. These characteristics are considered established based on the literature review and the analysis of aggregated clinical trial data and are not at discussion.

In the previous qualification advice it was agreed that that the presented domains are among the important disability domains in MS (EMEA/H/SAB/040/1/FU/1/QA/2014, see figure above).

Further, the domain of cognition was broader and did not only include pace of thought (SDMT) but also memory (California verbal learning test [CVRT]; Benton visual retention test [BVRT]; 7/24 Spatial Recall Test [SRT]) and attention. The focus on speed of processing as cognition parameter needs to be more extensively justified.

However, disability refers to the inability to execute activities, less involvement in life situation and ability in performing social roles. The T25FW, 9HPT, LCLA and SDMT are sole performance tests. How changes in test performance translate to an effect on daily functioning and/or disability remains unclear. In other words whether the connections between the yellow boxes drawn in the figure above are substantiated by data is not clear from the above. Change in speed (T25FW, 9HPT) or scores (LCLA, SDMT) of the performance tests cannot be accepted to reflect disability at face value. Hence, whether these tests reflect the concept of interest can only be determined when the connections mentioned are further substantiated.

The most common assessment scale used to evaluate disability in multiple sclerosis so far is the EDSS. It is acknowledged that the EDSS dominantly focusses on motor function and within that ambulation. Therefore, convergent validity of the LCLA and SDMT performance to the EDSS score is not expected. Thus convergent validity of LCLA and SDMT has to be established by relating the LCLA and SDMT to other scales measuring the impact of visual and cognitive function on ALD. Also here the Voice of the Patient study is important.

As noted the T25FW, 9HPT, LCLA, SDMT tests do not incorporate fatigue, pain, sexual dysfunction, sensory outcomes. The result of the second project in the Voice of the Patient study confirms that fatigue (90.3%), incoordination (88.7%) and spasticity (75.6%) are severe problems in multiple

sclerosis impacting overlapping levels of ADL. These impairments are also considered important by the consortium but thought to be better covered by PRO measures. However, this raised the question if the concept of interest i.e. "disability in multiple sclerosis" is fully covered by the 4 dimensions selected. This begs the question whether a general questionnaire e.g. Multiple Sclerosis Impact Scale - 29 items (MSIS) incorporating all these dimensions is not an alternative way forward although it is acknowledged that PRO may be less objective and more subject to variability.

Initially (2014), the MSOAC proposed a global disability score that would be built as a weighted sum of these its 4 components. The decision has been made not to pursue a global disability score is a change of concept i.e. the four measures are now considered as a battery of tests, all or some of could be used in a clinical trial as the primary endpoint. This seems to drift away from the concept of interest.

Context of use

The consortium defines the intended context of use for the COA instrument to serve as a primary, co-primary or key secondary endpoint to assess efficacy in clinical trials at various stages, including proof of concept, dose-ranging, confirmatory and registration trials. The four performance measures are considered as a battery of tests, some or all of which could be used by sponsors in a clinical trial. If used in registration trials, the ultimate language included in product labelling will reflect which measures were used in the trials and would describe the effect of treatment on each measure. Several constructions of endpoints could be envisioned and the clinical interpretation of results if the construct would be used as primary efficacy endpoint would markedly differ. The single measures could be used as single endpoints, as co-primary endpoints (also in combination with established instruments as e.g. EDSS), or as a composite in a time-to-event analysis (with AND and OR conditions, e. g. time to worsening of T25FW AND 9HPT or time to worsening of T25W OR EDSS), and the list of options is by far not exhaustive.

Methods

Three approaches were used to substantiate the relation between the test performances and functional impact i.e. 1) A review of the literature, 2) A voice of the Patient study and 3) an integrated analysis of aggregated clinical trial data.

1. Ad review of the literature

The value of the literature review is limited as the data dominantly concern cross-sectional data. The integrated analysis of aggregated clinical trial data is based on a large number of subjects (n=12776). Therefore it is not unexpected that small to modest correlations between different assessment scales used are statistical significant.

In the Voice of the Patient study an attempt was made to relate the test performances of the T25FW, 9HPT, LCLA and SDMT to an impact on ADL (Patient-Related Level of Inference in Daily activities).

2. Ad voice of the patient study

In this Voice of the Patient study, patients with Multiple Sclerosis first executed each performance measure and then were presented with examples of everyday activities related to that performance measure and asked to indicate the impact of their impairment on those daily activities. Participants ranked interference with activities of daily living (ADLs) in close alignment with their performance test results in the physical domains of mobility (T25FW) and upper extremity function (9HPT). Thus basically, even though the data are not longitudinal, this is the major study were the hypothesised linkage above can be substantiated.

There were two projects incorporated in the VOPS study.

In the first step of Project 1 each participant was evaluated using each of the 4 performance measures in random order. In a second step, immediately following administration of each of the four performance measures, participants were presented with 5 examples of everyday activities that have increasing levels of difficulty to judge interference with their ADLs. In the third step the extent to each of the 4 performance measures relate to perceived interference with related ADLs were estimated.

Rank order sets of 5 daily activities for each performance measure based on increasing levels of difficulty were created. Each set of 5 items was calibrated using the related domains of the Neuro-QoL assessment platform. As there is no Neuro-QoL domain for vision, the rank-ordered set of five activities was created based on the literature review as described above.

The aim of Project 2 was to obtain the perspectives of PwMS concerning the impact on their ADLs of impairment in areas of functioning not necessarily measured by the four candidate measures. Five candidate symptoms were evaluated and all five were endorsed by majorities of PwMS. The most frequently cited symptom was fatigue (90.3%) followed by incoordination (88.7%), spasticity (75.6%), bladder dysfunction (69.4%) and pain (54.8%). In addition, participants also added paresthesias, sensory loss, and weakness. Paresthesia was the most frequently identified unprompted symptom (22.6%). Other symptoms mentioned had lower frequencies, ranging from 16.1% to 3.2%.

In the voice of patient study for each dimension (ambulation, arm functioning vision, cognition) a limited number of functional questions were rated by the patient. For instance in the Mobility ADL there are 5 simple questions rated by patient on a 1-10 point severity scale. As an example 1 of the 5 questions concerns difficulties getting up from the floor. Point is that questions used in the Voice of Patient study and overall score are not anchored i.e. not fully validated e.g. against the MSIS.

Unfortunately the EDSS was not measured as part of this study. Since patients were required to be ambulatory it is assumed that the EDSS had it been measured would be ≤6.5. The lack of these data and the lack of inclusion of patients with higher EDSS disability scores is a missed opportunity to estimate the added value of the meaningfulness of the T25FW, 9HPT, LCLA and the SDMT to PwMS in comparison to the EDSS and the meaningfulness of 9HPT, LCLA and the SDMT to patients with higher disability (i.e. non-ambulatory PwMS).

3. Ad integrated analysis of aggregated clinical trial data.

Performance tests

Timed 25 foot walk (T25FW)

T25FW-literature review

The T25W is objective, reliable, and sensitive to change and reflects functional changes. A 20% change in T25FW performance is related to clinical relevant changes in the Physical Component Summary (PCS) score of the SF-36.

Based on the review of the literature the T25FW test correlated to the EDSS score.

The T25FW has been used as primary end-point for clinical research targeting ambulation in MS. However, the T25FW is not universally applicable across the MS spectrum, as ceiling effects are anticipated for those with an EDSS above 6.5 (i.e., not able to walk 25 feet). Apart one dimension of walking is measured, i.e. gait speed. Other dimensions are also important as balance.

However, in the ENHANCE study with Fampyra the clinical meaningfulness of the T25FW was confirmed by consistent effects in the Multiple Sclerosis Walking Scale (MSWS-12 scale) which is a patient reported outcome measure. The results of the literature review indicate that that a 20% change in

T25FW performance represents a meaningful change in walking performance in MS. Also in the EPAR of Fampyra (EMA/55566172011) it is mentioned that the re-examination SAG Neurology considered the 20% improvement based on walking speed to be of potential relevance, if correlated to patient-reported outcome measures.

However, the context of use of the T25FW was symptomatic treatment not for assessing disability.

T25FW- voice of patient

Based on the Voice of Patient study test performance of the T25FW was reasonably correlated to the Patient-Rated Level of Interference in Daily Activities-mobility score. Regression analyses indicated that participants with a T25FW score <8.5 seconds, each one second increase in T25FW score resulted in a 1 to 2-point interference score increase in each of the 5 individual mobility ADLs and a 7-point increase in summed mobility interference score ($p<.001$). Among those participants scoring > 8.5 seconds on the T25FW, a 1 second increase in the T25FW score resulted in a 0.1 to 0.2 score increase in each of the 5 mobility interference scores and a 0.7 increase in summed mobility interference score ($p = 0.02 - 0.28$).

T25FW- Aggregated clinical data

Based on the analysis of aggregated clinical trial data there is no concordance in agreement between Disability Worsening as defined by EDSS and worsening as defined by T25FW at baseline and end of the study as the Kappa coefficient is 0.02) Further whereas the correlation between the absolute values of the T25FW and absolute EDDS values is relatively high (0.39-0.62) the correlation between the change in T25FW and change in EDSS was only around 0.25. This is unexpected considering that in the paper Bosma et al. (2012) it was shown that early changes in EDSS and T25FW are independently good predictors of long term EDSS (3 years). This is what would be expected as the two scales focusing on ambulation. It sets some doubt on the reliability of the aggregated clinical trial data analyses.

T25FW- summary

Nevertheless considering the literature and Voice of Patient study, the connection between T25FW test performance and functionality may be considered reasonably established. Main weight in this assessment is given by the "Voice of the Patient" and correlation of the T25FW with the related to clinical relevant changes in the Physical Component Summary (PCS) score of the SF-36. Thus the connection between T25FW test performance and ADL (see figure above) is considered established.

Nine hole peg test (9HPT)

9HPT-literature review

The 9HPT is objective, reliable and sensitive to change and reflects functional changes.

It is agreed that the 9HPT detects progression over time, is responsive to treatment, is reliable within and between test sessions, discriminates between healthy subjects and MS patients with different levels of upper limb impairment, and shows high convergent validity with other manual dexterity as well as more comprehensive upper limb measures. Ecological validity is indicated by its relation to perceived upper limb use in daily life and perceived difficulty in performing activities of daily living.

Based on the literature review a 15%-20% change in 9HPT performance is claimed to be related to clinical meaningful changes of the Guys Neurological rating test, MS Impact scale score and EDSS. However, that a 15-20% difference in 9HPT is clinically relevant has not been convincing demonstrated as the information in the literature review is anecdotal. Quantitative data that relates a change in 9HPT test performance to a change in for instance MSIS-score are not presented.

9HPT-Voice of patient

Based on the Voice of Patient study test performance of the 9HPT was reasonably correlated to the Patient-Rated Level of Interference in Daily Activities-mobility score . The regression for the upper extremity domain demonstrated that for every ten seconds increase in 9HPT score, there was an associated increase of only 1 to 2 points of interference in each of the 5 upper extremity interference scores and an 8-point increase in summed score.

9HPT- aggregated clinical trial

Based on the analysis of aggregated clinical trial data there is no concordance in agreement between Disability Worsening at Endpoint as defined by EDSS and as defined by 9HPT as the Kappa coefficient is 0.01. Further whereas there is a rather modest correlation between the absolute values of the 9HPT and absolute EDDS values (0.37-0.59), the correlation between the change in 9HPT and in change in EDSS was only around 0.20. Also the correlation between 9HPT test performance and Physical Component Summary (PCS) score of the SF-36 was low (<0.20).

9HPT- Summary

Nevertheless considering the literature and Voice of Patient study, for the 9HPT the connection between 9HPT test performance and functionality may be considered reasonably established although to a lesser extent as compared to the T25FW. Again main weight in this assessment is given by the "Voice of the Patient" study. Thus the connection between 9HPT test performance and ADL (see figure above) may be considered established.

Low contrast letter acuity (LCLA)

The LCLA assessment is objective, reliable and sensitive to change. However, the impact of changes in LCLA on ADL is less clear.

LCLA literature review

Visual dysfunction is one of the most common manifestations of MS. Loss of low-contrast vision has been shown to be an important contributor to impairment and disability in MS and seems to captures visual loss not seen in high-contrast visual acuity (HCVA) measurements.

Based on the literature review LCLA performance correlates to the Retinal Neuronal and T2 lesion volume. The change in LCLA from baseline to 1 year was a predictor of change in EDSS score between year 1 and 2 in the IMPACT study. Reduction in LCLA reflect worse scores in vision specific QOL e.g. NEI-VFQ-25 and IVIS. Finally LCLA performance is responsive to therapy.

LCLA voice of patient

However, in the Voice of Patient Study the correlation between Visual function and 2.5% LCLA score was weak to modest. A linear relationship could not be established.

LCLA aggregated clinical trial data

In the analysis of aggregated clinical trial data there was limited concordance in agreement between Disability Worsening at Endpoint as defined by EDSS and worsening as defined by LCLA (Kappa coefficient around 0.10). Correlation between LCLA and the physical component of the SF-36 is more than weak i.e. 0.02-0.04).

LCLA summary

Thus the connection between LCLA and ADL/function as suggested by the literature review, was not reflected in the results of the Voice of Patient study and aggregated data analysis. Considering this all for the LCLA the connection between LCLA and functionality is not considered established.

Symbol digit modalities test (SDMT)

SDMT literature review

From the literature review it appears that the SDMT score is objective, reliable and sensitive to change. It is claimed to measure information processing speed. The SDMT is a strong predictor of central atrophy.

While it is agreed that the SDMT (and also the PASAT) is a valid measure of processing speed, the justification that processing speed is the most important cognitive domain affected in MS patients and correlation with memory and higher cognitive functions is rather weak. This is already acknowledged by the Consortium. A second potential weakness is that SDMT does not have similar face validity as tests of motor function, in the sense that SDMT test results may not seem on the surface to be directly related to common activities of daily living. Moreover, SDMT performance can be influenced e.g. by visual acuity and ocular motor functions and there are learning effects (Benedict 2017).

SDMT voice of patient

Based on the Voice of Patient Study the correlation between Cognitive Functioning and SDMT score was weak to modest. A linear relationship between SDMT and patient related level of interference in daily activities could not be established.

SDMT aggregated clinical trial data

Based on the analysis of aggregated clinical trial data there is no concordance in agreement between Disability Worsening at Endpoint as defined by EDSS and worsening as defined by SDMT (Kappa coefficient around 0. Correlation between the absolute values of the SDMT and absolute EDDS values was modest at best 0.34 (table 38 p 125/205 of the briefing document). However, the correlation between change in SDMT and change in EDSs was less i.e. 0.12. This is not unexpected as the correlation between EDSS and SDMT a priori is remote as the EDDS has no cognitive dimension. More important is the modest correlation between SDMT and the mental component of the SF-36.

SDMT summary

Thus the connection between SDMT and ADL/function as suggested by the literature review was not reflected in the results of the Voice of Patient study and aggregated data analysis. Considering this all for the SDMT the connection between SDMT and functionality is not considered established.

<u>Summary overall discussion</u>

The concept of interest measuring disability in progressive MS is clear and not at discussion.

The intended context of use has however changed from a global disability score that would be built as a weighted sum of these 4 components into four separate measures that can be used in combination or on a single primary endpoint in support of a descriptive indication e.g. Product A demonstrated to delay the accumulation of disability in information processing as measured by of Symbol Digital Modalities Test. This seems to drift away from the original concept of interest.

The T25FW, 9HPT, LCLA, SDMT tests do not incorporate fatigue, pain, sexual dysfunction and sensory outcomes. These impairments are also considered important by the consortium but thought to be

better covered by PRO measures. However, this raised the question if the concept of interest i.e. "disability in multiple sclerosis" or impact on ADL is fully covered.

The relationship between (changes in) test performance in T25FW and 9HPT test and impact on daily functioning is considered reasonably established based on the data submitted. This is more based on the literature review and to some extent the Voice of Patient study than on the Aggregated Data Analysis of clinical studies. The almost absence of concordance in agreement of worsening of EDSS and worsening on the T25FW or 9HPT in the aggregated data analysis is unexpected and sets doubts on the reliability of the aggregated clinical trial data analyses. It is speculated that the aggregated clinical trial data set is more heterogeneous than expected. The relationship between (changes in) test performance in the LCLA and SDMT and impact on daily functioning is not considered established based on the data submitted. Convergent validity was not shown in the Voice of Patient study or aggregated data analysis of the clinical studies. Convergent validity with the EDSS is not expected. However, a relationship between these test performances and patient related level of inference in daily activity was not shown. Similar there was a more than modest correlation with the SF-36 in the aggregated data analysis.

So far there is limited experience with the SDMT as endpoint in clinical studies in MS. Speed of information processing is important for cognitive function but whether it covers cognitive function in multiple sclerosis is not made clear. The quality of cognitive processing e.g. executive functioning is not assessed. Whereas inclusion of cognitive impairment scales as endpoint in MS trials is generally endorsed the usefulness/validity/relevance of the SDMT as representative measure for cognitive function is still at discussion.

Apart from that, literature data (Borghi et al., Front Hum Neurosci 2016) suggest differences in cognitive scoring as assessed by PASAT for patients affected with different courses of the disease (SPMS vs. RRMS). The transferability to later stage MS needs to be justified since only data in early stage patients are available for the SDMT. Only one randomized double- blind controlled study was analysed (ADVANCE) that contained data on both the SDMT and the PASAT.

Moreover, there are learning effects and the SDMT performance can be influenced e.g. by visual acuity and ocular motor functions (Benedict 2017). Apart from that the type and degree of cognitive impairment in MS is highly dependent on the location of the lesions.

