Title: Walking measures to evaluate assistive technology for foot drop in multiple sclerosis: A systematic review of psychometric properties.

Abstract (249 words)

Background

Foot drop in people with multiple sclerosis (pwMS) often managed with assistive technologies, such as functional electrical stimulation and ankle foot orthoses. No evidence synthesis exists for the psychometric properties of outcomes used to evaluate the efficacy of these interventions.

Objective

This systematic review aimed to identify the outcome measures reported to assess the benefits of assistive technology for pwMS and then synthesize the psychometric evidence in pwMS for a subset of these measures.

Methods

Two searches in eight databases were conducted up to May 2017. Methodological quality was rated using the COSMIN guidelines. Overall level of evidence was scored according to the Cochrane criteria.

Results

The first search identified 27 measures, with the 10m walk test, gait kinematics and Physiological Cost Index (PCI) most frequently used. The second search resulted in 41 studies evaluating 10 measures related to walking performance. Strong levels of evidence were found for the internal consistency and test-retest reliability of the Multiple Sclerosis Walking Scale-12 and for the construct validity for Timed 25 Foot Walk. No psychometric studies were identified for gait kinematics and PCI in pwMS. There was a lack of evidence for measurement error and responsiveness.

Conclusion

Although a strong level of evidence exists for some measures included in this review, there was an absence of psychometric studies on commonly used measures such as gait kinematics. Future psychometric studies should evaluate a wider range of walking related measures used to assess the efficacy of interventions to treat foot drop in pwMS.

Key words

multiple sclerosis, COSMIN, walking performance, psychometrics, FES, assistive technology

Introduction

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system that typically strikes adults [1]. There is a wide variability among the symptoms, with gait impairments being one of the most common [2]. People with MS (pwMS) rate the impairment of their gait as being an inhibiting factor in their everyday life, sometimes even in relatively early stages of the disease [3,4].

One of the most common gait impairments is foot drop, which is the reduced dorsiflexion of the ankle during the swing phase of gait, potentially leading to trips or falls. Foot drop can be caused by weakness of the dorsiflexor muscles, impaired neural control causing co-contraction of agonist and antagonist muscles and increased tone in the plantarflexor muscles [5]. In pwMS foot drop can also be caused by increased motor fatigability, which is described as the exercise-induced reduction in the ability of the muscles to produce force or power [6]. Two common interventions to treat foot drop are functional electrical stimulation and ankle foot orthoses. The most commonly used ankle foot orthoses restrain the movement of the foot and thus reduce foot drop, but they do not allow active control of the ankle, which may result in an abnormal gait pattern [7]. On the contrary, functional electrical stimulation involves electrical stimulation that is applied to the common peroneal nerve, eliciting the desired contraction to produce ankle dorsiflexion during the swing phase of gait. The advantage of functional electrical stimulation is that it facilitates a more normal gait pattern, increases walking speed and decreases the physiological cost of gait [8,9].

The effects of functional electrical stimulation and ankle foot orthoses on walking performance is currently evaluated via a wide variety of outcome measures including, for example, timed

walking tests [e.g. 6-Minute Walk test (6MWT), Timed 10-Meter Walk (10mWT), Timed 25 Foot Walk (T25FW)] or patient or clinician reported instruments and rating scales [e.g. Multiple Sclerosis Walking Scale-12 (MSWS-12), Hauser Ambulation Index (HAI), Dynamic Gait Index (DGI)]. Instrumental motion analysis techniques are also used to objectively quantify the gait pattern. A comprehensive assessment of three-dimensional kinematics and kinetics can reveal minimal changes that cannot be observed visually [10]. For this reason, three-dimensional (3D) gait analysis is widely used to discriminate between normal and abnormal gait patterns and to evaluate responses to interventions in a variety of populations, such as stroke [11], cerebral palsy [12] and Parkinson's disease [13,14].

The outcome measures used to assess the efficacy of interventions such as assistive technology to treat foot drop need to be valid, reliable and responsive to change. Several studies have evaluated the psychometric properties of outcome measures used to assess the effects of ankle foot orthoses and functional electrical stimulation to treat foot drop (e.g. Goldman et al [15], Stellman et al [16], Learmonth et al [17,18]). However, no systematic review exists that has evaluated both the evidence and the methodological quality of studies describing the psychometric properties of such outcome measures.

We, therefore, aimed to (i) identify studies that evaluated the effects of ankle foot orthoses and functional electrical stimulation in pwMS and then (ii) synthesize the available psychometric evidence for the designated subset of, walking performance, effort of walking and lower limb function, outcome measures identified. In so doing, we hoped to augment the evidence-base available to optimize the appropriate selection of outcome measure(s) to evaluate the efficacy of assistive technology to treat foot drop in pwMS.

Methods

First search: overview of outcome measures

The purpose of the first search of the literature was to identify those studies that assessed the effects of either functional electrical stimulation or ankle foot orthoses used to treat foot drop in pwMS. From these studies we identified the outcome measures used and the frequency of their use.

Search strategy and study selection

A comprehensive search of eight databases, including MEDLINE (1963-5/2017), CINAHL (1969-5/2017), EMBASE (1974-5/2017), SCOPUS (1963-5/2017), PsycINFO (1963-5/2017), AMED (1967-5/2017), SPORTDiscus (1963-5/2017) and Web of Science (1967-5/2017) was conducted in order to identify the articles that met the inclusion criteria. The search strategy included synonyms and keywords for functional electrical stimulation (e.g. 'Functional Electrical Stimulation', 'foot drop stimulation' and 'common peroneal stimulation') and ankle foot orthoses (e.g. 'Ankle Foot Orthoses' and 'splints') and the population of interest (e.g. 'multiple sclerosis' and 'demyelinating disease'). The full strategy has been included as supplementary material. The inclusion criteria for this search were: a) studies that have assessed the use of functional electrical stimulation or ankle foot orthoses to treat foot drop in pwMS and b) studies that included outcome measures that evaluate function, walking performance, fatigue and quality of life (QoL). The exclusion criteria were: a) studies that used other forms of electrical stimulation (i.e. not functional) and those that evaluated orthoses for other joints than the ankle, b) studies that were reviews (i.e. systematic, meta-analysis, etc.), conference abstracts and editorials and c) studies in languages other than English, Greek or Dutch.