The correlation with perceived interference in daily activities related to SDMT was relatively weak (-0.21). A number of reasons are discussed (less perception of vision and cognition impairment to affect ADLs by patients depending on profession and education, compensation methods). It is argued that the relative lack of alignment should not lead to the interpretation that the measure is not clinically meaningful, but that PwMS do not relate scores on an unfamiliar test to interference with their ADLs related to those disease dimensions. It needs however also to be justified, that this lack of alignment is not an issue in clinical trials. The same lack of correlation was also observed for the LCLA (-0.28).

There are practice effects that may hamper the use of the SDMT in clinical trials. Performance characteristics of the SDMT were only discussed in relation to the PASAT and not in relation to other cognitive scales. Correlation of SDMT with PASAT was moderate to strong at baseline and endpoint, which is expected. However, changes from baseline at the endpoint for both measures were only weakly correlated. This is rather unexpected and suggests that these two measures do not detect the same characteristics of longitudinal changes of abilities (Brochet et al., Multiple Sclerosis 2008). This finding needs further discussion.

The relationship of performance measure scores to relapse or EDSS for the SDMT did not always go into the expected direction. Although there was a worsening of SDMT on relapse, it was not a statistically significant change. Actually the data also indicate that 51% of the patients showed an

improvement on relapse (Table 58). On recovery from relapse, there was a statistically significant improvement. Although worsening was expected on EDSS worsening, in fact the SDMT score showed a statistically significant improvement. A greater improvement was seen on EDSS improvement, which was as expected. This is unexpected and difficult to interpret.

All validation work was done retrospectively with exception of the VOPS study. Unfortunately the EDSS or MSIS was not measured as part of this study. Since patients were required to be ambulatory it is assumed that the EDSS had it been measured would be ≤6.5. The lack of these data and the lack of inclusion of patients with higher EDSS disability scores is a missed opportunity to estimate the added value of the meaningfulness of the T25FW, 9HPT, LCLA and the SDMT to PwMS in comparison to the EDSS and the meaningfulness of 9HPT, LCLA and the SDMT to patients with higher disability (i.e. non-ambulatory PwMS).

The intended context of use has changed from a global disability score into four separate measures that can be used in combination or as a single primary endpoint in support of delay the accumulation of disability. This makes it even more prudent to firmly establish the relationship of a single dimension performance and ADL/Disability.

The attractiveness of the performance tests chosen lies in their objectivity, reproducibility, reliability sensitivity to change and that they are easy to perform. They lack the subjectivity of a PRO (e.g. MSIS). However, the assessment of what an effect means in terms of clinical significance is an ongoing discussion. Prospective studies that firmly establish this connection are limited.

The consortium defines the intended context of use for the COA instrument to serve as a primary, co-primary or key secondary endpoint to assess efficacy in clinical trials at various stages, including proof of concept, dose-ranging, confirmatory and registration trials. The four performance measures are considered as a battery of tests, some or all of which could be used by sponsors in a clinical trial. If used in registration trials, the ultimate language included in product labelling will reflect which measures were used in the trials and would describe the effect of treatment on each measure. It remains however unclear how exactly this "toolbox" is intended to be used. Several constructions of endpoints could be envisioned and the clinical interpretation of results if the construct would be used as primary efficacy endpoint would markedly differ. The single measures could be used as single endpoints, as co-primary endpoints (also in combination with established instruments as e.g. EDSS), or as a composite in a time-to-event analysis (with AND and OR conditions, e. g. time to worsening of T25FW AND 9HPT or time to worsening of T25W OR EDSS), and the list of options is by far not exhaustive. However, the relationship of these test performances either as single test or in different combinations to functioning (e.g. MSIS, MSWS-12, PRO-developed) and thus the interpretation of the clinical relevance of the test performances remains to be established. This precludes accepting these tests as primary endpoint in support of an effect on disability. It is considered unlikely that a claim like ""delay of disability as measured by arm dexterity as measured by the 9 Hole Peg Test" is sufficient to support a disability claim.

CHMP overall conclusion

While the validation work is acknowledged, the Timed 25-foot walk (T25FW), hand dexterity (9 Hole peg Test, 9HPT), visual function (Low contrast Letter acuity, LCLA) and mental tests assessing processing speed (Symbol Digit Modalities Test, SDMT) can neither be used as single variable or in combination with each other as primary endpoint for measurement of disability without including functional scales as well in the primary endpoint. The T25FW and 9HPT could be included in a composite primary endpoint combined with a functional endpoint (e.g. EDSS). All components should contribute to the overall effect and the overall effect should not be predominantly driven by the performance tests in the composite. It is highly advisable to included simultaneously another

functional endpoint in order to relate the effect size on the composite to a clinically relevant change (e.g. MSIS) within the same study. Further it is considered that subjects, after meeting the composite event, should be followed up for all the components of the composite endpoint. The inclusion of these tests in clinical studies as secondary endpoints in comparison to functional scales is accepted.

Background information as submitted by the Applicant

New Clinical Outcome Assessment Instrument to Assess Disability for Use in Clinical Trials of Medicinal Products to Treat Multiple Sclerosis (MS)

Multiple Sclerosis Outcome Assessments Consortium (MSOAC)
Summary as Submitted by the Applicant for Public Comment

Evidence to Support Qualification

MSOAC executed three approaches to establishing the optimal performance measures for inclusion in a COA instrument to assess treatment benefit: 1) review of the literature; 2) incorporation of the patient voice; and 3) analysis of aggregated MS clinical trial data. These datasets represent a comprehensive, in-depth analysis of information on MS clinical endpoints over 5 years. The work performed by MSOAC to demonstrate that these measures are clinically valid, highly reliable, practical, cost-effective, and meaningful to persons with MS (PwMS) included representatives from advocacy organizations, the National Institute of Neurological Disorders and Stroke (NINDS), academic institutions, regulators and industry partners along with persons living with MS, all collaborating to develop improved and meaningful measures of MS – related disability.[1]

I. Review of the literature revealed the significance of the different domains for PwMS, the ways in which each domain can be measured, the psychometric properties of extant performance measures for each domain, and the suitability of each performance measure as a clinical trial endpoint. To assess the extant literature, MSOAC members first formulated a set of questions to be addressed and selected relevant domains from the International Classification of Functioning, Disability, and Health (ICF) core and comprehensive set: ambulation, arm dexterity, vision, and cognition. Based on reliability, construct and predictive validity, discriminative validity, criterion validity, and clinical relevance, four measures were selected for inclusion in the outcome assessment instrument: T25FW as a measure of walking speed, the 9HPT as an upper extremity dexterity measure, the LCLA as a vision measure, and the SDMT as a measure of information processing speed. The features of these four measures identified by the literature review include the following:

- The **T25FW** is conducted by timing the PwMS as he/she walks for 25 feet (7.6 m). Walking is one of the most important and valued functions for PwMS, and the T25FW is the best characterized measure of walking disability in MS. The T25FW is capable of detecting changes in walking and worsening of at least 20%, and this degree of change is clinically meaningful.
- The **9HPT** requires PwMS to repeatedly place and then remove nine pegs in nine holes, one at a time, as quickly as possible. Of the upper limb outcome measures applied in MS studies, the 9HPT is considered the gold standard metric for manual dexterity. The 9HPT detects progression over time, is sensitive to treatment, is reliable within and between test sessions, discriminates between different levels of upper limb impairment, and relates to perceived difficulties in performing activities of daily living. A 20% change in the 9HPT test score is commonly used to define clinically-meaningful worsening.
- The **LCLA** is conducted by asking PwMS to read aloud the letters on the Sloan Low Contrast Letter Acuity chart until they can no longer see the letters. The LCLA has excellent test-retest reliability and is the most sensitive test for visual function in MS. A loss of 7 letters on the LCLA is considered to be clinically meaningful.
- The **SDMT** presents a key, consisting of nine abstract symbols. Each symbol is paired with a number ranging from 1 to 9. The test consists of 120 abstract symbols presented in random order. PwMS are asked to associate the symbols with the correct corresponding number, as shown in the key. PwMS respond orally as quickly as possible and the number of correct responses is recorded. Processing speed is a basic, elemental cognitive function. A systematic review of the literature revealed one cognitive measure, the SDMT, as being particularly sensitive to the slowed processing of information that is commonly seen in MS. Published evidence supports the reliability and validity of this test, its relevance to daily activities, and recently has supported a responder definition of a change in the SDMT score as approximately 4 points or 10% in magnitude.

II. **Voice of the Patient** In addition to conducting an extensive literature review, MSOAC members designed and carried out an innovative approach to incorporate the "Voice of the Patient" (VOP) from PwMS recruited from the Mellen Center at the Cleveland Clinic Foundation. In the VOP study, PwMS first executed each performance measure listed above, and were then presented with examples of everyday activities related to that performance measure and asked to indicate the impact of their impairment on those daily activities. This combination of direct reports with evaluation of actual performance was useful in assessing the clinical relevance of candidate performance measures. Participants ranked interference with activities of daily living (ADLs) in close alignment with their performance test results in the physical domains of mobility (T25FW) and upper extremity function (9HPT). For non-motor performance measures (LCLA and SDMT), participants did not rank interference with ADLs related to vision or cognition as closely correlated with their scores on the LCLA and SDMT, respectively. Input on symptoms not related to these four performance measures was also collected. The most frequently cited symptom was fatigue, which is the focus of the PRO instrument under development by the Critical Path Institute's PRO Consortium.

III. **Analysis of Clinical Trial Data** In order to analyze available data on performance measures in trials of MS therapies, MSOAC member organizations contributed patient-level treatment and control arm clinical data from approximately 12,766 participants from 14 trials. A Clinical Data Interchange Standard Consortium (CDISC) data standard for MS was developed, published, and applied, in order to integrate the data from different sources (http://www.cdisc.org/therapeutic#MS). The MSOAC database is a very large and rich source of data on functional assessments, including performance measures such as the T25FW, 9HPT, LCLA, and SDMT. Also, a well-validated, standardized patient reported functional outcome measure, the Short Form Health Survey (SF-36), was used in many trials, allowing performance measures to be correlated. The following attributes of each performance measure were assessed: floor and ceiling effects, test-retest reliability, change over time, construct validity, convergent validity, extent of practice effects, known group validity, sensitivity to change, treatment effects, and the minimum clinically important change scores. Analysis of the pooled data revealed 1) a normal distribution of LCLA and SDMT scores, with no evidence of floor or ceiling effects; positively skewed frequency distributions of T25FW and the 9HPT, yet both measures distinguish gradations of performance in the middle of the scale; 2) modest practice effects of the SDMT and no evidence of practice effects for the T25FW, 9HPT, and LCLA; 3) excellent test-retest reliability for all four measures; 4) support for the construct validity of all four measures; no significant correlation of SDMT scores with depression, suggesting that SDMT scores are less likely to be confounded by emotional factors; 5) meaningful relationships of all four measures with duration and severity of MS; 6) significant correlations between worsening on T25FW, 9HPT, and SDMT and meaningful worsening on the Physical Component Summary from the SF-36 patient report, which is strong evidence for clinical meaningfulness.

Drawing on the wealth of published studies and accumulated clinical trial data together with a proactive MS community, MSOAC assessed the literature, directly engaged PwMS to assess patient preferences, and analyzed available data to determine the suitability of individual performance measures of MS disability. It is the totality of the evidence from these three approaches that demonstrate the excellent psychometric properties and clinical meaningfulness of the performance measures of ambulation, dexterity, vision, and cognition: T25FW, 9HPT, LCLA, and SDMT, respectively. As a measure of ambulation, the T25FW is capable of detecting clinically meaningful changes in walking and worsening of at least 20%. As a measure of dexterity, the 9HPT is capable of detecting clinically meaningful changes of at least 20%. As a measure of vision, the LCLA is capable of detecting a clinically meaningful change of 7 letters. Finally, as a measure of the processing speed of the cognition domain, the SDMT is capable of detecting a change of approximately 4 points or 10% in magnitude. Altogether, the MSOAC analyses provide evidence indicating that a change in one or more of these performance measures constitutes a clinically meaningful change. The MSOAC Members conclude that the combined data obtained for each performance measure from three robust sources of data – the comprehensive literature review, the VOP study, and analyses conducted on the large MSOAC clinical trial database – provide a strong preponderance of evidence for the use of these 4 measures as acceptable primary disability outcome measures for MS clinical trials.

Background on the Disease

MS is a chronic disorder characterized by central nervous system (CNS) inflammation with associated damage to neurons, axons, and myelin. The most important pathologies in MS are 1) inflammation, which leads to new lesions in the brain that are detectable by magnetic resonance imaging (MRI) and neurological symptoms (relapse) if the pathology hits certain pathways; and 2) neurodegeneration, which leads to brain scarring, shrinking, and permanent disability. According to an international consensus panel, the diagnosis of MS is based on typical clinical manifestations supplemented by MRI findings in an individual without an alternative diagnosis explaining the illness. Three common subtypes of MS, based on disease course, have been described by the MS Phenotype Group under the auspices of the International Advisory Committee on Clinical Trials in MS (supported by the European Committee for Treatment and Research in Multiple Sclerosis [ECTRIMS] and the National Multiple Sclerosis Society [NMSS]).2 These subtypes are relapsing remitting MS, secondary progressive MS, and primary progressive MS. Most people begin with an RRMS course, which is characterized by new or worsening neurological symptoms, lasting days to weeks (relapse), followed by full or partial recovery (remission), followed by subsequent relapses occurring unpredictably over the ensuing years. Relapses are separated by periods of neurologic stability. Over time, people with RRMS may evolve to a SPMS course, characterized by continued, gradual worsening of disability with or without superimposed relapses. By definition, SPMS always occurs following RRMS. People with MS in whom disability progresses from the onset of disease are diagnosed with PPMS. The biological differences between SPMS and PPMS, if any, are a matter of debate in the scientific community. An additional subtype of MS was included in earlier clinical course definitions: progressive relapsing MS (PRMS).3 People with MS displaying gradual progression from onset with subsequent superimposed relapses were considered to have PRMS. This subtype was eliminated by the MS Phenotype Group,2 because subjects categorized in the past as PRMS would now be classified as PPMS with relapses as evidence for disease activity.

Background on the Performance Measures

The COA instrument that MSOAC is proposing for qualification is intended to reflect the impact of an intervention on disease worsening as it relates to disability due to MS. The focus of the COA is on the major aspects of function that are objectively quantifiable, relate closely to the MS pathological process, and are thought to relate to important aspects of daily life for MS patients, such as walking, manual dexterity, vision, and cognition. The qualified COA instrument could be used as a primary outcome measure in clinical trials, including registration trials. Its use could be in conjunction with a range of other performance measures as well as other secondary outcome measures, including imaging, relapse assessment, and PRO instruments including measures of Quality of Life (QoL).

Inclusion of a cognitive performance measure in particular is considered critical for future MS clinical trials, because cognitive impairment is well established as an important contributor to overall impact of MS on the individual, and it has not been adequately assessed in most prior registration trials. Cognitive decline occurs in some individuals in the absence of worsening physical or visual disability, and it has been identified by regulatory agencies as a critical unmet need in disability assessment for MS trials.

AMBULATION: Walking is one of the most valued functions for PwMS and walking dysfunction represents a primary burdensome feature of MS for quality of life.[4–6] The T25FW test is considered the best characterized objective measure of walking disability in MS[7] and served as the primary endpoint for improved walking in the Phase III trial of dalfampridine.[8,9]

Timed 25 Foot Walk (T25FW)

a. Reporter, if applicable: The administrator/reporter of the measure manually records the time elapsed between the start and completion of the 25-foot walk.
b. Item content or description of measure: The subject is instructed to walk as fast and safely as possible across a clearly marked, linear 25-foot (or 7.62 meter) course.
c. Mode of administration: (detailed in the Multiple Sclerosis Functional Composite [MSFC] manual released by the National MS Society's Clinical Outcomes Assessment Task Force, available from www.nationalmssociety.org.
d. Data collection method: The time in seconds is recorded when the subject lifts one foot for starting the T25FW and ends upon breaking the plane of the end point with a foot. The test is

performed twice, and time in seconds is averaged between trials. The score is expressed by dividing 25 feet by the time in seconds, which represents walking velocity.

MANUAL DEXTERITY: Impaired hand function is one of the most frequently reported symptoms in the first year of MS.[10] The 9HPT is widely considered a gold standard metric for manual dexterity and is the most frequently used measure in MS rehabilitation.[11]

9 Hole Peg Test (9HPT)

a. Reporter, if applicable: The administrator/reporter of the measure manually records the time required to complete the task.
b. Item content or description of measure: Subjects repeatedly place and then remove nine pegs into nine holes, one at a time, as quickly as possible
c. Mode of administration: (detailed in the MSFC manual released by the National MS Society's Clinical Outcomes Assessment Task Force, available from www.nationalmssociety.org.
d. Data collection method: Subjects' responses are recorded during the task. The number pegs placed per second is calculated within the time limit of 300 seconds. Four trials are conducted (two trials for each hand).

VISION: Visual dysfunction is one of the most common manifestations of MS and consequently sensitive visual outcome measures are important in assessing treatment benefit. LCLA charts capture visual loss that is undetectable when using high-contrast visual acuity (HCVA) charts. Of the LCLA measures, the Sloan charts perform better than the other chart.

Low Contrast Letter Acuity (LCLA)

a. Reporter, if applicable: The administrator/reporter of the measure manually records the examinee's verbal response on a separate scoring sheet throughout the administration of the test.
b. Item content or description of measure: Subjects read aloud the letters on the Sloan letter chart.
c. Mode of administration: Sloan charts at 2.5% and 1.25% contrast are administered binocularly or each eye can be tested individually.
d. Data collection method: Subjects' responses are recorded during the task. The score for each chart is quantified as the number of letters identified correctly with a maximum score of 70 letters.