Two independent researchers (GA, MvdL) were involved in the screening of the articles for inclusion. After exclusion of irrelevant articles based on the titles and abstracts, the full-text of the remaining articles was examined for their eligibility. Reference lists of articles included in the review were searched for potentially relevant articles that were not retrieved in the original search. If any differences in opinion existed, consensus was made through discussion and a third reviewer (TM) was available if consensus between the primary two reviewers was not reached. From the eligible articles, we extracted the outcome measures that were employed to assess the effects of functional electrical stimulation or ankle foot orthoses and recorded the frequency of these measures being used.

Principal search: systematic review of the psychometric properties of outcome measures

The second and principal search was conducted to identify studies that evaluated the psychometric properties of outcome measures that assess walking performance, effort of walking and lower limb function in pwMS.

Search strategy and study selection

A similar protocol for the second search was followed as the one described above. A comprehensive search of MEDLINE (1976-5/2017), CINAHL (1995-5/2017), SCOPUS (1999-5/2017), EMBASE (1974-5/2017), PsycINFO (1963-5/2017), AMED (1967-5/2017), SPORTDiscus (1963-5/2017) and Web of Science (1967-5/2017) databases was conducted by combining the outcome measures of walking performance, effort of walking and lower limb function which were identified in the first search. The search strategy included keywords and synonyms of the population of interest (see first search), a subset of the identified outcome measures (e.g. ‘3D gait analysis’, ’10m walk test’, etc.) and a search filter for identifying studies

evaluating measurement properties, developed by Terwee et al [19]. The full search strategy is included as supplementary material.

The inclusion criteria for our second search were: studies that assessed the psychometric properties of a subset of the outcomes identified in the first search, namely those assessing walking performance, lower limb function and effort of walking. Although we acknowledge the importance of outcome measures such as QoL and fatigue, we decided to restrict the outcome measures in this review to those measures that are potentially directly affected by the use of functional electrical stimulation and ankle foot orthoses. Further, the psychometric evidence for fatigue measures used in MS has been the subject of a previous review [20]. The exclusion criteria were: a) studies that were reviews (e.g. systematic and meta-analyses), abstracts from conferences or editorials, and b) full texts in peer reviewed journals published in languages other than English, Greek or Dutch. The procedures used to select the final set of papers were the same as those described for the first search.

Methodological quality

The methodological quality of the studies identified in the second search was assessed using the Consensus-based Standards for the selection of Health Measurement Instruments (COSMIN). We chose the COSMIN checklist since is used to obtain a score for the methodological quality of a study evaluating one or more measurement properties of a particular outcome measure [21,22]. The COSMIN checklist has been assessed for the inter-rater agreement and reliability of each item, with the percentage agreement being appropriate, but the kappa coefficients for each item being relatively low [23]. However, to overcome low inter-rater agreement in scoring items, we familiarized with the grading process and developed specific guidelines as recommended by the developers of COSMIN. The COSMIN-checklist consists of nine boxes (internal consistency,

reliability, measurement error, content validity, structural validity, construct validity, cross-cultural validity and responsiveness) with each box including 5-18 items. The reviewer selects the measurement properties evaluated in the study and scores the specific item-lists with 'poor', 'fair', 'good' or 'excellent' depending on the design and execution. The lowest score from the rated items determines the methodological quality of the measurement property [24]. Two reviewers (GA, MvdL) used the COSMIN checklist to rate the methodological quality of the measurement properties in all studies. Any disagreements in ratings were resolved through discussion.

As previously mentioned, in order to be consistent in our ratings we developed guidelines for the rating of specific questions/items in each of the measurement properties in the COSMIN checklist. For example, all studies that used the EDSS as a comparator instrument were rated under the measurement property of construct validity, even if the authors stated that criterion validity was assessed. The questions for missing items and how they were handled was scored as 'not applicable' for measures that were not self-reported scales. For studies assessing within-day test-retest reliability, the items for patients being stable and the time interval being appropriate were rated as excellent.

The quality of the results of the psychometric properties of the outcome measures was assessed using the quality criteria by Terwee et al [25], which were recently revised by the authors [26]. The quality of the results of the psychometric properties was rated as 'positive' (+), 'indeterminate' (?) or 'negative' (-) depending on the methods and results of the studies (Table 1).

Data synthesis

The overall level of evidence for each outcome measure was reported according to the recommendations of the Cochrane Back Review Group. This overall score was given in relation to the methodological quality of the study and the results of the measurement properties. The evidence was rated as ‘strong’ (consistent (positive) findings in multiple studies of good methodological quality or in one study of excellent methodological quality), ‘moderate’ (consistent (positive) findings in multiple studies of fair methodological quality or in one study of good methodological quality), ‘limited’ (one study of fair methodological quality), ‘conflicting’ (both positive and negative findings), ‘unknown’ (only studies of poor methodological quality) [27]. For instance, if the intra-rater reliability for a particular outcome measure had one study of poor quality and one of good quality showing positive results, the overall score was ‘moderate’. Likewise, if there were four studies of fair methodological quality but only one with having a positive score for the quality of the results, the overall score was ‘limited’.