COGNITION: While there are several subdomains of cognition (e.g., memory, executive function), information processing speed is the cognition subdomain that is the focus of this COA measure qualification. Processing speed is a basic cognitive function, and deficits in information processing speed explain a significant proportion of the variance in the limitation in activities, participation, or roles that are understood to be important by persons with MS. Cognitive impairment and deficient information processing speed, is a common, often early manifestation of MS. Cognitive impairment has an established negative impact on how persons with MS feel, function, or participate in their societal and family roles.

This domain has been most commonly measured using the SDMT or the Paced Auditory Serial Addition Test (PASAT), two well-established cognitive performance measures used in the MS field. Based on the superior measurement properties of the SDMT, MSOAC seeks qualification of the SDMT as one of the performance measures for inclusion in the COA, but also presents analyses of PASAT data from the MSOAC database studies where SDMT was not administered, to further support the importance of information processing speed in MS patients and to compare the performance of the two information processing measures.

Symbol Digit Modalities Test (SDMT)

a. Reporter, if applicable: The administrator/reporter of the measure manually records the examinee's verbal response on a separate scoring sheet throughout the administration of the test.
b. Item content or description of measure: The SDMT is a measure of information processing speed. Participants are provided an 8 ½ x 11-inch sheet of paper consisting of nine unique symbols, each paired with a number (single digits 1-9) on top of the page (testing key). The remainder of the page presents a pseudo-randomized sequence of 120 of these symbols with empty boxes underneath. The first 10 symbols are used for the learning phase. Patients are asked to respond orally with the number that corresponds with each symbol as rapidly as

possible, without skipping any. The dependent variable is the total number correct in 90 seconds.

c. Mode of administration: The SDMT is owned by Western Psychological Services, and is administered according to the company instructions. The instructions to the participant for verbal administration of the test include directions to "tell me the number" rather than "fill in the number" as for the written version of the SDMT. The subject is timed to determine how many correct responses can be made in a 90 second period.

d. Data collection method: Subjects' responses are recorded during the task. The number of total correct responses is calculated and serves as the primary data point. Scores range from 0 – 110.

In summary, the goal of this qualification proposal is to provide data to support a qualified COA instrument to the MS field that is able to detect clinically meaningful changes in aspects of ambulation, dexterity, vision and cognition that are associated with limitations that are caused by MS in activities, participation, or roles and considered important by persons with MS.

CONTEXT-OF-USE (COU) STATEMENT

For regulatory qualification, the following COU components are defined so that the qualified measures are used within the proper context, as supported by the qualification data.

Target Population for Use

The target population is adults with a diagnosis of MS, and a relapsing-remitting (RRMS), secondary progressive (SPMS), or primary progressive (PPMS) clinical course.

Stage of Drug Development for Use

The intent is to use this proposed battery of performance tests as a dysconjugate composite endpoint. One or more of the performance measures could serve as a primary, co-primary or key secondary endpoint to assess efficacy in clinical trials at various stages, including proof of concept, dose-ranging, confirmatory and registration trials. A qualified disability performance measure could be used in clinical trials of MS as a primary outcome measure if the target is disability worsening, or as a co-primary or secondary measure, with other outcome measures – EDSS, relapse rate, etc. RRMS trials are increasingly designed as active comparator trials where poor sensitivity and reliability of EDSS-based endpoints become major obstacles to feasible trial design with respect to disability comparisons.

There is a strong precedent for use of such an endpoint in MS. EDSS also functions as a dysconjugate composite endpoint for RRMS trials. At EDSS levels below 4, disability worsening can occur by worsening on pyramidal, cerebellar, sensory, bowel/bladder, or visual change (or some combination). Thus, EDSS functions in most RRMS patients the same as proposed for the MSOAC test battery but without the ability to determine exactly what systems were meaningfully impacted by therapeutic intervention. This proposed battery of neuroperformance tests clearly supports the nature of MS with multiple areas of disability. It would facilitate conversation with patients about a combination of symptoms and impairments. The power and flexibility of the proposed battery would allow Sponsors to customize trials, tailoring the performance test battery to the population being studied, to better address the disease heterogeneity and unmet medical needs.

Role in Drug Development

The Concept of Interest (COI) for Meaningful Treatment Benefit is "disability in multiple sclerosis", characterized as neurological or neuropsychological impairments that result in limitation in activities, participation, or roles, which are understood to be important by persons with MS.

The performance measures for which qualification is sought are measures of ambulation, manual dexterity, vision and cognition. These components of disability were selected as the focus because they represent common problems for people with MS, reflect the effects of MS disease activity on common functions, are well studied as demonstrated by a significant literature on measuring these dimensions, and these dimensions of disability lend themselves to quantitative assessment of patient performance by clinicians in an office setting. The components of MS disability included here reflect limitations experienced by people with MS that negatively affect their ability to participate in activities or roles important to them[1]. They also reflect disabilities that often require special adaptations (e.g. walking aids, large print books, ramps).

Although other manifestations such as fatigue, pain, sexual dysfunction, sensory symptoms, and bowel dysfunction are important contributors to disability in MS, they are best assessed by PRO measures and thus are out-of-scope for this qualification. Our vision is that the performance measures developed by MSOAC could be complemented by the use of PRO measures in clinical trials. Measures of bladder and vestibular physiology, hearing loss or bulbar dysfunction are quantifiable, but the physiology is complex, and the tests require sophisticated instrumentation and thus would not be practical for widespread inclusion in multicenter clinical trials. For this reason, these were not the focus of the COA instrument developed by MSOAC.

Targeted Labeling or Promotional Claim(s) will reflect which of the four performance measures that were used in the trial produced a significant change in performance. The first example below represents labeling for a clinical trial in which SDMT was used as an endpoint and showed a significant difference between treatment and placebo. The second example illustrates labeling when both the SDMT and 9HPT were used with positive results.

TREATMENT is indicated for the treatment of people with relapsing or progressive MS.

Example 1: TREATMENT was demonstrated to delay the accumulation of disability in information processing speed as measured by the Symbol Digital Modalities Test.

Example 2: TREATMENT was demonstrated to delay the accumulation of disability in information processing and arm dexterity as measured by of Symbol Digital Modalities Test and the 9 Hole Peg Test, respectively.

Applicable Study Settings for Future Clinical Trials

a. Geographic location with language/culture groups

   The four measures have been widely used internationally with multi-language support and validation that will further enhance reliability and ease of administration.

b. Other study setting specifics

   The four measures are used in outpatient settings and are administered by trained healthcare professionals. The examiner uses a scorer form for each on which he/she records the subject's scores.

Impact of a Battery of Performance Measures for Different Functional Domains

Qualification of a battery of measures would allow non-ambulatory PwMS to be eligible for clinical trials based on outcomes not well captured by the EDSS at these levels of disability. These patients are excluded from present trials in which EDSS is the primary endpoint; this lack of access could be addressed by assessing disability in different domains.

PART I Review of the Literature
Methodology

The MSOAC initiative began by defining the concept of interest for meaningful treatment benefit as "MS disability", or simply "disability", characterized as neurological or neuropsychological impairments that result in limitations in activities and restrictions in participation or life roles, caused by MS, that are understood to be important by the person with MS. This work involved identifying disability dimensions common to MS.[1] The domains were selected from the International Classification of Functioning, Disability, and Health (ICF) core sets.

A literature review was then conducted for the four domains that were selected from the ICF core sets: ambulation, arm dexterity, vision, and cognition. The questions to be addressed by the literature review were formulated and the search parameters for the systematic literature review were established[1]. The searches were limited to publications since 1990 and through 2016 and included multiple languages. This time period was deemed by MSOAC subject matter experts (SMEs) to contain all the relevant publications. The literature review was performed in three levels: Level 1 involved the

defining the parameters and search terms; Level 2 applied abstract filtering criteria to review 447 cognition-related publications and 511 non-cognition related (for the ambulation, dexterity, and vision domains); Level 3 involved describing in detail the information from a total of 564 publications in the Data Extraction Table (supplementary material published in LaRocca et al[1]). At the Level 2 stage, SMEs identified a number of key papers that had been missed because key words and abstracts did not always include the performance measure search terms. An enrichment technique to allow the addition of SME-recommended papers was adopted. This combined "enriched search identified approximately 3000 publications that were evaluated.

Literature reviews for each of these domains have been published[12–15] and the results are summarized below. In addition to the T25FW, 9HPT, LCLA, and SDMT, alternate measures used in the four domains were included in the literature search as well as articles that would combine domains in a disability assessment.

Review of T25FW

Walking is defined as advancing or traveling on foot such that there is always one foot on the ground in bipedal locomotion. Walking has historical and clinical underpinnings as well as patient centrality and importance in MS. Walking dysfunction was recognized as a cardinal feature of MS in the earliest historical accounts of the disease,[16] and currently represents a primary construct for monitoring patients with MS in clinical research and practice.[17] Of note, walking is one of the most important and valued functions for patients with MS[4,5], and its dysfunction represents a primary burdensome feature of the disease for quality of life and participation.[5,6] Such observations underscore the importance of walking as an outcome in clinical research and practice involving MS patients.

Walking can be readily measured in MS. The EDSS,[18] which is the most common scale to measure disability in MS, classifies walking or ambulatory dysfunction based on EDSS scores of 4.0 or greater (e.g., able to walk 500 meters versus 300 meters without aid or rest differentiates a 4.0 and 4.5, respectively, on the EDSS). To that end, scores above 4.0 on the EDSS are primarily based on gait dysfunction, particularly scores of 6.0-7.5, and this makes the EDSS and 500-meter walk a poor choice for measuring ambulation in clinical research and practice at earlier stages of MS, and in addition, the EDSS has well-recognized limitations related to reliability and sensitivity.[17,19] The T25FW was the primary endpoint in Phase II and Phase III trials of extended release, oral dalfampridine (4-aminopyridine) for improving walking in MS.[8,9] The T25FW represents a primary outcome for trials of rehabilitation interventions[6] including exercise training.[20]

MSOAC's review of the literature[13] documented that the T25FW has a wide array of desirable measurements characteristics:

- *The T25FW has shown high reliability over both short and long periods and is more reliable than many other measures of walking used in MS.*
- *The content validity of the T25FW is strong since it measures aspects of walking that reflect essential characteristics of walking that are manifest in daily activities such as the need for speed over relatively short periods of time, e.g., getting to the bathroom on time or crossing streets.*
- *The literature review documented many ways in which the T25FW shows strong construct validity. It clearly distinguishes between individuals with MS and healthy controls. Individuals with higher EDSS scores perform more poorly on the T25FW. Individuals with MS who are unemployed do more poorly on the T25FW.*
- *The convergent and discriminant validity of the T25FW was also documented by the literature review. The T25FW shows higher correlations with other measures of walking than it does with measures of other functions such as manual dexterity.*
- *The clinical meaningfulness and relevance of the T25FW was reflected in the literature in more than one way. The T25FW has been shown to reflect improvement following the resolution of relapses as well as interventions such as steroid treatment. In addition, the widely accepted clinical meaningfulness of a 20% change in the T25FW has been documented in terms of the relationship of such a change to clinically meaningful changes in PROs such as the PCS of the SF-36.*

In summary, the T25FW possesses a broad array of desirable characteristics in a brief but sound and meaningful measure of walking that lends itself well to utilization in both small and large, multicenter clinical trials. In addition, the literature provides strong support for the clinical meaningfulness of a 20% change or difference.

Review of 9HPT

Like walking disability, visual problems and cognitive deficits, upper limb dysfunction is a core deficit affecting MS patients. A combination of predominantly motor and sensory symptoms causes upper limb disability, which hampers the ability to perform ADLs and social activities, resulting in a decreased quality of life.[21] Upper limb disability in MS patients may present in the proximal or distal parts of the upper limb. Distal upper limb dysfunction is frequently referred to as impaired manual dexterity or hand dysfunction. Impaired sensory function (85%), fatigue (81%), impaired hand function (60%), and mobility (50%) were the most frequently reported symptoms in the first year of the disease[10] Recently, Bertoni *et al* reported that 75% of their study population (n=110, median Expanded Disability Status Scale, EDSS 6.5) had bilaterally (minimally) impaired manual dexterity as measured with the 9HPT.[22]

An overview of upper limb outcome measures according to body function and structures as well as activity levels of the ICF included 1) capacity measures that assess the person's maximal ability in manual dexterity, gross motor function or both, at a given moment in time, measured in a standardized environment; and 2) patient-reported outcome measures that address upper limb use and perceived difficulty of performing ADLs requiring one or both arms.11 A review on upper limb measures applied in MS rehabilitation documented that the 9HPT was by far the most frequent measure, utilized in 63% of published studies11. As such, the 9HPT is widely considered a gold standard metric for manual dexterity. Besides the 9HPT, other manual dexterity assessment tools such as the Purdue Pegboard test, the Box and Block test, and Coin Rotation Test23 are less frequently used, and only limited studies have addressed their psychometric properties in MS patients.11
MSOAC's review of the literature[12] provided strong documentation supporting utilization of the 9HPT in clinical studies.

- Both inter-rater agreement and test-retest reliability were found to be high across a wide range of levels of disability, indicating the broad utility of the 9HPT in MS.
- The literature provided strong support for the validity of the 9HPT including its ability to discriminate PwMS from healthy controls.
- In terms of convergent validity, the literature review found that the 9HPT correlated modestly with other tests of hand function and highly with actual daily activities that involve hand function.
- The 9HPT also correlates well with most patient-reported outcomes that involve hand function.
- The clinical meaningfulness and relevance of the 9HPT was supported in various ways. Changes in the 9HPT parallel progression of disability in other measures such as the EDSS. A 15-20% change in the 9HPT was found to be related to changes in a wide array of measures such as the EDSS, the MSIS, and others.
- Lastly, the 9HPT has been shown to be responsive to treatments such as steroids.

In summary, the 9HPT has exhibited a wide array of strong measurement properties in a brief, valid, and clinically meaningful measure that can be administered quickly and at low cost in clinical studies. In addition, the literature provides strong support for the clinical meaningfulness of a 15-20% change or difference.

Review of LCLA

As visual dysfunction is one of the most common manifestations of MS, sensitive visual outcome measures are important in examining the effect of treatment. LCLA captures visual loss not seen in high-contrast visual acuity (HCVA) measurements.

Testing of LCLA using Sloan charts was first implemented as an exploratory outcome measure in the IMPACT (International MS Progressive Avonex Clinical Trial) study of interferon beta-1a for secondary progressive MS[24] Both in this study and in a heterogeneous convenience sample cohort of MS patients, it was demonstrated that LCLA was more sensitive than HCVA, L'Anthony D-15 DS color test, and Esterman binocular visual field test in MS patients. Although both Sloan and Pelli-Robson methods distinguished MS subjects from healthy controls significantly better than HCVA, Sloan charts performed better than Pelli-Robson charts with odds ratios for worse visual function scores in MS patients of 2.41 (95% CI 1.77, 3.29; $p<0.001$) for Sloan LCLA versus 1.77 (95% CI 1.38, 2.26 $p<0.001$) for Pelli-Robson contrast sensitivity. Furthermore, only Sloan LCLA was able to distinguish MS subjects from healthy controls in the two lowest age quartiles (18-32 and 33-43 years).[24] MS patients have significantly lower LCLA scores than disease-free controls, a difference that is most pronounced at the lowest contrast levels.[24–26] Importantly, MS and disease-free controls have similar median Snellen VA

scores,[24] supporting previous clinical observations that LCLA and other contrast measures capture aspects of visual function missed by HCVA. Information from these pivotal studies set the stage for use of LCLA as an outcome measure in MS research, clinical trials, and practice.

MSOAC's review of the literature[14] provided strong support for the utility of LCLA in clinical studies.

- LCLA has high inter-rater reliability in both PwMS and healthy controls. Moreover, this holds true across a wide range of LCLA scores.
- The content validity of LCLA has been supported in a variety of ways. Deficits in LCLA are related to deficits in reading, facial recognition, and driving.
- The validity of LCLA has been demonstrated in a variety of ways. LCLA correlates modestly but significantly with the EDSS, indicating that it captures something not captured by the EDSS. For example, PwMS can show worsening on LCLA but not the EDSS.
- Validity of LCLA has been studied using a variety of imaging methods. LCLA is correlated with retinal nerve fiber layer (RNFL) thinning on optical coherence tomography (OCT), showing that LCLA is a marker of neuropathology. LCLA is also related to macular volume reduction, T1 and T2 lesion burden, and cerebral atrophy.
- The validity of LCLA has also been supported by its correlation with increased latency on visual evoked potentials (VEPs) and slower King-Devick scores.
- Sensitivity to change has been supported by the fact that changes in LCLA are predictive of changes in the EDSS.
- Clinical meaningfulness has been supported by the fact that LCLA is related to a variety of both visual and non-visual PRO's such as the IVIS and the SF-36. In addition, changes in LCLA have demonstrated sensitivity to treatments including interferon and natalizumab.
- The literature has provided considerable evidence to support a 7-letter difference or change as clinically meaningful. This is based in part on analysis of the threshold of variability in test-retest reliability and to the fact that a 7-letter loss is related to significant worsening on PROs such as the National Eye Institute Visual Functioning Questionnaire (NEI-VfQ-25) as well as RNFL shown on OCT.

In summary, the review of LCLA has shown that this simple, quick and inexpensive instrument has all the qualities desired in a measure of visual function for use in MS clinical trials where efficiency and sensitivity are important. In addition, the literature provides strong support for the clinical meaningfulness of a 7 letter change or difference.

Review of SDMT

The Literature Review was designed to assess domains of ability in MS, including cognition, and the performance measures that have been used to measure aspects of this domain. Processing speed is a measure of cognitive function underpinning "higher cognitive processes", such as executive function and memory, two other cognitive dimensions commonly affected in persons with MS. Slowed cognitive processing was identified as a core symptom of MS in 1877 by Charcot who presciently observed that in many patients "conceptions are formed slowly and the intellectual and emotional faculties are blunted in their totality."[27] Cognitive dysfunction occurs in all types of MS, all durations, and all severities of disability. Neuropsychological assessment in MS has undergone an evolution as understanding of MS has progressed. Specifically, the major focus has shifted from retrieval of stored information to the original encoding of information in storage, to working memory, and processing speed. Processing speed is considered the essential substrate for all cognitive processes. For example, in order to effectively understand and remember information, one must continuously process the incoming information. Deficits in processing speed can lead to failure to properly take in and process incoming information, thus leading to failure to store that information which leads to an inability to retrieve it at a later time. There is agreement in the MS community that the one best single measure of cognition in MS is processing speed.

Beginning with the work of Stephen Rao in the 1980s,[28,29] cognitive processing speed (CPS) was formally quantified primarily using two neuropsychological tests, the PASAT[30] and the SDMT.[15] The SDMT is the single test common to all recommended cognitive batteries for MS patients including the Brief Repeatable Battery of Neuropsychological Tests (BRB), the Minimal Assessment of Cognitive Function in MS (MACFIMS), NINDS Common Data Elements MS-Cog, and the Brief International Cognitive Assessment for MS (BICAMS).[15] The correlation between processing speed test results and work, school, activity participation, activities of daily living, coping, and quality of life are well documented by many investigators.[31] Recently issued recommendations for cognitive screening of PwMS from a multidisciplinary group of researchers, clinicians, and PwMS chosen by the National Medical Advisory Committee of the National MS Society[31] endorse the use of the SDMT. Likewise, an international consensus on processing speed testing in MS[32] recommended SDMT for clinical trials and clinical care. Slower performance on the SDMT is correlated, at the group level, with ADLs and

employment status. SDMT was the most robust neuropsychological predictor of employment from a comprehensive test battery, with a large effect size on the order of d = 0.80–0.90.[33–35]

MSOAC's review of the literature on cognition[15] provided strong support for the utility of the SDMT in clinical studies.

- The SDMT has excellent test-retest reliability over both short and long periods. This is attenuated somewhat by mild practice effects, but not to a degree that compromises the utility of the test. In addition, the availability of equivalent alternate forms helps to reduce practice effects and maintain consistency in scores over time.
- The literature has considerable evidence for the construct validity of the SDMT. The SDMT loads on a general processing factor and to some extent on a memory factor. It has been shown to be the best test to discriminate PwMS from healthy controls and to predict subsequent cognitive decline.
- There is also substantial evidence for criterion-related validity. The SDMT is the single best predictor of cerebral atrophy, diffusion abnormalities, and lesion burden.
- The ecological validity of the SDMT is probably its greatest shortcoming. Although PwMS are highly accepting of it and in some cases enjoy the task, its significance to them is not obvious. The task entailed in the SDMT does not resemble anything familiar to most people, although PwMS will often report symptoms that are suggestive of processing speed problems, e.g., inability to do things as quickly as before, "brain-fog", etc.
- Despite its lack of intuitive significance, the clinical relevance and meaningfulness of the SDMT has been amply documented in the literature along with estimates of what constitutes a clinically meaningful change or difference. Scores on the SDMT are correlated with instrumental activities of daily living such as cooking, managing finances, and using the Internet. Among cognitive measures, the SDMT is the best predictor of employment status. A 3 or 4 point difference on the SDMT reliably discriminates those who stopped work from those still working. In the course of a relapse, scores on the SDMT are likely to decline by 2 or 3 points and in one study stable vs. relapsing PwMS differed by 5 points on the SDMT.
- Studies have also shown that between 15 and 20% of relapses are exclusively cognitive relapses and that such relapses are missed if cognitive testing is not used.
- Lastly, the SDMT has been shown to be sensitive to the effects of MS disease-modifying therapies based on a 3 or 4 point difference.

In summary, the review of the SDMT has shown that this simple, quick and inexpensive test, among the brief cognitive tests available, stands out as offering the best array of the qualities desired in a measure of cognitive function for use in MS trials. Moreover, the literature provides strong support for the clinical meaningfulness of a 3 to 5 point change or difference.

Overall Conclusions of the Literature Review

Based on review of published, peer-reviewed literature, each of these four performance measures demonstrates strong reliability, validity, sensitivity to change, clinical meaningfulness, and evidence concerning what constitutes a clinically meaningful change or difference. In addition, each performance measure is inexpensive, easily administered in many settings, languages, and cultures, readily accepted by PwMS, requires minimal equipment, and takes very little time to administer. In addition, each of the performance tests can be administered based on clear operating instructions, and thus can produce scores that are directly comparable between evaluators. In conjunction with appropriate PRO measures, this suite of four measures will provide a powerful tool for the evaluation of treatments in MS clinical trials.

PART II VOICE OF THE PATIENT

Background and Methodology of the VOP Study

MSOAC members understood the critical importance of direct input from PwMS about the meaningfulness of performance-based measurement tools. Those individuals living with MS are best positioned to help researchers understand the clinical meaningfulness of measures, to explore any significant gaps in content validity, and to help with estimates of meaningful transitions in levels of functional status. The goal of the Voice of the Patient Study was to contribute evidence towards the meaningfulness to PwMS of the measures of walking speed, manual dexterity, visual acuity, and speed of information processing, which will be used to quantify disability in MS clinical trials.

Project #1 The first VOP project was to (a) estimate the extent to which gradations of activity limitation arising from impairments in each of these 4 performance measures (T25FW, 9HPT, LCLA and SDMT) are judged by PwMS to interfere with their Activities of Daily Living (ADLs); and (b) estimate the extent to which scores on each of the 4 performance measures relate to perceived interference with related ADLs. For walking, manual dexterity, vision and processing speed, a set of 5 daily activities, rank ordered based on increasing level of difficulty, was created for each.

Step 1: Each participant was evaluated using each of the 4 performance measures, administered in random order.

Step 2: Immediately following administration of each of the four performance measures, participants were presented with the set of 5 examples of everyday activities that have increasing levels of difficulty.

Step 3: Immediately following the presentation of each of the 5 everyday activities in Step 2, participants were asked to indicate the impact that their impairment in each of the 4 performance measures had on the associated 5 ADLs using a 0-10 scale (0 indicating no interference and 10 indicating the greatest possible amount of interference).

Project #2 The second project was to obtain the perspectives of PwMS concerning the impact on their ADLs of impairment in areas of functioning other than walking speed, manual dexterity, vision, and speed of information processing. This Project was completed in three steps:

Step 1: Participants were presented with a checklist of five common MS-related symptoms - fatigue, spasticity, incoordination, pain, and bladder problems and were asked to indicate which of these symptoms they experience.

Step 2: Next, participants were asked to identify up to five other MS symptoms that they experience that were not included in the 5-item checklist.

Step 3: For each MS-related symptom they endorsed in Steps 1 and 2, participants were then asked to identify one ADL that was difficult to perform because of that symptom.

Subject recruitment The subjects (n = 62) were drawn from among the MS patients regularly followed at the Mellen Center for Multiple Sclerosis at the Cleveland Clinic Foundation, in Cleveland, Ohio, USA. Participants with a confirmed diagnosis of MS according to the McDonald criteria[39] were selected to ensure diversity in age, type of MS, disability, duration of diagnosis, sex, and ethnicity. Participation required a single visit to the Mellen Center for an approximately 1.5-hour session that is described below. Participants were provided a parking voucher and paid an incentive of $75 upon completion of the study.

Results of the VOP Project #1:

Data in Table 1 demonstrate that as the time to complete walking 25 feet increased, so too did the participant's estimate of the extent to which MS interfered with walking in daily life. For example, participants rated standing up from a chair as the function least affected by MS. However, participants rated climbing several stairs as the walking function most affected by MS out of the 5 used in the study. Moreover participants who needed more time to complete walking 25 feet tended to rate each of the five daily activities as more challenging due to their MS. This same pattern held true for the 9HPT. Participants who needed more time to complete the 9HPT tended to rate daily activities involving manual dexterity as more challenging for them.

Table 1: Performance Measures and Other Factors Related to Patient-Rated Level of Interference in Daily Activities for Mobility and Upper Extremity

Mobility ADL (T*<8.5/T*≥=8.5)				**Upper Extremity ADL**			
	Bα	**Betaβ**	**P value**		**Bα**	**Betaβ**	**P value**
Individual ADL Activity Score							
Univariate							
1.Stand up from chair	**1.15** *0.14*	-	**<.001**/*<0.18*	1.Brush teeth	0.10	-	<.001
2.Short walk	**1.06**/*0.09*	-	**0.005**/*0.28*	2.Cut food with utensil	0.19	-	<.001
3.Get up from floor	**1.45**/*0.21*	-	**<.001**/*0.02*	3.Write with pen	0.23	-	<.001
4.Jump up and down	**1.92**/*0.12*	-	**<.001**/*0.08*	4. Pick coins from table	0.18	-	<.001
5.Climb several stairs	**1.71**/*0.11*	-	**<.001**/*0.19*	5. Change bulb	0.15	-	<.001

Data for individuals with T25FW**<** 8.5 is shown **in boldface**; *Data for individuals with T25FW≥8.5 is shown in italics*

Results of the VOP Project #2:

In Table 2 the symptoms that were presented to respondents included fatigue, incoordination, spasticity, bladder and pain. The remaining columns represent symptoms that PwMS volunteered as being of concern to them. The first column, labeled "Activities of Daily Living Affected by Listed Symptoms", are the ADL's the participants identified as being interfered with by the symptoms in the first row. The "X" in the body of the table indicates those symptoms which affect performance of ADLs and, conversely, for any given ADL, which symptoms cause interference. For example, the symptom "Fatigue" was reported by at least one participant to interfere with one of the following ADL: walking, exercise, working, household chores, any physical activity, thinking/mental activities, shopping/errands, managing stress, staying awake or playing with grandchildren. The table also indicates that the ADL "Working" was reported by at least one participant to be affected by one of the following symptoms: Fatigue, Spasticity, Bladder, or Pain.

Table 2: Patient Reported Additional Symptoms and ADLs Affected by Them*, **

	Symptoms Presented to PwMS					Other Symptoms Recommended by PwMS		
Activities of Daily Living Affected by Listed Symptoms	**Fatigue**	**Incoordination**	**Spasticity**	**Bladder**	**Pain**	**Parasthesias**	**Sensory loss**	**Weakness**
Total report symptom: N (%)	56(90.3%)	55(88.7%)	47(75.8%)	43(69.4%)	34(54.8%)	13(22.6%)	9(26.1%)	7(11.3%)
Walking	X	X	X	X	X			X
Exercising	X	X	X		X			
Working	X		X	X	X			
Household chores	X	X	X		X	X		X
Any physical activity	X							
Thinking/mental activities	X							
Shopping/errands	X			X			X	
Managing stress	X		X					
Staying awake	X							
Playing with grandkids	X							
Getting out of bed			X		X			
Sitting			X			X	X	
Going to bed			X					
Driving			X					
Standing		X	X		X			
Relaxing			X					
Sleeping				X	X			
Climbing stairs		X			X			
Yard work		X						
Writing		X						
Transferring		X						
Drinking liquids				X				
Taking long trips				X				
Taking short trips				X				
Coughing/sneezing				X				

*All respondents endorsed at least one of the presented or endorsed sympotms (N=62). An X indicates at least 1 subject reported a symptom affecting this ADL

**Symptoms that were endorsed by less than 10% of respondents: Psychological (6.5%), Sexual (6.5%), Spasms (4.8%), Headache (4.8%), Dizziness (4.8%), Dysarthria (4.8%), Bowel (4.8%), Lhermittes (3.2%)

Overall Conclusions of the VOP Study

The purpose of this study was to gain input from PwMS concerning the clinical meaningfulness of four performance measures in relation to daily activities, as well as to explore what additional MS -related symptoms were problematic to the participants, and what activities were affected by those symptoms. In this sample, participants ranked interference with ADLs in close alignment with their performance test results in the physical domains of mobility and upper extremity function. It is clear that the mobility and dexterity domains represent important areas of compromise in terms of daily functioning. Furthermore, the relationships between gradations shown by scores on the performance measures and the gradations in interference in daily activities suggests that these performance measures are not only measuring something important to patients but to a great extent are capturing the patient's perception of severity. That is, these domains are meaningful to patients and the underlying metrics of the performance measures are quantitatively meaningful to them as well.

By comparison, participants did not rank interference with ADLs related to vision or cognition as closely correlated to test results from the vision and cognition performance measures. The discrepancy between the motor and non-motor performance measures, as related to PwMS ADL interference ratings is not currently well-understood. It is possible that PwMS do not perceive the specific performance measure for vision and cognition (Low Contrast Letter Acuity and Symbol Digit Modalities

Test) as relevant to their daily activities. It is also possible that the degree of interference in daily activities is much more nuanced for vision and cognition than it is for walking and manual dexterity, and may be related to other factors such as employment or hobbies. For example, visual problems for a surgeon would entail more disruption of daily activities than for a psychotherapist. Mild cognitive problems may be more disruptive for a crossword or card player than a gardener or swimmer. Another confounding factor is the issue of ability to compensate. For example, individuals with mild visual impairments could compensate for such deficits in a variety of ways, particularly when such deficits emerge gradually. There are a number of reasons why we found less alignment between the vision and cognition measures compared to perception of daily activities. This relative lack of alignment should not lead to the interpretation that the measure is not clinically meaningful, as the literature data demonstrate otherwise, but only that PwMS do not relate scores on two unfamiliar tests to interference with their ADLs related to those disease dimensions.

Evidence of the ecological validity of LCLA was obtained through MSOAC's review of the literature.[14] The 25-question National Eye Institute Visual Functioning Questionnaire (NEI-VFQ-25) is a widely used and well-validated measure of vision specific quality of life (QOL) that captures activity limitations in patients with MS and in a variety of ocular disorders. To better assess some unique features of visual dysfunction in MS and other neuro-ophthalmologic conditions, a 10-Item Neuro-Ophthalmic Supplement to the NEI-VFQ-25 was designed with participation of MS patients in focus groups. Both the NEI-VFQ-25 and 10-Item Supplement have been implemented in MS clinical trials. It is now well established that reductions in LCLA reflect worse scores for vision-specific QOL. Two-line (10-letter) differences in LCLA are associated with 4-point or greater reductions in NEI-VFQ-25 composite scores.[40] This is important since 4-point differences in overall score are considered clinically meaningful for the NEI-VFQ-25 (Submacular Surgery Trials Research Group, 2007).

MSOAC members acknowledge the disadvantages of the SDMT. First, the test measures information processing speed, which is only one component of the cognition domain. Even though SDMT explains some of the variance in memory or higher executive functioning, SDMT does not assess memory or higher executive function *per se*. MSOAC has carefully weighed the pros and cons of including other cognitive measures into the COA being developed for cognitive disability, and elected to recommend a single cognitive test – the SDMT – as opposed to a test battery, such as the Brief International Cognitive Assessment for MS (BICAMS). Including multiple cognitive tests, which themselves are significantly correlated, would add significant analytical and practical complexity to the necessary inclusion of cognitive assessments in MS clinical trials. Also, motor or visual disability related to MS are measured with a single measure, which may not entirely capture the construct of interest in a comprehensive manner. For example, upper extremity function is captured with the 9HPT, when tests of strength and hand sensation could be added to provide a more comprehensive assessment. A timed 25-foot walk is typically used to characterize ambulation, when distance walk, axial sway, and other measures could be used to provide a more comprehensive assessment. Use of a single cognitive test is conceptually similar. A second weakness of SDMT, which was evident in the VOP results, is face validity: the SDMT may not seem directly related to common activities of daily living. This limitation is inherent to structured, validated tests of neuropsychological performance testing, the goal of which is to quantify cognitive performance relative to age, gender, and education matched controls. This possible shortcoming could be addressed through label language, e.g. "cognitive disability as measured by information processing speed testing."

The second part of the project provided evidence that there are a number of domains, in addition to the four addressed in the first part of the study, which are also important to patients and which result in compromise of daily activities. Most of these, such as fatigue, do not lend themselves to objective measurement but can be validly assessed using existing patient reported outcome measures. The results from this study provide further evidence to support the inclusion of such patient reported outcomes as companion measures to the objective performance measures that were the subject of this study. Ideally a clinical trial should incorporate performance measures such as those used in this study in conjunction with PROs in order to achieve a holistic picture of the impact of MS and the efficacy of proposed therapies.

Altogether, the VOP study provided evidence that the four performance measures included in this investigation represent important areas of functioning for PwMS. While associations among the Cognition and Vision performance measures and their related interference scores were not strong, the correlations were in the expected direction. The T25FW and 9HPT, as measures of mobility and dexterity respectively, represent important areas of compromise in terms of daily functioning. In addition, the relationships between gradations shown by scores on the performance measures and the gradations in interference in daily activities suggests that these performance measures are not only

measuring something important to patients but to a great extent are capturing the patient's perception of severity.