Results

First search: overview of outcome measures

After a systematic search of the eight databases, we retrieved 1393 titles for screening according to our inclusion criteria (Figure 1). We retained 34 articles and identified 27 outcome measures evaluating lower limb function, walking performance, effort of walking, fatigue and QoL. These outcomes measures were either self-reported measures [seven measures e.g. Fatigue Severity Scale (FSS), MSWS-12] or objective assessments [20 measures e.g. 6MWT, Multiple Sclerosis Functional Composite (MSFC), spatiotemporal gait parameters]. The most frequently used outcome measures were walking speed (mostly recorded over 10 meter distance), 3D gait kinematics and the Physiological Cost Index (PCI) (Figure 2).

Principal search: systematic review of the psychometric properties of outcome measures

Description of included studies

The systematic search of eight databases resulted in the identification of 2488 potentially relevant titles. After independent screening according to the inclusion and exclusion criteria, we retained 36 articles with further reference citation tracking resulting in five additional articles (Figure 1). Four studies [28-31] were excluded at the full-text screening stage because they aimed at validating a previously reported minimal clinically important difference (MCID) or cut-off points for a certain outcome measure and did not validate the outcome measure itself. Although of interest, the methodology of these studies is different from those reporting the psychometric properties of the outcome measures themselves and are therefore not appropriate to be assessed using the COSMIN checklist and Terwee criteria. In total, we included 41 articles reporting the psychometric properties of 10 outcome measures [MSFC, MSWS-12, spatiotemporal parameters, 10mWT, T25FW, 2 Minute Walk Test (2MWT), 6MWT, Rate of Perceived Exertion (RPE), peak oxygen uptake (VO_2 peak) & reaction time/movement time (RT/MT)] which all have been used to assess the effects of assistive technology to treat foot drop. Using the COSMIN taxonomy the following measurement properties were evaluated: reliability was assessed in 18 studies [(intra-rater n=3; inter-rater n=3; test-retest n=14), 8 outcome measures], measurement error in four studies (six outcome measures) and internal consistency in six studies (one outcome measure). Hypothesis testing/construct validity was evaluated in 15 studies (nine outcome measures) and responsiveness was assessed in 15 studies (seven outcome measures). Most studies assessed the MSWS-12 (n=12), followed by the 6MWT (n=11) and the T25FW (n=11). The agreement between the two raters (GA & MvdL) in the items of all the measurement

properties was 94.8% and for the final scores of each property the agreement was 94.7%. Upon discussion any disagreement regarding the rating of the items or the total score of each property was resolved. Studies included pwMS with relapsing remitting (RR), secondary progressive (SP), primary progressive (PP) and clinically isolated syndrome (CIS) with EDSS levels ranging from 0-8.5, with some studies not reporting this information [32-36]. The majority of the studies reported a mean of EDSS of four or more and five studies reported a mean EDSS of six [17,37-40]. The sample size was 6796 in total for the 41 studies, with the number of females (n=2109) exceeding the number of males (n=972) and with some studies not reporting the gender of the participants [35-34,41-42]. Table 2 presents an overview of the results together with the COSMIN rating and the rating of the quality of the results according to the revised Terwee criteria [24-26].

Methodological quality and strength of evidence

Reliability

The methodological quality of the studies was rated according to the COSMIN checklist as 'good' (n=3) [18,33,43], as 'fair' (n=3) [15,34,44] and 'poor' (n=12) [17,35-36,38,40,45-51]. The main reasons for a lower score included not reporting the intraclass correlation coefficient (ICC) or weighted Kappa, not describing the ICC model used, small sample size and the lack of an explicit statement that the repeated measurements were independent. Using the revised Terwee quality criteria [25-26], the evidence for reliability in 13 studies (seven outcome measures) [15,17-18,33-34,36,43-45,47-48,50-51] were rated as 'positive' and the remaining five (four outcome measures) [35,38,40,46,49] were rated as 'indeterminate' because neither ICC nor weighted Kappa were reported. From the eight outcome measures that were evaluated for

reliability (intra- & inter-rater, test-retest), seven of them demonstrated good and excellent values of ICC ranging from 0.86-0.96 and only for RPE the ICC values were moderate (0.706).

Measurement error

Of the four studies that evaluated measurement error (six outcome measures), the methodological quality of three [17,48,50] was rated as 'poor' due to a small sample size (n<30) and due to testing conditions not being similar. The methodology in one study [18] was rated as 'fair' because it was unclear whether the patients were stable in the interim period. The quality of the results for measurement error in all four studies was rated as indeterminate ('?') because in none of the studies the Minimal Important Change (MIC) values was reported, which is required to interpret whether the measurement error is acceptable [52].

Internal consistency

There were six studies that evaluated internal consistency. The methodological quality of four [33-34,53-54] was rated as 'excellent', for one [55] it was rated as 'good' and one [56] as 'poor' due to a small sample size. All six studies evaluated the MSWS-12 and were rated as positive ('+') for the quality of their results.

Hypothesis testing/construct validity

Fifteen studies assessed construct validity, with only one [40] with an 'excellent' methodological quality and one in which was rated as 'good' [57]. The methodological quality of seven [33,41,43,55,58-60] was rated as 'fair' either because the hypotheses were vague or due to limited information regarding the comparator instruments and its psychometric properties. The other studies [16,37,51,56,61-62] were rated as 'poor' due to a small sample size or the absence of information regarding the comparator instruments. Applying the Terwee quality criteria [25-

26], the quality of the results reported in 10 studies [16,33,37,40,43,55-59] was rated as positive ('+') and in five studies [41,51,60-62] as indeterminate ('?') as the correlations presented were with unrelated constructs. The construct validity of seven of the studies reporting on six laboratory based measures and one self-perceived scale of walking performance used the EDSS as a comparator instrument. The comparator instrument in other studies were outcome measures such as the MSWS-12, Multiple Sclerosis Impact Scale-29 (MSIS-29), accelerometry and O_2 cost of walking. Table 2 includes information regarding the comparator instruments and correlation coefficients presented in studies assessing construct validity.