PART III ANALYSIS OF AGGREGATED CLINICAL TRIAL DATA

Methodology

Data Sources *Table 3 provides a summary list and description of the data sources by name and ClinicalTrial.gov number (https://clinicaltrials.gov) that were aggregated for the instrument development and validation. This MSOAC database represents the largest pooled analysis of prospectively acquired clinical trial data in MS to date. De-identified patient-level trial data from 14 clinical trials totaling 12,776 subjects were mapped to the standard Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDMT) with the aid of available translation instruments. The CDISC foundational standards, as well as the MSOAC-developed CDISC Therapeutic Area Data Standard for MS (version 1.0) were utilized. All submitted data were validated by a Quality Control process before being made available to Premier Research, the CRO conducting the statistical analysis.*

Table 3: Source Datasets in the MSOAC Database

Study	CT.gov #	n	Type	EDSS	FSS	T25FW	9HPT	PASAT	SDMT	LCLA	SF-36	BDI-II
ADVANCE	NCT00906399	1512	RRMS	√	√	√	√	√	√	√	SF-12	√
AFFIRM	NCT00027300	939	RRMS	√	√	√	√	√	No	√	√	No
CARE-MS 1	NCT00530348	563	RRMS	√	√	√	√	√	No	√	√	No
CARE-MS 2	NCT00548405	798	RRMS	√	√	√	√	√	No	√	√	No
CombiRx	NCT00211887	1008	RRMS	√	√	√	√	√	No	√	√	No
FREEDOMS	NCT00289978	1272	RRMS	√	√	√	√	√	No	No	No	No
FREEDOMS II	NCT00355134	1083	RRMS	√	√	√	√	√	No	√	No	No
IMPACT	N/A	434	SPMS	√	√	√	√	√	No	No	√	√
MAESTRO	NCT00869726	610	SPMS	√	√	√	√	√	No	No	√	No
PROMISE	N/A	943	PPMS	√	√	√	√	√	No	No	√	No
SENTINEL	NCT00030966	1196	RRMS	√	√	√	√	√	No	√	√	√
STRATA	NCT00297232	1094	RRMS	√	√	No	No	No	√	No	No	BDI-FS
TEMSO	NCT00134563	1086	RRMS	√	√	√	√	√	No	No	√	No
TRANSFORMS	NCT00340834	1292	RRMS	√	√	√	√	√	No	√	No	No

Analyses focused on the four performance measures (T25FW, 9HPT, LCLA, and SDMT) and their relation to other measures in the database: 1) the EDSS is an ordinal scale ranging from 0-10 based on the severity of findings on the neurological examination, walking ability, and ability to carry out activities of daily living, with higher scores indicating worse disability; 2) the Paced Auditory Serial Addition Test (PASAT); 3) the Beck Depression Inventory (BDI) is a 21-item self-report measure of depression with scores ranging from 0 to 62 and higher score indicating more severe depression symptoms;[41] and 4) the Short Form-36 (SF-36) is a 36-item questionnaire that includes eight multi-item health concepts (Physical Functioning, Role-Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role-Emotional, and Mental Health).[42] Scores are a mean of subsetted questions and range from 0 to 100; higher scores indicate better health-related quality of life (HRQoL). The SF-36 has two summary scales, the Physical Component Summary (PCS) and the Mental Component Summary (MCS) whose calculation produces a T-score, with a mean score of 50 and SD of 10, representing the reference score for the United States general population.

Limitations of this work include the availability of somewhat fewer data for LCLA, SDMT, and self-reported measures. Relatively few datasets contained all four performance measures and EDSS, limiting the analyses of their relative sensitivity. The ability to fully explore clinical meaningfulness of the performance measures using self-report measures also was restricted. Other measures of self-report have been applied to the MS population, but these analyses were limited by the surveys in the existing data set, i.e., the SF-36. Also, although the dataset included the full range of disability, the majority of subjects had RRMS and relatively mild disability, with median EDSS of 2.5, reflecting the over-representation of clinical trials in RRMS in the MS field at the time the database was constructed. This point may limit a full understanding of the performance tests in more disabled, progressive populations. Finally, for these analyses, pooled treatment groups and focus on three-month confirmed disability worsening (rather than six-month) could have affected the results.

Statistical methods The Statistical Analysis Plan (SAP) was developed by an expert MSOAC group to assess the following attributes of each of the four performance measures: 1) magnitude of floor or

ceiling effect; 2) test-retest reliability; 3) whether scores decrease with time as the course of MS progresses; 4) construct validity; 5) convergent validity by assessing correlation with overall disability measured by EDSS; 6) extent to which the performance scores are affected by practice effects; 7) known group validity by comparing performance scores in patients with short vs long disease duration, with high vs low EDSS scores; 8) sensitivity to change by comparing scores before and after disease worsening or improvement assessed by EDSS, and before and after disease relapses; and 9) the minimum clinically important change in performance scores.

No imputation was done for missing data other than for participants unable to complete the T25FW or 9HPT because of disability. Following convention, imputation for patients unable to perform was 180 sec for T25FW and 300 sec for 9HPT.[36] The MSFC administration and scoring manual states that for T25FW testing patients should use their usual assistive devices and an effort should be made to use the same device over the course of the study. Summary scores of the SF-36 MCS and PCS were calculated using standard methods which provide T-scores for analysis. For the SF-36 eight health concept scores, QualityMetrics Health Outcomes™ Scoring Software was utilized. The maximum data recovery method was used to handle missing data. If any individual item was missing for the BDI score, the total score was not calculated for that participant and time point.

Test-retest reliability was assessed by intraclass correlation coefficient (ICC) of all administrations of each test (2-6 compared with test 1) based on periods in which patient status on the EDSS did not change and not exceeding six months from baseline.[37] Correlations among the EDSS and performance tests were assessed by Spearman rank correlation coefficient. Time to confirmed clinically meaningful worsening was analyzed by Kaplan-Meier methods. Cohen's kappa coefficient was used to assess agreement in worsening in different disability measures. The baseline score for each performance measure was compared between the groups of patients based on disease duration and EDSS score using an ANOVA model adjusting for age in five-year age bands.

For these analyses, worsening was defined as follows: T25FW (20% increase)[13]; 9HPT (20% increase);[12] LCLA with 2.5% contrast (20% or seven-letter decrease);[13] SDMT (four-point decrease);[15] EDSS (baseline score 0: 1.5-point increase, baseline score 1.0-5.5: 1-point increase, baseline score ≥6.0: 0.5-point increase);[38] and SF-36 PCS Score (five-point worsening).[39] For all variables except SF-36 PCS, the worsening had to be sustained for at least three months.

Results of the Analysis of Aggregated Clinical Trial Data

Table 4 summarizes the data available, and baseline demographics, disease characteristics, EDSS score, performance test results, and participant self-reported measures. Overall, the population was relatively young with a recent diagnosis of MS, predominantly relapsing-remitting (RR) course, and mild disability. Although fewer studies included LCLA, SDMT, and self-reported measures, substantial data were available for all outcome measures.

Table 4: Baseline Characteristics for the Aggregated Dataset Derived from 14 Clinical Studies.

PARAMETER	N	VALUE				
AGE (years)	N	Mean	Standard Deviation	Minimum	Median	Maximum
	12776	39.5 years	9.92	17	40.0	72
POOLED AGE GROUP		< 35 years	35-45 years	> 45 years		
	12727	4148 (32.6%)	4864 (38.2%)	3715 (29.2%)		
GENDER		Male	Female			
	12776	3977 (31.1%)	8799 (68.9%)			
ETHNIC ORIGIN		Native American	Asian	African American	White	Other
	9118	12 (0.1%)	228 (2.5%)	263 (2.9%)	8326(91.3%)	289 (3.2%)
GEOGRAPHIC REGION		Europe	North America	Other Regions		
	6568	3463 (52.7%)	2118 (32.2%)	987 (15.0%)		
TREATMENT ARMS		Placebo	Glatiramer Acetate or Interferon Beta	Other Drug		
	12776	2614 (20.5%)	4093 (32.0%)	6069 (47.5%)		
MS SUBTYPE		RRMS	SPMS	PPMS		
	12776	10789 (84.4%)	1044 (8.2%)	943 (7.4%)		
DISEASE DURATION AT BASELINE		Mean	Standard Deviation	Minimum	Median	Maximum
	6641	6.5 years	7.26	0 years	4.0 years	48 years
Duration						

PARAMETER	N		VALUE			
AGE (years)	N	Mean	Standard Deviation	Minimum	Median	Maximum
Category		≤ 10 years	≥ 10 years			
	6641	5016 (75.5%)	1625 (24.5%)			
BASELINE PERFORMANCE MEASURES	N	Mean	Standard Deviation	Minimum	Median	Maximum
EDSS	12776	2.9	1.63	0	2.5	8
SDMT	2583	47.9	15.90	0	48.0	110
PASAT	11609	48.1	11.42	0	52.0	60
9HPT	11653	24.3	14.30	5	21.3	331
T25FW	11649	7.6	9.84	1	5.4	231
LCLA (2.5%)	5669	34.6	11.65	0	37.0	60
BDI	2824	8.9	8.62	0	7.0	53
MCS	7766	47.7	11.53	-5	49.5	74
PCS	7766	41.5	9.95	10	40.9	73
BASELINE CATEGORIES						
EDSS		0 - 3.5	4 - 10			
	12746	9279 (72.8%)	3467 (27.2%)			
9HPT		Below median (≤ 21.3)	Equal to or above median (≥ 21.3)			
	11653	5805 (49.8%)	5848 (50.2%)			
T25FW		Below median (≤ 5.4)	Equal to or above median (≥ 5.4)			
	11649	5798 (49.8%)	5851 (50.2%)			
LCLA (2.5%)		Below median (≤ 37)	Equal to or above median (≥ 37)			
	5969	2781 (49%)	2888 (51%)			

A comparison of the data for two cognition measures, SDMT and PASAT, showed that SDMT was superior to PASAT in several aspects: 1) the PASAT showed a severely negatively skewed distribution, indicative of pronounced ceiling effects; 2) practice effects were larger with the PASAT; 3) the SDMT was correlated with physical measures to a higher degree than PASAT; SDMT was correlated with the PCS of the SF-36 to a higher degree than PASAT. Importantly, patient experience is favorable concerning the SDMT. Many MS patients appear to enjoy completing this test, in distinct contrast to the Paced Auditory Serial Addition Test, which MS patients report to be distressing. Consequently, SDMT became the focus of MSOAC's qualification for a measure of cognition.[40]

The frequency distributions of the T25FW and 9HPT were positively skewed and showed floor effects, with scores tending to be clustered at shorter times (Figure 1). Both possessed the ability to distinguish gradations of performance in the middle of the scale. LCLA distribution appeared mildly negatively skewed without floor or ceiling effects. SDMT scores showed no evidence of skewing, or floor or ceiling effects.

Figure 1
Distribution of performance measure scores at baseline. A Timed 25-Foot Walk (sec). B Nine-Hole Peg Test (sec). C Low Contrast Letter Acuity with 2.5% contrast (number correct). D Symbol Digit Modalities Test (number correct).

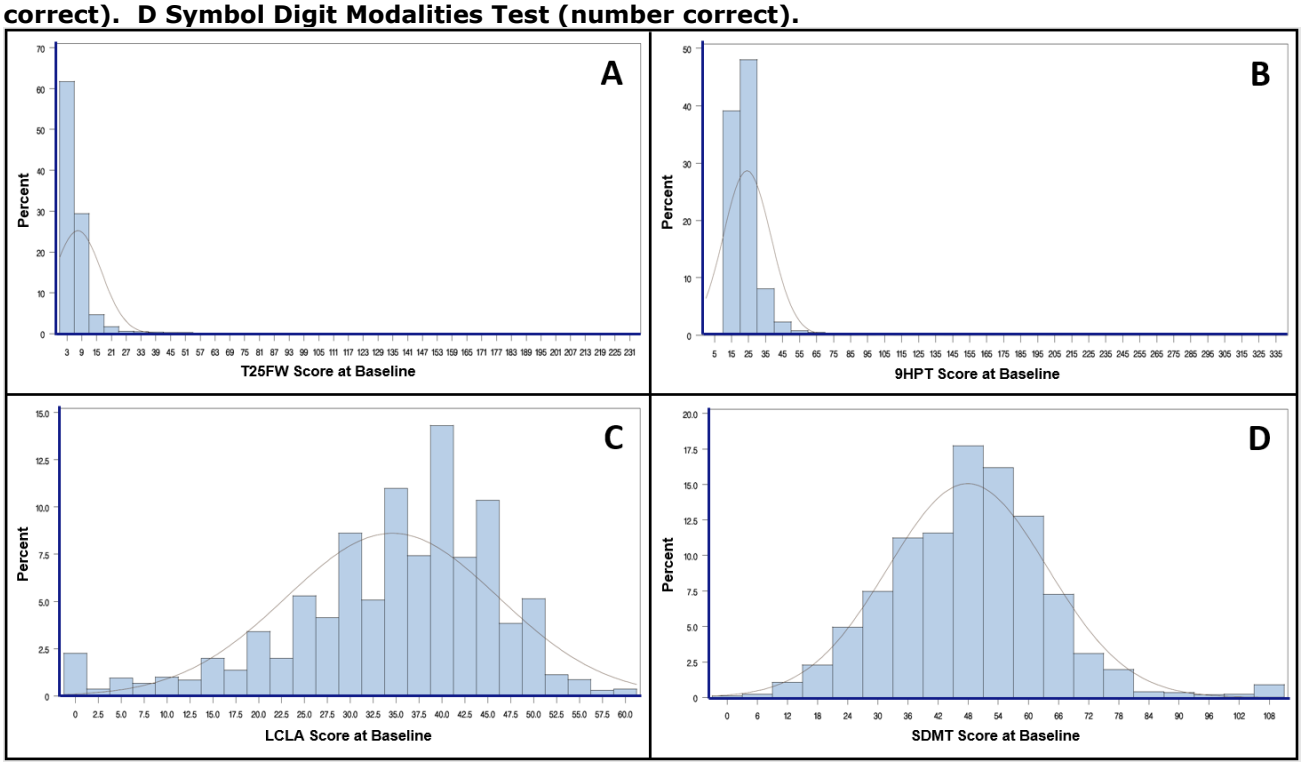


Table 5 summarizes trends over the first six assessments for the performance tests. T25FW, 9HPT, and LCLA tended to worsen over time and showed minimal or no practice effects, while the SDMT demonstrated modest practice effects. Test-retest reliability was estimated by calculating the ICC, accounting for practice effects where needed. All measures showed good test-retest reliability, though the ICC for T25FW was somewhat lower (0.71) compared to the other tests (0.84-0.88).

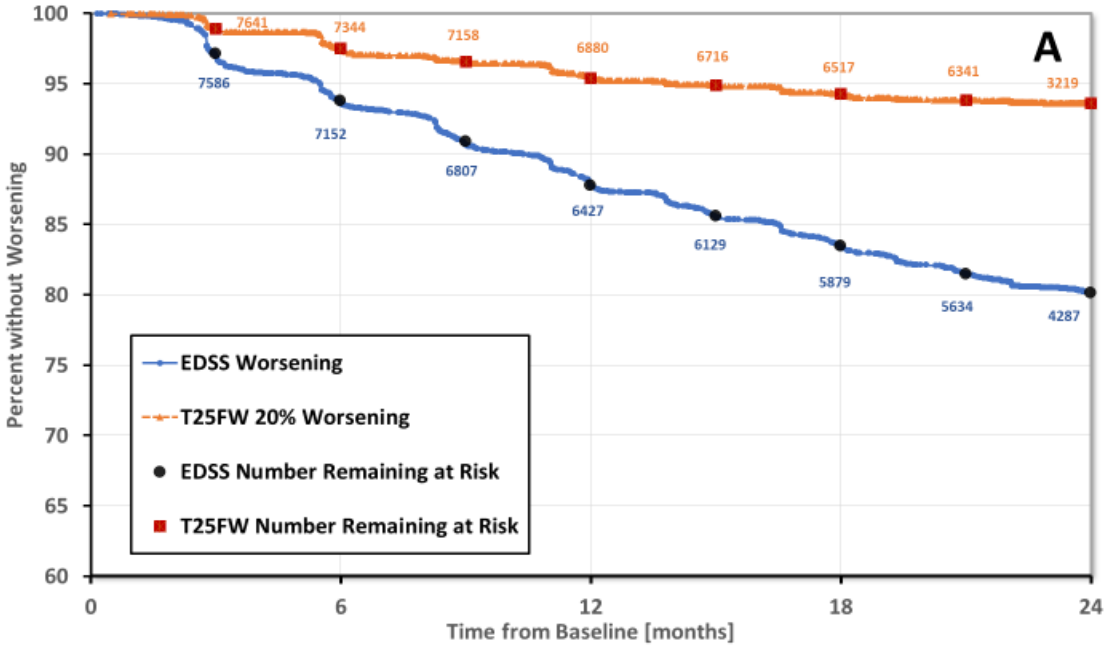
Table 5: Practice Effects & Test-retest Reliability Measures with Tests 2-6 each Compared to Test 1.