Responsiveness

Responsiveness was evaluated in 15 studies, with the methodological quality of nine [32-33,39,41,63-67] rated as 'fair' and the remainder classed as 'poor' [35,38,42,60,68-69]. Most of the studies had a vague hypothesis or did not use appropriate statistical methods and this lowered their rating. Only four studies [39,64-66] investigating the MSWS-12, 10mWT, T25FW, 2MWT and 6MWT received a 'positive' rating and the remaining 11 [32-33,35,38,41-42,60,63,67-69] were rated as 'indeterminate' due to correlations with unrelated constructs or the lack of differences between relevant groups. Of the 15 studies evaluating responsiveness, only two studies [39,65] reported on the MCID (MIC) for the 10mWT, T25FW and 6MWT.

Level of evidence – data synthesis

The overall levels of evidence for the psychometric properties of each outcome measure are summarized in Table 3. It was found that the MSWS-12 has strong positive evidence for its internal consistency and test-retest reliability, moderate positive evidence for its construct validity when compared to MSIS-29 and O_2 cost of walking and limited positive evidence for its

responsiveness. The MSFC showed moderate positive evidence for its intra-rater reliability and construct validity, while for the remaining measurement properties, including responsiveness, the evidence was 'unknown'. For lower limb reaction/movement time, there was limited positive evidence for construct validity, but for responsiveness the evidence was 'unknown'. Strong evidence was found for the construct validity of the T25FW while for responsiveness and for test-retest reliability the evidence was moderately positive. Spatiotemporal parameters were classed as having a limited positive level of evidence for construct validity. For the 10mWT the level of evidence for its responsiveness was moderately positive, while for the other measurement properties this was 'unknown'. Limited positive evidence was found for the responsiveness of the 2MWT. For the 6MWT, the level of evidence for responsiveness and test-retest reliability was moderately positive, while the evidence for the inter-rater reliability was limited positive. The level of evidence for the measurement properties assessed for VO_2 peak and RPE were all 'unknown'.

Discussion

The first search of the present systematic review identified 27 outcome measures, assessing self-reported and objectively measured walking performance, self-perceived fatigue, effort of walking, QoL, balance, falls and lower limb function, that had been used in studies assessing the effects of assistive technology to treat foot drop in pwMS. The most frequently used measure was the 10mWT (n=19), followed by 3D gait kinematics (n=12) and PCI (n=10). Interestingly, although 3D gait kinematics was one of the most frequently used outcome measures to assess the effects of assistive technology to treat foot drop, its psychometric properties have not yet been reported for this specific population [10]. Similarly, there were no psychometric studies identified for PCI for the MS population. However, studies into the psychometric properties for 3D gait

kinematics have demonstrated that 3D gait analysis is a reliable, valid and responsive tool for characterizing gait in stroke sufferers [70], CP [71-73] and many musculoskeletal disorders [74-75]. Similarly, the construct validity of the PCI has been assessed in the subacute stroke population and its reliability documented in children with cerebral palsy [76-77].

The second, and main, search for studies assessing the psychometric properties of the 20 outcome measures related to walking performance, lower limb function and effort of walking identified in the first search, revealed 41 studies that evaluated only 10 of these twenty outcomes. Of those 10 measures, the MSWS-12 was found to have a strong level of evidence for its internal consistency and test-retest reliability and the T25FW for construct validity. Moderate evidence was found for the test-retest reliability and responsiveness of the 6MWT and the responsiveness of the 10mWT.

Short distance walking tests, such as the 10mWT and T25FW have been classified as reliable owing to ICC values of 0.7 and over. However, there are indications that walking speed, as measured over such short distances, may not be appropriate to assess the benefits of functional electrical stimulation for community walkers with relatively low levels of disability. For example, Miller et al [78] found that pwMS who walked faster than 0.8m/s did not increase their walking speed in the T25FW with the assistance of functional electrical stimulation, while those with a slower walking speed than 0.8m/s did.

De Vet et al [79] distinguished two aspects of reliability, namely consistency (or relative reliability), which is assessed by the ICC and secondly measurement error (or absolute reliability), which is reported by measures like standard error of measurement (SEM), minimal detectable change (MDC) and the Limits of Agreement (LoA). Although ICC values are informative, these are greatly dependent on inter-subject variance in the outcome measure. The

knowledge of the measurement error of a particular measure is essential for both researchers and clinicians when selecting a reliable outcome as both need to establish whether an "improvement" in a patient's walking performance, with the use of assistive technology, is due to measurement error or a 'true' change as a result of the intervention [80]. This is best achieved via the implementation of MDC data, the value beyond which, in this instance, a difference between performance with and without assistive technology can be considered a true change. In our review, of the 18 studies evaluating 'relative' reliability, only four also reported the measurement error of six outcomes (MSWS-12, 10mWT, T25FW, 6MWT, RPE and VO2peak). The MDC was reported to be 22 points, 2.7s and 88 meters for the MSWS-12, T25FW and 6MWT respectively [18]. Paltamaa et al [48] reported an SEM of 0.09 for the 10mWT that indicates an MDC (95%) of 2.4s. Heine et al [50] reported an SEM of 1.1 and 0.131 for RPE and VO_2peak respectively, indicating an MDC (95%) of 3.04 for RPE and 0.36 $L \cdot min^{-1}$ for VO_2peak. However, the strength of the results in these studies rated as 'indeterminate' because the MIC values were not reported. According to Terwee et al [81], the value of the measurement error needs to be considered in relation to MIC (also referred to as the MCID) values in order to determine whether the measurement error of an outcome measure is acceptable for use in research or clinical practice. If the measurement error is exceeding MIC, it is difficult to interpret whether the observed changes are clinically relevant and are not just because of measurement error [52,81]. Another issue to consider is that patient-related factors, such as medications and comorbidities, can influence clinical outcome measurement findings by contributing to measurement error. Many people with MS using medications and have co-morbidities and symptoms such as fatigue, which may change over a period of several weeks or even days [82-83]. These factors are likely to affect outcome measures, both in test-retest reliability studies and clinical trials. One of the items in COSMIN checklist for reliability and measurement error is: 'Were patients stable in the interim period on

the construct to be measured?' For an 'excellent' score for this item authors need to provide evidence that the patients were stable. However, none of the papers, including those with repeated assessment over more than two weeks [18,50] reported this evidence.

The methodological quality of the 41 studies rated according to the COSMIN criteria revealed that both the analysis and reporting of the psychometric properties of outcome measures is often inappropriate. For example, the methodological quality of responsiveness studies was often only rated as 'fair' and 'poor' because the hypotheses were not reported or because there was a lack of information regarding the comparator instruments (often EDSS) and their psychometric properties. Another potential problematic issue with evaluating responsiveness was that in eight of the 14 studies there was no intervention and the (often assumed) hypothesis was that pwMS would deteriorate over the time frame of the study, which ranged from one to two years.

The comparator instrument in seven out of the 15 studies that evaluated validity was the EDSS, which has been widely accepted as a gold standard to measure disability in pwMS. However, its use as a gold standard to validate outcomes of walking performance may be less appropriate. The EDSS [84] is a scale that was developed over 30 years ago and even though studies have reported high inter- and intra-rater reliability and high correlations for face validity [85-86], there are other studies raising issues regarding its reliability and objectivity and whether it can be considered a 'gold standard' [87-88].

It should be noted that the aforementioned methodological issues in the studies included in this review do not imply that the outcome measures are not appropriate but instead that more psychometric studies with higher methodological quality are needed. When planning studies to assess the psychometric properties of outcome measures, researchers should consult standard

guidelines such as COSMIN in relation to the selection of appropriate study design, statistical analysis and reporting of methods and results.

To our knowledge, this is the first review that evaluated the evidence for the psychometric properties of walking performance related measures used to assess the effect of assistive technology in pwMS. We used standardized criteria to evaluate both the methodological quality (COSMIN) and quality of the results [22,25]. To date, only two reviews have tried to highlight which are the most useful tools for walking assessment in pwMS. However, one was a narrative review of available outcome measures and offered little detail about psychometric properties [10]. The other was a topical review including some details of the psychometric properties of measures to assess walking disability, but which did not employ specific criteria to evaluate the evidence for their use [89]. Work has been published on the stroke population that evaluated, also using COSMIN criteria, the psychometric properties of walking performance measures [90]. This review concluded that most of the outcome measures were reliable and valid for use in the stroke population, but it was observed, similar to our findings, that there was a lack of evidence for the minimally important change and responsiveness. Two COSMIN reviews into the functional outcomes in the cerebral palsy population came to similar conclusions [91-92].

This review has several limitations. Firstly, the COSMIN checklist was originally designed for patient-reported outcome measures and not for performance-based measures such as the majority of those included in our review. However, as there is no specific checklist for performance-based measures we opted to use the COSMIN checklist since most of the items scored are also highly relevant to performance-based measures. Additional rules were specified for the ratings of items that were only applicable to patient-reported outcome measures. Another limitation is that only studies published in English, Greek or Dutch were included, which means that eligible studies in

other languages will likely have been excluded. Finally, in the majority of the included studies, the mean EDSS was four or more and five studies involved participants with a mean EDSS of six. The responsiveness and reliability of walking performance measures in pwMS with EDSS > 4 may be different from those who are less affected by MS.

Conclusion

The present systematic review reported on the psychometric properties of outcome measures used to assess the effects of assistive technology to treat foot drop. Forty-one studies were identified which reported information on the psychometric properties of only 10 of the previously identified 20 measures related to walking performance. Strong levels of evidence were found for internal consistency and test-retest reliability of the MSWS-12 and the construct validity of the T25FW. Moderate evidence was found for the test-retest reliability and responsiveness of the 6MWT and for the responsiveness of the 10mWT. None of the outcome measures that were evaluated for measurement error had an acceptable level of evidence for this measurement property. Our findings do not indicate that the existing outcome measures included in this review are poor, but that there is a need for more high quality studies evaluating the psychometric properties of these measures. Future research should (i) investigate the psychometric properties, and in particular measurement error and responsiveness, of a wider range of walking performance related measures and (ii) use standard guidelines such as the COSMIN to increase methodological quality enabling clinicians and researchers to select appropriate outcome measures to assess the effects of assistive technology to treat foot drop.

Declaration of Conflicting Interests

The authors declare no conflict of interest with respect to authorship, and/or publication of this article.

References

[1] W.I. McDonald, T.A. Sears. The effects of experimental demyelination on conduction in the central nervous system. Brain 93 (1970) 583-598.

[2] N.G. LaRocca. Impact of walking impairment in multiple sclerosis: perspectives of patients and care partners. Patient 4 (2011) 189-201.

[3] C. Heesen, J. Böhm, C. Reich, J. Kasper, M. Goebel, S.M. Gold. Patient perception of bodily functions in multiple sclerosis: gait and visual function are the most valuable. Mult. Scler. 14 (2008) 988-991.

[4] A. Kalron, A. Achiron, Z. Dvir. Muscular and gait abnormalities in persons with early onset multiple sclerosis. J. Neurol. Phys. Ther. 35 (2011) 164-169.

[5] C.L. Barrett, G.E. Mann, P.N. Taylor, P. Strike. A randomized trial to investigate the effects of functional electrical stimulation and therapeutic exercise on walking performance for people with multiple sclerosis. Mult. Scler. 15 (2009) 493-504.