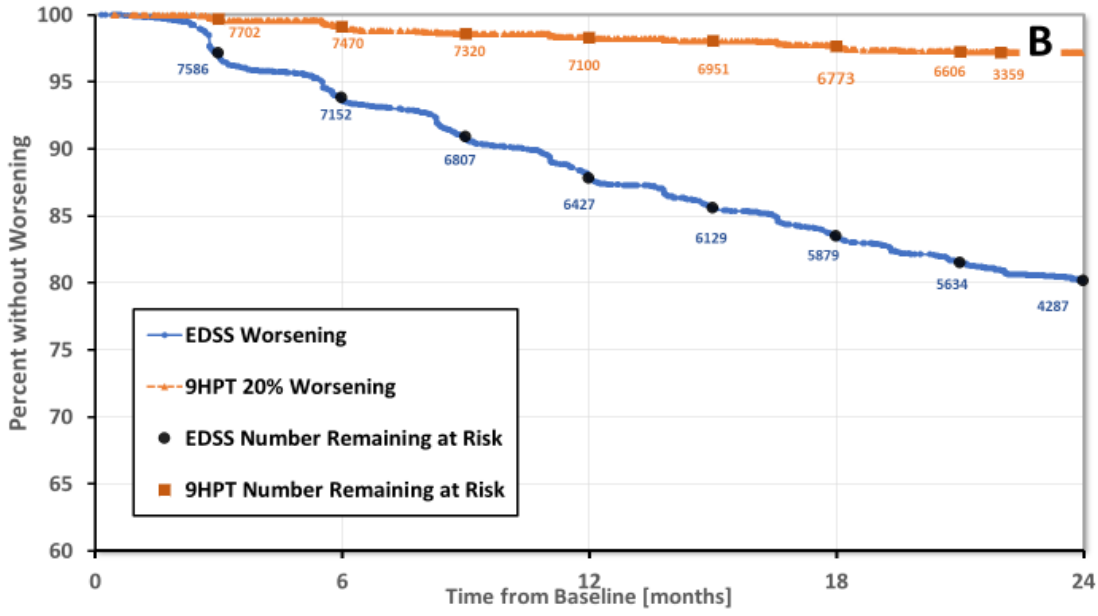
Measure	N	Test 2	Test 3	Test 4	Test 5	Test 6	ICC
T25FW	7,971	0.08	0.08	0.05	0.08	0.13	0.71
9HPT	7,973	0.02	0.00	-0.03	-0.04	0.00	0.84
LCLA	4,611	-0.02	-0.03	-0.01	0.00	0.00	0.88
SDMT	2,094	0.03	0.10	0.15	0.28	0.37	0.85

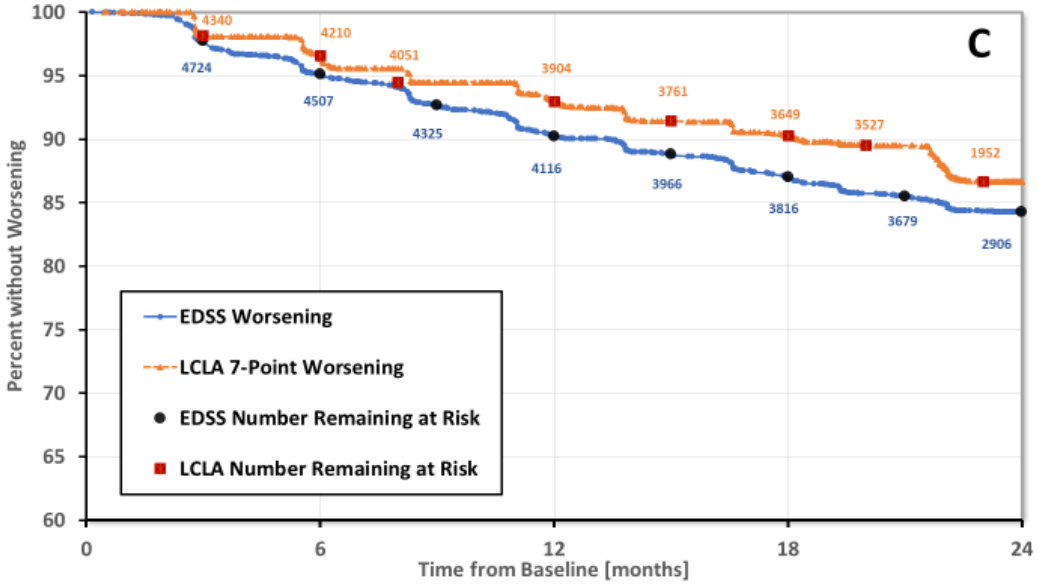
The values for Test 2-Test 6 are the regression coefficients for the 2nd to 6th test, expressed as an effect size to make them comparable. For example, with T25FW, the 2nd test was on average 0.08 standard deviations higher than the first test. ICC = intraclass correlation coefficient (a measure of reliability, higher is better, 1 is the maximum possible score); LCLA = Low Contrast Letter Acuity with 2.5% contrast; N = number of subjects.

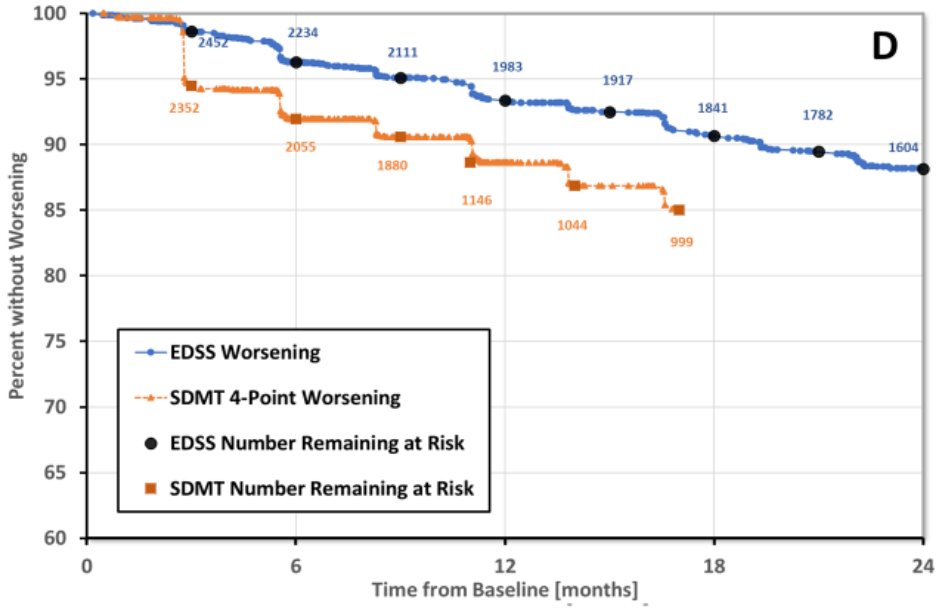
To compare sensitivity to change of the performance measures with EDSS, time from baseline to three-month confirmed worsening over 24 months was analyzed (Figure 2). The study populations available for each comparison differed, leading to differing proportions with three-month confirmed worsening on EDSS. Using a 20% threshold for T25FW, 6.5% worsened compared to 20.2% on EDSS. Using a 20% threshold for 9HPT, 2.9% worsened compared to 20.2% on EDSS. Using seven-point threshold for LCLA, 13.1% worsened compared to 16.1% on EDSS. Using four-point threshold for SDMT, 15.0% worsened compared to 14.5% on EDSS. Thus, progression rates were lower for T25FW and 9HPT compared to that of EDSS, while progression rates for LCLA and SDMT were similar to that of EDSS. When the performance tests were combined into a multidimensional outcome measure, the proportion of participants worsening on any one performance test was greater to the proportion worsening on EDSS. When worsening on two performance tests was required, sensitivity to disability progression was somewhat reduced compared to the EDSS. The progression events defined by the performance tests were weakly associated with or independent of those defined by the EDSS: T25FW (Cohen's κ=0.02, 95% confidence interval [CI] -0.00 to 0.03), 9HPT (κ=0.00, 95 CI -0.01 to 0.01), LCLA (κ=0.11, 95% CI 0.08 to 0.14), and SDMT (κ=-0.02, 95% CI -0.06 to 0.02).

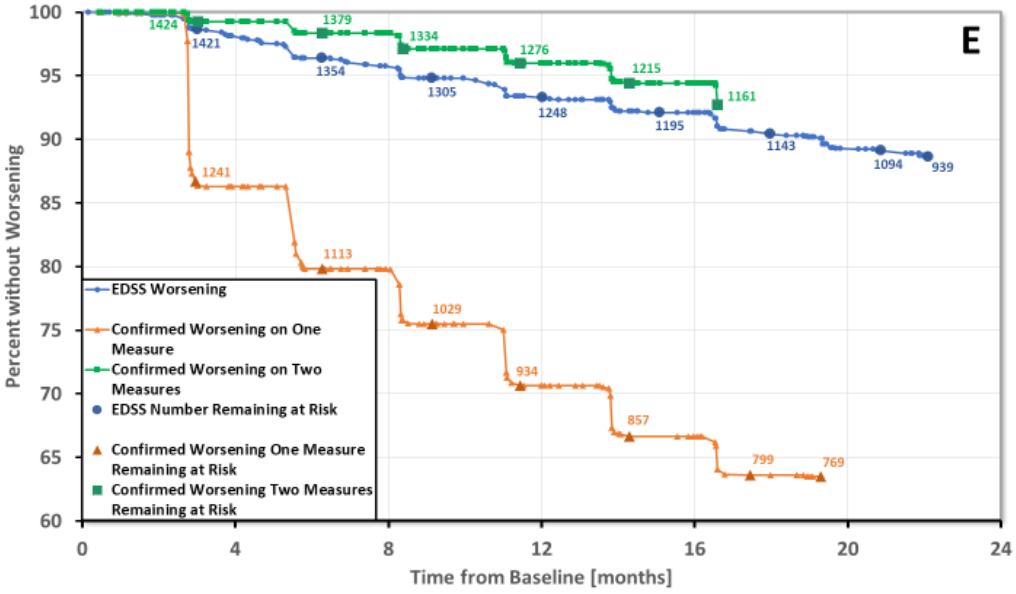
Figure 2: Kaplan-Meier Graphs of Time to Three-month Confirmed Disability Worsening of Performance Measures Compared to Expanded Disability Status Scale. A T25FW. B 9HPT. C LCLA with 2.5% contrast. D SDMT. E any one or two of the performance measures.






C
Percent without Worsening
Time from Baseline [months]
EDSS Worsening
LCLA 7-Point Worsening
EDSS Number Remaining at Risk
LCLA Number Remaining at Risk
100
95
90
85
80
75
70
65
60
0
6
12
18
24
4340
4210
4051
3904
3761
3649
3527
1952
4724
4507
4325
4116
3966
3816
3679
2906


D
Percent without Worsening
Time from Baseline [months]
EDSS Worsening
SDMT 4-Point Worsening
EDSS Number Remaining at Risk
SDMT Number Remaining at Risk
100
95
90
85
80
75
70
65
60
0
6
12
18
24
2452
2234
2111
1983
1917
1841
1782
1604
2352
2055
1880
1146
1044
999


E
Percent without Worsening
Time from Baseline [months]
EDSS Worsening
Confirmed Worsening on One Measure
Confirmed Worsening on Two Measures
EDSS Number Remaining at Risk
Confirmed Worsening One Measure Remaining at Risk
Confirmed Worsening Two Measures Remaining at Risk
1424
1421
1379
1354
1334
1305
1276
1248
1215
1195
1161
1143
1094
939
1241
1113
1029
934
857
799
769

To investigate construct and convergent validity, correlations between the performance measures and EDSS were analyzed (Table 6). The T25FW and 9HPT correlated strongly with one another and demonstrated the strongest correlation to the EDSS relative to other performance measures. LCLA and SDMT were weakly correlated to the other performance measures and EDSS. Between the two, the SDMT had somewhat stronger correlation to the EDSS. Cross-sectional correlations among outcomes at baseline were notably stronger than the correlations among changes from baseline-to-endpoint, which had a similar pattern of correlative strength (T25FW > 9HPT> SDMT > LCLA), but wholly weaker in magnitude.

Table 6: Correlations between Outcome Measures

Baseline correlations							
	9HPT	**LCLA**	**SDMT**	**EDSS**	**PCS**	**MCS**	**BDI**
T25FW	**0.52** (0.51 to 0.53)	**-0.30** (-0.32 to -0.27)	**-0.42** (-0.46 to -0.38)	**0.56** (0.55 to 0.58)	**-0.40** (-0.42 to -0.38)	**-0.13** (-0.16 to -0.11)	**0.22** (0.18 to 0.26)
9HPT		**-0.33** (-0.35 to -0.31)	**-0.47** (-0.51 to -0.43)	**0.54** (0.53 to 0.56)	**-0.33** (-0.36 to -0.31)	**-0.14** (-0.16 to -0.11)	**0.20** (0.16 to 0.24)
LCLA			**0.34** (0.30 to 0.39)	**-0.29** (-0.31 to-0.27)	**0.12** (0.09 to 0.14)	**0.19** (0.16 to 0.22)	**-0.16** (-0.20 to -0.12)
SDMT				**-0.34** (-0.38 to -0.29)	**0.36** (0.32 to 0.41)	**0.21** (0.16 to 0.26)	**-0.20** (-0.24 to -0.15)
Correlations of change from baseline to endpoint							
	9HPT change	**LCLA change**	**SDMT change**	**EDSS change**	**PCS change**	**MCS change**	**BDI change**
T25FW change	**0.30** (0.28 to 0.32)	**-0.08** (-0.11 to 0.06)	**-0.14** (-0.19 to 0.09)	**0.29** (0.27 to 0.31)	**-0.20** (-0.23 to 0.18)	**-0.09** (-0.12 to 0.06)	**0.10** (0.05 to 0.14)
9HPT change		**-0.06** (-0.09 to -0.04)	**-0.20** (-0.25 to -0.15)	**0.23** (0.22 to 0.25)	**-0.16** (-0.19 to -0.13)	**-0.07** (-0.10 to -0.05)	**0.11** (0.07 to 0.16)
LCLA change			**0.06** (0.01 to 0.11)	**-0.11** (-0.13 to -0.08)	**0.02** (-0.01 to 0.05)	**0.06** (0.03 to 0.10)	**-0.02** (-0.07 to 0.03)
SDMT change				**-0.12** (-0.16 to -0.08)	**0.00** (-0.01 to 0.05)	**0.06** (0.03 to 0.10)	**-0.09** (-0.13 to -0.04)

Values are Spearman correlation coefficients (95% CI). CI = confidence interval; LCLA = Low Contrast letter Acuity with 2.5% contrast.

Known group validity was assessed as a function of disease duration and disability level (Table 7). At baseline, values for all four performance measures were better in participants with MS of shorter duration (<10 years since symptom onset) compared to those with disease of longer duration (≥10 years). Similarly, the results on all four performance tests were better in participants with lower EDSS scores (0-3.5) versus those with higher EDSS scores (4.0-10).

Table 7: Known Group Analysis of Baseline Values based on Disease Duration and Disability Level

Measure	**Disease duration (years)**			**EDSS**		
	<10	**≥10**	**Difference (95% CI) P-value**	**0-3.5**	**4.0-10**	**Difference (95% CI) P-value**
T25FW (sec)	**7.7**	**13.3**	**N=5597** 5.57 (4.74 to 6.40) P<0.0001	**6.1**	**12.7**	**N=11595** 6.63 (6.21 to 7.06) P<0.0001
9HPT (sec)	**24.3**	**29.9**	**N=5599** 5.57 (4.48 to 6.65) P<0.0001	**21.7**	**31.8**	**N=11594** 10.10 (9.48 to 10.72) P<0.0001
LCLA (number	**33.2**	**30.7**	**N=3579** -2.50	**34.8**	**27.4**	**N=5787** -7.46

Measure	Disease duration (years)			EDSS		
correct)			(-3.75 to -1.25) P<0.0001			(-8.33 to -6.60) P<0.0001
SDMT (number correct)	**48.5**	**45.2**	**N=2543** -3.31 (-4.85 to -1.77) P<0.0001	**49.8**	**41.2**	**N=2583** -8.60 (-10.09 to -7.12) P<0.0001

CI = confidence interval; LCLA = Low Contrast Letter Acuity with 2.5% contrast.

To explore clinical meaningfulness, correlations were calculated between performance measures and participant self-reported measures of HRQoL and depression (Table 8). At baseline, T25FW, 9HPT, and SDMT correlated moderately with SF-36 PCS and significantly but weakly with MCS and BDI. LCLA correlated weakly with SF-36 PCS and MSC, and BDI. Correlations between change baseline-to-endpoint in the performance measures and change on SF-36 PCS or MCS, or BDI were generally not significant and weak at best. Among participants with worsening from baseline-to-endpoint on the T25FW, 9HPT, or SDMT, the mean SF-36 PCS also worsened (P<0.001, P<0.001, and P=0.0308, respectively). Similarly, among participants who showed baseline-to-endpoint worsening on the T25FW, 9HPT, or SDMT, the proportions of participants with five-point PCS worsening on SF-36 PCS were greater. Non-significant trends were seen for mean PCS change and the proportion with five-point PCS change among participants who did or did not experience baseline-to-endpoint worsening on LCLA. Mean PCS worsened among participants with baseline-to-endpoint worsening in each of two groups: those with worsening on any one measure and those worsening on two or more performance measures. The SF-36 PCS was improved or stable, respectively, among participant who did not worsen on one or on two or more performance measures. Similarly, the proportions of participants with five-point SF-36 PCS worsening were greater among participants who showed baseline-to-endpoint worsening on one or on two or more performance measures.