[6] B.M. Kluger, L.B. Krupp, R.M. Enoka. Fatigue and fatigability in neurologic illnesses: proposal for a unified taxonomy. Neurology 80 (2013) 409-416.

[7] C. Bulley, T.H. Mercer, J.E. Hooper, P. Cowan, S. Scott, M.L. van der Linden. Experiences of functional electrical stimulation (FES) and ankle foot orthoses (AFOs) for foot-drop in people with multiple sclerosis. Disabil. Rehabil. Assist. Technol. 10 (2015) 458-467.

[8] R.B. Stein, S. Chong, D.G. Everaert, et al. A multicenter trial of a footdrop stimulator controlled by a tilt sensor. Neurorehabil. Neural. Repair 20 (2006) 371-379.

[9] L. Paul, D. Rafferty, S. Young, L. Miller, P. Mattison, A. McFadyen. The effect of functional electrical stimulation on the physiological cost of gait in people with multiple sclerosis. Mult. Scler. 14 (2008) 954-961.

[10] F. Bethoux, S. Bennett. Evaluating walking in patients with multiple sclerosis: which assessment tools are useful in clinical practice? Int. J. MS Care 13 (2011) 4-14.

[11] D.S. Stokic, T.S. Horn, J.M. Ramshur, J.W. Chow. Agreement between temporospatial gait parameters of an electronic walkway and a motion capture system in healthy and chronic stroke populations. Am. J. Phys. Med. Rehabil. 88 (2009) 437-444.

[12] H. Kainz, C.P. Carty, S. Maine, H.P.J. Walsh, D.G. Lloyd, L. Modenese. Effects of hip joint centre mislocation on gait kinematics of children with cerebral palsy calculated using patient-specific direct and inverse kinematic models. Gait Posture 57 (2017) 154-160.

[13] M. Roiz Rde, E.W. Cacho, M.M. Pazinatto, J.G. Reis, A. Jr. Cliquet, E.M. Barasnevicius-Quagliato. Gait analysis comparing Parkinson's disease with healthy elderly subjects. Arq. Neuropsiquiatr. 68 (2010) 81-86.

[14] M. Pistacchi, M. Gioulis, F. Sanson, E. De Giovannini, F. Rossetto, S. Zambito Marsala. Gait analysis and clinical correlations in early Parkinson's disease. Funct. Neurol. 32 (2017) 28-34.

[15] M.D. Goldman, R.A. Marrie, J.A. Cohen. Evaluation of the six-minute walk in multiple sclerosis subjects and healthy controls. Mult. Scler. 14 (2008) 383-390.

[16] J.P. Stellmann, A. Neuhaus, N. Gotze, et al. Ecological validity of walking capacity tests in multiple sclerosis. PLoS One 10 (2015) e0123822.

[17] Y.C. Learmonth, L. Paul, A.K. McFadyen, P. Mattison, L. Miller. Reliability and clinical significance of mobility and balance assessments in multiple sclerosis. Int. J. Rehabil. Res. 35 (2012) 69-74.

[18] Y.C. Learmonth, D.D. Dlugonski, L.A. Pilutti, B.M. Sandroff, R.W. Motl. The reliability, precision and clinically meaningful change of walking assessments in multiple sclerosis. Mult. Scler. 19 (2013) 1784-1791.

[19] C.B. Terwee, E.P. Jansma, I.I. Riphagen, H.C. de Vet. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual. Life Res. 18 (2009) 1115-1123.

[20] R.G. Elbers, M.B. Rietberg, E.E. van Wegen, et al. Self-report fatigue questionnaires in multiple sclerosis, Parkinson's disease and stroke: a systematic review of measurement properties. Qual. Life Res. 21 (2012) 925-944.

[21] L.B. Mokkink, C.B. Terwee, D.L. Patrick, et al. The COSMIN checklist manual. 2012.

[22] C.B. Terwee, L.B. Mokkink, D.L. Knol, R.W. Ostelo, L.M. Bouter, H.C. de Vet. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual. Life Res. 21 (2012) 651-657.

[23] L.B. Mokkink, C.B. Terwee, E. Gibbons, et al. Inter-rater agreement reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Med. Res. Methodol. 10 (2010) 82.

[24] L.B. Mokkink, C.B. Terwee, D.L. Patrick, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual. Life Res. 19 (2010) 539-549.

[25] C.B. Terwee, S.D. Bot, M.R. de Boer, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 60 (2007) 34-42.

[26] Systematic reviews of measurement properties: Assessment of the results of the included studies.http://www.cosmin.nl/Systematic%20reviews%20of%20measurement%20properties.html Accessed October 2, 2016.

[27] M. van Tulder, A. Furlan, C. Bombardier, et al. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. Spine 28 (2003) 1290-1299.

[28] J.J. Kragt, F.A. van der Linden, J.M. Nielsen, B.M. Uitdehaag, C.H. Polman. Clinical impact of 20% worsening on Timed 25-foot Walk and 9-hole Peg Test in multiple sclerosis. Mult. Scler. 12 (2006) 594-598.

[29] E.L. Hoogervost, N.F. Kalkers, B.M. Uitdehaag, C.H. Polman. A study validating changes in multiple sclerosis functional composite. Arch. Neurol. 59 (2002) 113-116.

[30] M.D. Goldman, R.W. Motl, J. Scagnelli, J.H. Pula, J.J. Sosnoff, D. Cadavid. Clinically meaningful performance benchmarks in MS: timed 25-foot walk and the real world. Neurology 81 (2013) 1856-1863.

[31] R.W. Motl, Y.C. Learmonth, L.A. Pilutti, D. Dlugonski, R. Klaren. Validity of minimal clinically important difference values for the multiple sclerosis walking scale-12? Eur. Neurol. 71 (2014) 196-202.

[32] J. Freeman, R. Walters, W. Ingram, A. Slade, J. Hobart, J. Zajicek. Evaluating change in mobility in people with multiple sclerosis: relative responsiveness of four clinical measures. Mult. Scler. 19 (2013) 1632-1639.

[33] J.C. Hobart, A. Riazi, D.L. Lamping, R. Fitzpatrick, A.J. Thompson. Measuring the impact of MS on walking ability: The 12-item MS Walking Scale (MSWS-12). Neurology 60 (2003) 31-36.

[34] R.W. Motl, E. McAuley, S. Mullen. Longitudinal measurement invariance of the Multiple Sclerosis Walking Scale-12. J. Neurol. Sci. 305 (2011) 75-79.

[35] M. Kaufman, D. Moyer, J. Norton. The significant change for the Timed 25-Foot Walk in the multiple sclerosis functional composite. Mult. Scler. 6 (2000) 286-290.

[36] E. Toomey, S. Coote. Between-rater reliability of the 6-minute walk test, berg balance scale, and handheld dynamometry in people with multiple sclerosis. Int. J. MS Care 15 (2013) 1-6.

[37] J.J. Sosnoff, M. Weikert, D. Dlugonski, D.C. Smith, R.W. Motl. Quantifying gait impairment in multiple sclerosis using GAITRite technology. Gait Posture 34 (2011) 145-147.

[38] C. Vaney, H. Blaurock, B. Gattlen, C. Meisels. Assessing mobility in multiple sclerosis using the Rivermead Mobility Index and gait speed. Clin. Rehabil. 10 (1996) 216-226.

[39] C.I. Coleman, D.M. Sobieraj, L.N. Marinucci. Minimally important clinical difference of the Timed 25-Foot Walk Test: results from a randomized controlled trial in patients with multiple sclerosis. Curr. Med. Res. Opin. 28 (2012) 49-56.

[40] J.C. Hobart, A.R. Blight, A. Goodman, F. Lynn, N. Putzki. Timed 25-Foot Walk: Direct evidence that improving 20% or greater is clinically meaningful in MS. Neurology 80 (2013) 1509-1517.

[41] G.R. Cutter, M.L. Baier, R.A. Rudick, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 122 (1999) 871-882.

[42] J.J. Kragt, A.J. Thompson, X. Montalban, et al. Responsiveness and predictive value of EDSS and MSFC in primary progressive MS. Neurology 70 (2008) 1084-1091.

[43] J.A. Cohen, G.R. Cutter, J.S. Fischer, et al. Use of the multiple sclerosis functional composite as an outcome measure in a phase 3 clinical trial. Arch. Neurol. 58 (2001) 961-967.

[44] R.D. Larson, D.J. Larson, T.B. Baumgartner, L.J. White. Repeatability of the timed 25-foot walk test for individuals with multiple sclerosis. Clin. Rehabil. 27 (2013) 719-723.

[45] J.A. Cohen, J.S. Fischer, D.M. Bolibrush, et al. Intrarater and interrater reliability of the MS functional composite outcome measure. Neurology 54 (2000) 802-806.

[46] P. Feys, B. Bibby, A. Romberg, et al. Within-day variability on short and long walking tests in persons with multiple sclerosis. J. Neurol. Sci. 338 (2014) 183-187.

[47] D.K. Fry, L.A. Pfalzer. Reliability of four functional tests and rating of perceived exertion in persons with multiple sclerosis. Physiother. Can. 58 (2006) 212-220.

[48] J. Paltamaa, H. West, T. Sarasoja, J. Wikström, E. Mälkiä. Reliability of physical functioning measures in ambulatory subjects with MS. Physiother. Res. Int. 10 (2005) 93-109.

[49] S.R. Schwid, A.D. Goodman, M.P. McDermott, C.F. Bever, S.D. Cook. Quantitative functional measures in MS: What is a reliable change? Neurology 58 (2002) 1294-1296.

[50] M. Heine, L.E. van den Akker, O. Verschuren, A. Visser-Meily, G. Kwakkel, TREFAMS-ACE Study Group. Reliability and responsiveness of cardiopulmonary exercise testing in fatigued persons with multiple sclerosis and low to mild disability. PLoS One 10 (2015) e0122260.

[51] B.T. Cleland, B.A. Ingraham, M.C. Pitluck, D. Woo, A.V. Ng. Reliability and validity of ratings of perceived exertion in persons with multiple sclerosis. Arch. Phys. Med. Rehabil. 97 (2016) 974-982.

[52] H.C. de Vet, C.B. Terwee, R.W. Ostelo, H. Beckerman, D.L. Knol, L.M. Bouter. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual. Life Outcomes 4 (2006) 54.

[53] L.B. Mokkink, F. Galindo-Garre, B.M. Uitdehaag. Evaluation of the Multiple Sclerosis Walking Scale-12 (MSWS-12) in a Dutch sample: Application of item response theory. Mult. Scler. 22 (2016) 1867-1873.

[54] M.M. Engelhard, K.M. Schmidt, C.E. Engel, J.N. Brenton, S.D. Patek, M.D. Goldman. The e-MSWS-12: improving the multiple sclerosis walking scale using item response theory. Qual. Life Res. 25 (2016) 3221-3230.

[55] R.W. Motl, E.M. Snook. Confirmation and extension of the validity of the Multiple Sclerosis Walking Scale-12 (MSWS-12). J. Neurol. Sci. 268 (2008) 69-73.

[56] R.W. Motl, D. Dlugonski, Y. Suh, et al. Multiple Sclerosis Walking Scale-12 and oxygen cost of walking. Gait Posture 31 (2010) 506-510.