Table 8: Change in SF-36 PCS in Participants with and without Worsening on EDSS and Performance Measures

Disability measure	N	Absolute change in PCS (SD) among PwMS with disability measure worsening	Absolute change in PCS (SD) among PwMS without disability measure worsening	P-value	Percent (95% CI) with 5-point PCS worsening among PwMS with disability measure worsening	Percent (95% CI) with 5-point PCS worsening among PwMS without disability measure worsening	Odds ratio (95% CI) P-value
EDSS	**Total: 7455** Worse: 1479 Not worse: 5976	**-2.75** (8.21)	**0.43** (7.54)	**P<0.0001**	**36.6** (34.1 to 39.1)	**20.5** (19.5 to 21.6)	**2.23** (1.98 to 2.53) P<0.0001
T25FW (20%)	**Total: 7455** Worse: 1666 Not worse: 5789	**-2.18** (7.83)	**0.37** (7.67)	**P<0.0001**	**33.4** (31.1 to 35.7)	**20.9** (19.9 to 22.0)	**1.89** (1.68 to 2.13) P<0.0001
9HPT (20%)	**Total: 7455** Worse: 622 Not worse: 6833	**-2.86** (8.33)	**0.04** (7.68)	**P<0.0001**	**38.6** (34.7 to 42.5)	**22.3** (21.4 to 23.4)	**2.18** (1.84 to 2.59) P<0.0001
LCLA (7 point)	**Total: 4678** Worse: 570 Not worse: 4108	**0.03** (7.95)	**0.38** (7.49)	**P=0.2907**	**22.1** (18.8 to 25.7)	**20.0** (18.8 to 21.3)	**1.13** (0.92 to 1.40) P=0.2662
SDMT (4-point)	**Total: 1467** Worse: 288 Not worse: 1179	**-1.15** (8.19)	**-0.04** (7.69)	**P=0.0308**	**28.8** (23.7 to 34.4)	**22.2** (19.9 to 24.7)	**1.42** (1.06 to 1.89) P=0.0201
Worse on any 1 measure (T25FW or 9HPT or LCLA or SDMT)	**Total: 7455** Worse: 2478 Not Worse: 4977	**-1.63** (7.94)	**0.51** (7.60)	**P<0.0001**	**30.8** (29.0 to 32.7)	**20.2** (19.0 to 21.3)	**1.77** (1.58 to 1.97) P<0.0001
Worse on any 2 or more measures	**Total: 7455** Worse: 616 Not Worse: 6839	**-2.43** (8.26)	**-0.00** (7.71)	**P<0.0001**	**35.4** (31.6 to 39.3)	**22.6** (21.7 to 23.7)	**1.87** (1.57 to 2.23) P<0.0001

CI = confidence interval; EDSS = Expanded Disability Status Scale; LCLA = Low Contrast Letter Acuity with 2.5% contrast; PCS = Physical Component Summary; SD = standard deviation; SF-36 = Short Form-36.

Conclusions from Analysis of Aggregated Clinical Trial Data

Analyses were conducted to characterize the measurement properties; sensitivity; construct, convergent, and known group validity; and clinical meaningfulness of four performance measures – T25FW, 9HPT, LCLA, and SDMT – to permit use individually or combined into a multidimensional test battery as primary or co-primary outcome measures. The components of the proposed multidimensional test battery were assessed in relation both to the EDSS and self-reported measures of health-related quality of life and depression. These results, based on a database of 14 datasets comprising 12,776 participants, represent the largest pooled analysis of prospectively acquired clinical trial data in MS to date. The demographics of the pooled dataset largely reflect the type of patents historically enrolled in MS clinical trials, for which trials in RRMS have predominated.

The distributions of the T25FW and 9HPT were positively skewed, and both measures demonstrated floor effects. The potential for a high proportion of patients to perform these measure as well as can be performed by a healthy control can result in reduced ability to distinguish gradations of performance at the lower end of the scale (demonstrated by far left peaks in (Figure 1A and 1B). Baseline LCLA and SDMT scores were more normally distributed, without evidence of floor or ceiling effects. The T25FW, 9HPT, and LCLA showed no clear-cut evidence of practice effects. As is typical of most cognitive measures, the SDMT exhibited some practice effects, but these appeared not to affect the normality of the SDMT's frequency distribution. All four performance measures demonstrated good test-retest reliability, indicating they yield reproducible scores if there is no change in the subject's condition. As a result, changes in a score can be assumed due to the participant's condition rather than

measurement variability. These results support the advantageous measurement properties of the four performance measures.
These results provide a cautionary note regarding the population for which these measures will be most useful. The majority of participants represented in the pooled dataset had RRMS with relatively mild disability. In turn, the T25FW and 9HPT exhibited floor effects, which may explain the decreased sensitivity of three-month confirmed worsening of T25FW and 9HPT compared to EDSS. More sensitive tests may be needed in studies enrolling participants with mild gait and upper extremity impairments.[38]
At baseline, the T25FW and 9HPT had stronger correlations with the EDSS and with each other than with the other two performance measures. These results support the construct validity of the T25FW and 9HPT, as both are measures of physical functions that overlap with the EDSS in its lower range (EDSS 0–4.0) as seen in this population. In comparison, LCLA and SDMT correlated less strongly with EDSS and the other performance measures, supporting their additive value, to assess functions not captured by the other performance measures and EDSS. Compared to correlations at baseline, all the correlations for change from baseline-to-endpoint were much weaker. Cohen's kappa coefficients showed that the confirmed worsening events defined by the four performance measures were largely independent of those defined by EDSS. Taken together, these results suggest that the four performance measures assess overlapping but somewhat different aspects of disability and disability worsening than does the EDSS.
All four performance measures were worse in subjects with longer MS disease duration and with worse disability measured by EDSS, supporting known group validity. Exploratory analyses were undertaken to assess the clinical meaningfulness of worsening on the performance measures using the SF-36 PCS as an anchor. SF-36 PCS correlated moderately at baseline with T25FW, 9HPT, and SDMT and weakly with LCLA. SF-36 MCS and BDI correlated weakly with all four performance measures at baseline. Group aggregate changes from baseline-to-endpoint in the performance measures and self-report measures correlated weakly or not at all when directionality was not considered. However, importantly, for subjects experiencing confirmed worsening from baseline-to-endpoint on the T25FW, 9HPT, and SDMT, the SF-36 PCS was significantly worse. Similarly, for subjects who showed confirmed baseline-to-endpoint worsening on the T25FW, 9HPT, and SDMT, the proportions of subjects with a five-point worsening on SF-36 PCS, which is considered clinically meaningful were greater. The LCLA results mirrored these findings with non-significant trends. These results indicate that the T25FW, 9HPT, and SDMT assess clinically meaningful aspects of MS-related disability and that the proposed thresholds for clinically meaningful change for each are reasonable. The non-significant trend of concomitant worsening in the LCLA and SF-36 PCS provides some support for the clinical meaningfulness of seven letter change in LCLA.
These results confirm the advantageous measurement properties of the T25FW, 9HPT, LCLA, and SDMT and support their construct, convergent, and known group validity, and sensitivity, particularly when combined into a multidimensional test battery. The associations with established measures of disability (EDSS) and HRQoL (SF-36) indicate that they evaluate clinically meaningful aspects of MS-related disability. These findings support the use of these measures either alone or together as a multidimensional test battery as primary or key secondary endpoints in MS studies.

Floor or Ceiling Effects The frequency distributions of the T25FW and the 9HPT are positively skewed, mainly due to the practice of imputing very high scores for PwMS unable to perform one or both of these tests. Both measures also show some floor effects since scores tend to be clustered in the shorter time spans. This results in some loss of ability to distinguish gradations of performance at the lower end of the scale. However, both the T25FW and the 9HPT possess the ability to distinguish gradations of performance in the middle of the scale. Overall, the results indicate robust distributional characteristics for both scales.
LCLA scores were normally distributed with no evidence of floor or ceiling effects or skewing for LCLA (2.5%) and only a slight floor effect for LCLA (1.25%). In other words, there is a slight bunching of scores on LCLA (1.25%) at the lower end of the scale for the most severely impacted individuals. The results favor LCLA (2.5%) as a measure of visual function with the ability to distinguish a wide range to severity.
Baseline values for the SDMT show a near normal distribution with no evidence for either a floor or ceiling effect. In contrast the PASAT is not normally distributed but instead is negatively skewed with a pronounced ceiling effect. The result is that a high proportion of PwMS score near the high end of the PASAT, thereby constraining the ability of the PASAT to distinguish gradations of performance for less severely impaired individuals. The results therefore strongly support the utilization of the SDMT as a measure of processing speed.

Practice Effects The T25FW, 9HPT, and LCLA (2.5% and 1.25%) tend to worsen over time and show no clear-cut evidence of practice effects. This makes them excellent choices for detecting both disease progression and the positive effects of treatment.
Practice effects tend to be an issue for most cognitive measures, even if alternate forms are available. Both the SDMT and the PASAT are subject to practice effects. Practice effects are greater for the PASAT than for the SDMT and the practice effects seen with the SDMT have little effect on the distributional characteristics of SDMT scores. To some extent these practice effects can be attenuated by administering the test several times prior to baseline. From the standpoint of practice effects, the SDMT is clearly superior to the PASAT and therefore is likely to be more sensitive to change or treatment effects.

Reliability *Test-retest reliability analyses were conducted using data from stable patients over a period not exceeding 6 months from baseline. "Stable" was defined as absence of EDSS change. The test-retest reliability of the SDMT, PASAT, T25FW, 9HPT, and LCLA was estimated by calculating the intra-class correlation coefficient from a random effects linear regression analysis with a random subject effect and terms to account for practice effects (where needed). The results indicated that all of the measures have excellent test-retest reliability: SDMT (0.85), PASAT (0.86), 9HPT (0.84), and LCLA (2.5%) (0.88). The T25FW had an ICC slightly lower (0.71). This lower value was due in part to the use of imputed values for PwMS unable to walk. When these individuals were not included in the analysis, the ICC increased to 0.78. These analyses indicate that all of the measures produce scores that are reproducible over a period of $\leq$ 6 months assuming there is no actual change in the individual's condition. Therefore, over a similar period of time, the overwhelming proportion of variance in scores would be attributable to changes in an individual's MS rather than random variation.*

Construct Validity The T25FW and the 9HPT had stronger correlations with EDSS than with other measures and a strong correlation with one another. Correlations with the EDSS at baseline and the endpoint were higher than correlations for changes from baseline or changes from baseline to endpoint. These results support the construct validity of the T25FW and 9HPT given that they are physical measures of functions also captured to some extent by the EDSS and that they correlate with one another.

Results for LCLA (2.5% contrast) and LCLA (1.25% contrast) were strongly correlated with each other, while correlations with all other measures were modest. The strongest correlation was with the EDSS, which also includes a rating of visual function. Correlations with other measures for change from baseline or baseline to endpoint were small. These results support the construct validity of LCLA, which correlates modestly with physical and cognitive measures and more strongly with the EDSS.

Correlation coefficients between SDMT and other measures at both baseline and endpoint were statistically significant and in the expected directions, though modest, mostly in the range of 0.2 to 0.4. This shows that although patients more severely affected on one measure are also likely to be more severely affected on SDMT, the correlations are modest, so SDMT appears to be measuring something different. This confirms the construct validity of SDMT: although it is related to the other measures, it is not the same. Correlations between the PASAT and other measures tended to follow the same pattern as that for the SDMT, although the coefficients were smaller. Overall these analyses support the construct validity of both the SDMT and the PASAT as sharing some variance with the other, primarily physical measures but focusing largely on cognition. These results strongly support including SDMT as a cognitive test that provides important information about an important dimension of MS that is complementary to motor function.

Known Group Validity PwMS with longer MS duration had lower scores on both the SDMT and the PASAT, although these differences were more pronounced for the SDMT. PwMS with worse EDSS scores had lower scores on both the SDMT and the PASAT, with the SDMT showing larger differences. PwMS who were older had lower scores on both the SDMT and the PASAT and once again these differences were more pronounced for the SDMT. Altogether these results provide strong evidence for the known groups validity of the SDMT and support the selection of the SDMT over the PASAT as the cognitive measure for use in clinical trials.
The results for the T25FW, 9HPT, and LCLA (2.5% and 1.25%) were similar to those for the cognitive tests with worse scores associated with longer disease duration, worse EDSS, and higher age. Together these findings support the use of the SDMT, the T25FW, the 9HPT, and LCLA as a comprehensive and multifaceted suite of measures to track MS-related disability.Correlation of Performance Test Scores with EDSS Progression and Relapses (Table 9) summarizes the results for the each measure. As this table shows, results were mixed and varied among the measures. In populations, SDMT and the PASAT showed improvement over time, probably due to the presence of

practice effects for these two measures, as it seems unlikely that MS was getting less severe with time. The observed improvement over time cannot be interpreted as lack of cognitive worsening, since stability or slight improvement could represent masking of worsening by practice effects. SDMT scores improved during periods of time following relapse or EDSS progression events.
Not surprisingly, the T25FW and the 9HPT showed the strongest and most consistent correlations with relapse and EDSS progression events. This result was expected since these dimensions are typically affected by relapses and are aspects of EDSS worsening. These measures worsened with relapse or EDSS progression events and improved during the recovery phase.
LCLA (2.5%) was correlated with EDSS worsening but not relapses; while LCLA (1.25%) was correlated with both relapses and EDSS worsening. This is not surprising since LCLA was introduced to the clinical trial landscape in order to provide a measure that was more sensitive to visual disturbance than traditional measures often used in evaluating relapses and in the EDSS such as high contrast visual acuity.
The results indicate T25FW and 9HPT changes correlate in the expected direction with relapse and EDSS progression events, and with recovery from relapse and EDSS progression events. These findings are expected, because motor function is commonly affected by relapse, and is intrinsic to the EDSS scale. The cognitive and visual measures are not as strongly correlated with relapse or EDSS, suggesting that these dimensions provide complementary information to the traditional measures. This argues strongly for the need to incorporate the SDMT and LCLA as primary or co-primary endpoints in clinical trials. Use of the SDMT and LCLA would allow for the diagnosis of relapses that are currently missed using clinical diagnosis, the EDSS, the T25FW and 9HPT. Use of the SDMT and LCLA would also allow for the detection of differences in response to treatment that might otherwise be missed.

Table 9: Summary of the Results of the Analysis of Sensitivity to Change

Event	***SDMT***	***PASAT***	***T25FW***	***9HPT***	***LCLA 2.5%***	***LCLA 1.25%***
Relapse	*---*	*---*	*Yes*	*Yes*	*---*	*Yes*
Recovery from Relapse	*Yes*	*Yes*	*Yes*	*Yes*	*---*	*---*
EDSS Worsening	*Wrong Direction*	*---*	*Yes*	*Yes*	*Yes*	*Yes*
EDSS Improvement	*Yes*	*Yes*	*---*	*Yes*	*Yes*	*---*

Yes (in the expected direction and statistically significant); Wrong Direction (in the wrong direction and statistically significant); --- (not statistically significant)

Ability to Detect Worsening During the Course of a Clinical Trial *Confirmed worsening from baseline EDSS is the most common disability outcome used in contemporary clinical trials. In general, approximately 15-25% of patients in placebo arms of controlled clinical trials met this definition of disability worsening. The pooled database was interrogated to determine disability progression rates and compared with EDSS progression rates during the same interval. For SDMT, using the definition of confirmed 4-point worsening from baseline, 15% worsened on SDMT over 24 months, compared with 14.5% on EDSS. For T25FW, using the definition of confirmed 20% worsening from baseline, 6.5% worsened on T25FW over 24 months, compared with 20.2% on EDSS. For 9HPT, using the definition of confirmed 20% worsening from baseline, 2.9% worsened on 9HPT over 24 months, compared with 20.2%on EDSS. For LCLA (2.5%), using the definition of confirmed 20% worsening from baseline, 13.1% worsened on 9HPT in 24 months, compared with 16.1% on EDSS. Thus, progression rates for SDMT and LCLA (2.5%) were of similar magnitude to EDSS progression rates, and were somewhat lower for T25FW and 9HPT. Progression events defined with SDMT or LCLA (2.5%) are independent from EDSS progression, while progression events defined with T25FW and 9HPT are weakly associated with EDSS progression rates. This indicates that the four performance measures detect worsening in MS patients over the interval typical of a MS trial, and that patients with disability progression not detected by EDSS are identified by quantitative measures of cognition, vision, walking, and dexterity.*

Minimum Clinically Important Change or Difference Exploratory analysis of a clinically important change or difference in the candidate measures was undertaken using the PCS from the SF-36 as an anchor (Table 8). The PSC reflects several aspects of physical functioning, including limitations in work. It is generally recognized that a five-point change in the PCS represents a clinically meaningful difference or change. Based on evidence from the literature review, putative clinically important differences were proposed for the candidate measures: T25FW and 9HPT (20%), SDMT (4 points), LCLA (7 letters). The analyses examined the proportion of PwMS who experienced the aforesaid changes (20%, 4 points, 7 letters) and who also showed a 5-point worsening on the PCS. The T25W,

the 9HPT, and the SDMT all showed statistically significant differences in the proportion showing a 5-point worsening in the PCS between those with and without the respective putative differences.

SDMT had moderately strong and statistically significant correlations with the PCS of the SF-36 and smaller but statistically significant correlations with the MCS of the SF-36. The Mental Component Summary (MCS) of the SF-36 is not a good measure of cognitive functioning, as the items comprising the MCS are primarily related to affect and mood (refer to Hobart et al[48] for the list of the 36 items of the SF-36 measurement model). To determine the contribution of depression to the strength of the associations, analyses were conducted using the Beck Depression Inventory (BDI) data. Regression coefficients for the association between the SDMT and the MCS absolute values were substantially weaker after adjustment for Beck Depression Inventory (BDI), demonstrating that some of the association between the SDMT and the MCS can be explained by depression. On the other hand, BDI scores did not appreciably influence the correlation between the SDMT and the PCS. The regression coefficients were nearly identical whether unadjusted or adjusted for BDI for absolute values at baseline, and only changed to a small extent for absolute values at study end. These results indicate that the significant association between the SDMT and the PCS is not driven by depression and that cognitive dysfunction as measured by the SDMT is reflected in patient self-reports of quality of life as measured by the PCS.

While there are limited data from the MSOAC dataset to directly link SDMT to ADLs, there is abundant literature on this relationship, and the work of the Kessler Foundation has significantly expanded the available data to include common activities carried out in real world settings (e.g. ordering airline tickets on-line, preparing a meal, etc.). [41,42] *These studies control for other factors and demonstrate that processing speed has been the best overall indicator of cognitive functioning. The relationship between the SDMT and the PCS in the MSOAC database showed that PwMS who experienced a 4-point or greater decline in the SDMT were 42% more likely to also report a 5-point or greater worsening in the PCS, an amount of PCS change generally considered to be clinically important. Together these results provide strong support for the validity of the SDMT as a measure of cognitive functioning which is important to the quality of life of PwMS.*

Overall, these results provide support for the use of 20% for the T25FW and 9HPT and 4 points for the SDMT based on their relationship to a widely used PRO that measures health-related quality of life. Such a relationship indicates not only that these three scales are measuring functions that are important to PwMS but that the aforesaid differences in each measure represent a degree of change that is meaningful to PwMS. In summary, the results of the analysis of change for these three measures supports their use in clinical trials and the clinical meaningfulness of 20% for the T25FW and 9HPT and 4 points for the SDMT. Although this relationship was not observed for the LCLA, there is strong support in the literature defining a 7-letter decline in LCLA (2.5%) as clinically meaningful and the VOP Study also supports the importance of visual changes to PwMS.

Transferability of SDMT Data to All Forms of MS

Given that for the SDMT performance measure, the MSOAC database included only RRMS data, the following evidence from the literature is provided to support the sensitivity of the SDMT for SPMS and PPMS:

1. *Huijbregts*[43] *evaluated cognitive performance of patients with RRMS, SPMS, and PPMS compared to healthy controls using a wide variety of tests. They found that patients with all forms of MS performed significantly worse on the SDMT than controls. The mean scores for the SDMT were as follows: HC = 60.2; RRMS = 54.3; SPMS = 45.1; PPMS = 47.8. A 4-point difference in the SDMT is generally considered to be clinically meaningful.*

2. *Zaczanis*[44] *performed a meta-analysis of studies looking at cognitive performance in MS patients (N=1,845) compared to healthy controls (N=1,265). Results indicated that the SDMT was the most sensitive measure discriminating healthy controls from patients with both RRMS and chronic progressive MS (CPMS). Out of several dozen test scores examined by Zaczanis, the SDMT had the second largest effect size (-1.36) for both types of MS, second only to the Selective Reminding Test Delayed recall score.*

3. *Ruano*[45] *evaluated 1,040 patients including those with a diagnosis of clinically isolated syndrome (CIS), RRMS, PPMS, and SPMS. A wide range of assessments were utilized. Patients in all categories exhibited cognitive impairment ranging from 31.4% for CIS to 91.3% for PPMS. Information processing speed (incorporating the SDMT among other measures) was the most frequently affected cognitive domain with 47.9% affected.*

4. *Cognitive dysfunction as measured using the SDMT occurs in all types of MS, all durations, and all severities of physical disability. It has been utilized and studied in a wide variety of MS populations and cultures.[32] Moreover, it has been shown to be related to a variety of disease markers including central atrophy[46–48] and gray matter volume[49] in progressive and relapsing remitting MS. Of all cognitive tests that have been studied in MS, the SDMT has been shown to have the most robust relationship to important life activities such as employment and daily activities.[15]*

5. *Chow[50] assessed the degree to which QoL correlated with cognitive function as assessed using SDMT, PASAT, and the Trail Making Test-Part B in patients with PPMS and SPMS, and found that SDMT had the highest correlation with the SF36.*

Based on evidence from the literature and from our data, MSOAC suggests that SDMT reflects the similarities and commonalities of all MS types, in keeping with the recommendations of two international expert groups.[31,32] This type of evidence is reinforced and continues to emerge in other recent clinical trials in which SDMT was found to be sensitive to treatment effects in both RRMS and in secondary progressive MS. Kappos[51] recently published results from a trial of siponimod in SPMS. In a double-blind, placebo-controlled RCT study siponimod reduced EDSS disability progression in SPMS by 21%. Using data from the same study in a post-hoc analysis, Benedict reported a 21% reduction in the probability of $\geq$ 4 point worsening on SDMT (p = 0.015). This suggests that the treatment effect of siponimod for cognitive worsening is of similar magnitude to EDSS worsening in progressive MS, and supports use of SDMT in progressive forms of MS. Finally, 4 point change in SDMT following treatment with ocrelizumab was documented in RRMS patients.[52]

OVERALL CONCLUSIONS

MSOAC's goal is to achieve qualification of performance measures that possess the following characteristics: 1) the measure should address a common and important dimension of disability in the disease; 2) the dimension can be objectively quantified by a trained observer, 3) the measure has favorable psychometric properties, including precision and validity, 4) the measure reflects functional change that is perceived by the patients as important, and 5) the measure is practical, non-invasive, and acceptable to patients. MSOAC concludes that evidence from the literature and the newly generated evidence from analysis of contemporary RCTs from the MSOAC database are remarkably consistent in documenting excellent psychometric properties for T25FW, 9HPT, LCLA (2.5%), and SDMT, and for establishing validity as measures of four important dimensions commonly impacted in PwMS (walking, manual dexterity, vision, and cognition). Further, evidence collected for the MSOAC project in the VOP Study, which obtained data directly from PwMS, documented the importance of these four dimensions directly from the affected individuals. Thus, the combined data obtained for each performance measure from three robust sources of data – the comprehensive literature review, the VOP Study, and the large MSOAC clinical trial database – provide a strong preponderance of evidence for each of these 4 measures in future clinical trials.

MSOAC believes that the qualification of a COA instrument that measures multiple functional domains will enable sponsors to test potential disease-modifying interventions and accelerate the pace of clinical research for MS, particularly for people with MS who have a progressive disease course. A qualified COA instrument for cognition, ambulation, dexterity and vision, potentially used in combination with currently accepted clinical and patient-reported outcome measures, will provide a comprehensive approach to capturing disability in MS. The heterogeneity in symptoms reported both between PwMS and over time with individual PwMS, which stems from autoimmune attacks in different parts of the nervous system, calls for an instrument that measures disability in a number of domains. The MSOAC Members unanimously endorse these four measures as acceptable primary disability outcome measures for MS clinical trials.

ACKNOWLEDGMENTS
MSOAC gratefully acknowledges the longstanding, collective efforts of the 23 academic investigators, 12 companies, and 4 advocacy organizations in sharing data and expertise, starting with the launch of the consortium on April 1, 2013. Several colleagues from EMA and FDA served as regulatory liaisons and advisors. A list of the members is published.[1]

REFERENCES

1. LaRocca NG, Hudson LD, Rudick R, Amtmann D, Balcer L, Benedict R, et al. The MSOAC approach to developing performance outcomes to measure and monitor multiple sclerosis disability. Mult Scler Houndmills Basingstoke Engl. 2017 Aug 1;1352458517723718.

2. Lublin FD, Reingold SC, Cohen JA, Cutter GR, Sørensen PS, Thompson AJ, et al. Defining the clinical course of multiple sclerosis: the 2013 revisions. Neurology. 2014 Jul 15;83(3):278–86.

3. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology. 1996 Apr;46(4):907–11.

4. Heesen C, Böhm J, Reich C, Kasper J, Goebel M, Gold S. Patient perception of bodily functions in multiple sclerosis: gait and visual function are the most valuable. Mult Scler J. 2008 Aug 1;14(7):988–91.

5. Larocca NG. Impact of walking impairment in multiple sclerosis: perspectives of patients and care partners. The Patient. 2011;4(3):189–201.

6. Motl RW. Ambulation and multiple sclerosis. Phys Med Rehabil Clin N Am. 2013 May;24(2):325–36.

7. Kieseier BC, Pozzilli C. Assessing walking disability in multiple sclerosis. Mult Scler J. 2012 Jul 1;18(7):914–24.

8. Goodman AD, Brown TR, Cohen JA, Krupp LB, Schapiro R, Schwid SR, et al. Dose comparison trial of sustained-release fampridine in multiple sclerosis. Neurology. 2008 Oct 7;71(15):1134–41.

9. Goodman AD, Brown TR, Edwards KR, Krupp LB, Schapiro RT, Cohen R, et al. A phase 3 trial of extended release oral dalfampridine in multiple sclerosis. Ann Neurol. 2010 Oct 1;68(4):494–502.

10. Kister I, Bacon TE, Chamot E, Salter AR, Cutter GR, Kalina JT, et al. Natural History of Multiple Sclerosis Symptoms. Int J MS Care. 2013 Oct 1;15(3):146–56.

11. Lamers I, Kelchtermans S, Baert I, Feys P. Upper Limb Assessment in Multiple Sclerosis: A Systematic Review of Outcome Measures and their Psychometric Properties. Arch Phys Med Rehabil. 2014 Jun 1;95(6):1184–200.

12. Feys P, Lamers I, Francis G, Benedict R, Phillips G, LaRocca N, et al. The Nine-Hole Peg Test as a manual dexterity performance measure for multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2017 Apr;23(5):711–20.

13. Motl RW, Cohen JA, Benedict R, Phillips G, LaRocca N, Hudson LD, et al. Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2017 Apr;23(5):704–10.

14. Balcer LJ, Raynowska J, Nolan R, Galetta SL, Kapoor R, Benedict R, et al. Validity of low-contrast letter acuity as a visual performance outcome measure for multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2017 Apr;23(5):734–47.

15. Benedict RH, DeLuca J, Phillips G, LaRocca N, Hudson LD, Rudick R. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2017 Apr;23(5):721–33.

16. Orrell RW. Multiple Sclerosis: The History of a Disease. J R Soc Med. 2005 Jun;98(6):289.

17. Goldman MD, Motl RW, Rudick RA. Possible clinical outcome measures for clinical trials in patients with multiple sclerosis. Ther Adv Neurol Disord. 2010 Jul;3(4):229–39.

18. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology. 1983 Nov;33(11):1444–52.

19. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain J Neurol. 2000 May;123 (Pt 5):1027–40.

20. Pearson M, Dieberg G, Smart N. Exercise as a therapy for improvement of walking ability in adults with multiple sclerosis: a meta-analysis. Arch Phys Med Rehabil. 2015 Jul;96(7):1339-1348.e7.

21. Yozbatıran N, Baskurt F, Baskurt Z, Ozakbas S, Idiman E. Motor assessment of upper extremity function and its relation with fatigue, cognitive function and quality of life in multiple sclerosis patients. J Neurol Sci. 2006 Jul 15;246(1):117–22.

22. Bertoni R, Lamers I, Chen CC, Feys P, Cattaneo D. Unilateral and bilateral upper limb dysfunction at body functions, activity and participation levels in people with multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2015 Oct;21(12):1566–74.

23. Mendoza JE, Apostolos GT, Humphreys JD, Hanna-Pladdy B, O'Bryant SE. Coin rotation task (CRT): a new test of motor dexterity. Arch Clin Neuropsychol Off J Natl Acad Neuropsychol. 2009 May;24(3):287–92.

24. Balcer LJ, Baier ML, Cohen JA, Kooijmans MF, Sandrock AW, Nano-Schiavi ML, et al. Contrast letter acuity as a visual component for the Multiple Sclerosis Functional Composite. Neurology. 2003 Nov 25;61(10):1367–73.

25. Balcer LJ, Baier ML, Pelak VS, Fox RJ, Shuwairi S, Galetta SL, et al. New low-contrast vision charts: reliability and test characteristics in patients with multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2000 Jun;6(3):163–71.

26. Fisher JB, Jacobs DA, Markowitz CE, Galetta SL, Volpe NJ, Nano-Schiavi ML, et al. Relation of Visual Function to Retinal Nerve Fiber Layer Thickness in Multiple Sclerosis. Ophthalmology. 2006 Feb 1;113(2):324–32.

27. Charcot JM (Jean M, Sigerson G. Lectures on the diseases of the nervous system [Internet]. London, New Sydenham Society; 1881 [cited 2019 Apr 10]. 530 p. Available from: http://archive.org/details/lecturesondiseas00char

28. Rao SM, St Aubin-Faubert P, Leo GJ. Information processing speed in patients with multiple sclerosis. J Clin Exp Neuropsychol. 1989 Aug;11(4):471–7.

29. Peyser JM, Rao SM, LaRocca NG, Kaplan E. Guidelines for Neuropsychological Research in Multiple Sclerosis. Arch Neurol. 1990 Jan 1;47(1):94–7.

30. Gronwall DM. Paced auditory serial-addition task: a measure of recovery from concussion. Percept Mot Skills. 1977 Apr;44(2):367–73.

31. Kalb R, Beier M, Benedict RH, Charvet L, Costello K, Feinstein A, et al. Recommendations for cognitive screening and management in multiple sclerosis care. Mult Scler Houndmills Basingstoke Engl. 2018 Nov;24(13):1665–80.

32. Sumowski JF, Benedict R, Enzinger C, Filippi M, Geurts JJ, Hamalainen P, et al. Cognition in multiple sclerosis: State of the field and priorities for the future. Neurology. 2018; 90(6):278–88.

33. Strober LB, Christodoulou C, Benedict RH, Westervelt HJ, Melville P, Scherl WF, et al. Unemployment in multiple sclerosis: the contribution of personality and disease. Mult Scler J. 2012 May 1;18(5):647–53.

34. Strober L, Chiaravalloti N, Moore N, DeLuca J. Unemployment in multiple sclerosis (MS): utility of the MS Functional Composite and cognitive testing. Mult Scler Houndmills Basingstoke Engl. 2014 Jan;20(1):112–5.

35. Honan CA, Brown RF, Batchelor J. Perceived Cognitive Difficulties and Cognitive Test Performance as Predictors of Employment Outcomes in People with Multiple Sclerosis. J Int Neuropsychol Soc. 2015 Feb;21(2):156–68.

36. Fisher J, Jak A, Knicker J, Rudick R, Cutter G. Multiple Sclerosis Functional Composite (MSFC). Administration and Scoring Manual. National Multiple Sclerosis Society. 2001.

37. Winer BJ. Statistical principles in experimental design. McGraw-Hill; 1971. 936 p.

38. Cohen JA, Reingold SC, Polman CH, Wolinsky JS, International Advisory Committee on Clinical Trials in Multiple Sclerosis. Disability outcome measures in multiple sclerosis clinical trials: current status and future prospects. Lancet Neurol. 2012 May;11(5):467–76.

39. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003 May;41(5):582–92.

40. Strober L, DeLuca J, Benedict RH, Jacobs A, Cohen JA, Chiaravalloti N, et al. Symbol Digit Modalities Test: A valid clinical trial endpoint for measuring cognition in multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2018 Oct 18;1352458518808204.

41. Goverover Y, Strober L, Chiaravalloti N, DeLuca J. Factors That Moderate Activity Limitation and Participation Restriction in People With Multiple Sclerosis. Am J Occup Ther Off Publ Am Occup Ther Assoc. 2015 Apr;69(2):6902260020p1-9.

42. Goverover Y, Haas S, DeLuca J. Money Management Activities in Persons With Multiple Sclerosis. Arch Phys Med Rehabil. 2016;97(11):1901–7.

43. Huijbregts SCJ, Kalkers NF, de Sonneville LMJ, de Groot V, Reuling IEW, Polman CH. Differences in cognitive impairment of relapsing remitting, secondary, and primary progressive MS. Neurology. 2004 Jul 27;63(2):335–9.

44. Zakzanis KK. Distinct Neurocognitive Profiles in Multiple Sclerosis Subtypes. Arch Clin Neuropsychol. 2000 Feb 1;15(2):115–36.

45. Ruano L, Portaccio E, Goretti B, Niccolai C, Severo M, Patti F, et al. Age and disability drive cognitive impairment in multiple sclerosis across disease subtypes. Mult Scler J. 2017 Aug 1;23(9):1258–67.

46. Benedict RHB, Weinstock-Guttman B, Fishman I, Sharma J, Tjoa CW, Bakshi R. Prediction of Neuropsychological Impairment in Multiple Sclerosis: Comparison of Conventional Magnetic Resonance Imaging Measures of Atrophy and Lesion Burden. Arch Neurol. 2004;61(2):226–30.

47. Christodoulou C, Krupp LB, Liang Z, Huang W, Melville P, Roque C, et al. Cognitive performance and MR markers of cerebral injury in cognitively impaired MS patients. Neurology. 2003 Jun 10;60(11):1793–8.

48. Cocozza S, Petracca M, Mormina E, Buyukturkoglu K, Podranski K, Heinig MM, et al. Cerebellar lobule atrophy and disability in progressive MS. J Neurol Neurosurg Psychiatry. 2017 Dec;88(12):1065–72.

49. Sanfilipo MP, Benedict RHB, Weinstock-Guttman B, Bakshi R. Gray and white matter brain atrophy and neuropsychological impairment in multiple sclerosis. Neurology. 2006 Mar 14;66(5):685.

50. Højsgaard Chow H, Schreiber K, Magyari M, Ammitzbøll C, Börnsen L, Romme Christensen J, et al. Progressive multiple sclerosis, cognitive function, and quality of life. Brain Behav [Internet]. 2018 Jan 5 [cited 2019 Apr 10];8(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5822575/

51. Kappos L, Bar-Or A, Cree BAC, Fox RJ, Giovannoni G, Gold R, et al. Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. The Lancet. 2018 Mar 31;391(10127):1263–73.

52. Benedict RH, Cree B, Tomic D, Fox R, Giovannoni G, Bar-Or A, et al. Impact of Siponimod on Cognition in Patients With Secondary Progressive Multiple Sclerosis: Results From Phase 3 EXPAND Study (S44.004). Neurology. 2018;90(15 Supplement):S44.004.