[57] J.C. Hobart, N. Kalkers, F. Barkhof, B. Uitdehaag, C. Polman, A.J. Thompson. Outcome measures for multiple sclerosis clinical trials: Relative measurement precision of the Expanded Disability Status Scale and Multiple Sclerosis Functional Composite. Mult. Scler. 10 (2004) 41-46.

[58] R.R.G. Apache, D. Jackson, D.H. Mattson. Use of reaction and movement time as a measure of lower extremity functioning in multiple sclerosis. Clin. Kinesiol. 56 (2002) 35-41.

[59] L.A. Pilutti, D. Dlugonski, B.M. Sandroff, et al. Further validation of multiple sclerosis walking scale-12 scores based on spatiotemporal gait parameters. Arch. Phys. Med. Rehabil. 94 (2013) 575-578.

[60] C. McGuigan, M. Hutchinson. Confirming the validity and responsiveness of the Multiple Sclerosis Walking Scale-12 (MSWS-12). Neurology 62 (2004) 2103-2105.

[61] D.M. Miller, R.A. Rudick, G. Cutter, M. Baier, J.S. Fischer. Clinical significance of the multiple sclerosis functional composite: Relationship to patient-reported quality of life. Arch. Neurol. 57 (2000) 1319-1324.

[62] N.F. Kalkers, L. Bergers, V. De Groot, et al. Concurrent validity of the MS Functional Composite using MRI as a biological disease marker. Neurology 56 (2001) 215-219.

[63] P. Filipović Grčić, M. Matijaca, I. Lušić, V. Čapkun. Responsiveness of walking-based outcome measuresafter multiple sclerosis relapses following steroid pulses. Med. Sci. Monit. 17 (2011) 704-710.

[64] J.C. Kempen, V. de Groot, D.L. Knol, C.H. Polman, G.J. Lankhorst, H. Beckerman. Community walking can be assessed using a 10-metre timed walked test. Mult. Scler. 17 (2011) 980-990.

[65] J. Paltamaa, T. Sarasoja, E. Leskinen, J. Wikström, E. Mälkiä. Measuring deterioration in international classification of functioning domains of people with multiple sclerosis who are ambulatory. Phys. Ther. 88 (2008) 176-190.

[66] I. Baert, J. Freeman, T. Smedal, et al. Responsiveness and clinically meaningful improvement, according to disability level, of five walking measures after rehabilitation in multiple sclerosis: A European multicenter study. Neurorehabil. Neural. Repair 28 (2014) 621-631.

[67] L.M. van Winsen, J.J. Kragt, E.L. Hoogervorst, C.H. Polman, B.M. Uitdehaag. Outcome measurement in multiple sclerosis: Detection of clinically relevant improvement. Mult. Scler. 16 (2010) 604-610.

[68] R.R.G. Apache, D. Jackson, D.H. Mattson. Responsiveness of choice reaction and movement time in multiple sclerosis. Clin. Kinesiol. 59 (2005) 25-30.

[69] H.B. Jensen, S. Mamoei, M. Ravnborg, U. Dalgas, E. Stenager. Distribution-based estimates of minimum clinically important difference in cognition, arm function and lower body function after slow release-fampridine treatment of patients with multiple sclerosis. Mult. Scler. Relat. Disord. 7 (2016) 58-60.

[70] G. Yavuzer, O. Oken, A. Elhan, H.J. Stam. Repeatability of lower limb three-dimensional kinematics in patients with stroke. Gait Posture 27 (2008) 31-35.

[71] K.J. Noonan, S. Halliday, R. Browne, S. O'Brien, K. Kayes, J. Feinberg. Interobserver variability of gait analysis in patients with cerebral palsy. J. Pediatr. Orthop. 23 (2003) 279-287.

[72] A. Nieuwenhuys, E. Papageorgiou, G. Molenaers, D. Monari, T. de Laet, K. Desloovere. Inter- and intrarater clinician agreement on joint motion patterns during gait in children with cerebral palsy. Dev. Med. Child Neurol. 59 (2017) 750-755.

[73] H. Kainz, D. Graham, J. Edwards, et al. Reliability of four models for clinical gait analysis. Gait Posture 54 (2017) 325-331.

[74] D. Laroche, A. Duval, C. Morisset, et al. Test-retest reliability of 3D kinematic gait variables in hip osteoarthritis patients. Osteoarthritis Cartilage 19 (2011) 194-199.

[75] A.V. Bates, A.H. McGregor, C.M. Alexander. Reliability and minimal detectable change of gait kinematics in people who are hypermobile. Gait Posture 44 (2016) 37-42.

[76] A.S. Delussu, G. Morone, M. Iosa, M. Bragoni, S. Paolucci, M. Traballesi. Concurrent validity of Physiological Cost Index in walking over ground and during robotic training in subacute stroke patients. Biomed. Res. Int. (2014).

[77] K. Raja, B. Joseph, S. Benjamin, V. Minocha, B. Rana. Physiological cost index in cerebral palsy: its role in evaluating the efficiency of ambulation. J. Pediatr. Orthop. 27 (2007) 130-136.

[78] L. Miller, D. Rafferty, P. Mattison. The impact of walking speed on the effects of functional electrical stimulation for foot drop in people with multiple sclerosis. Disabil. Rehabil. Assist. Technol. 11 (2016) 478-483.

[79] H.C. De Vet, C.B. Terwee, D.L. Knol, L.M. Bouter. When to use agreement versus reliability measures. J. Clin. Epidimiol. 59 (2006) 1033-1039.

[80] A. Bruton, J.H. Conway, S.T. Holgate. Reliability: What is it, and how is it measured? Physiotherapy 86 (2000) 94-99.

[81] C.B. Terwee, L.D. Roorda, D.L. Knol, M.R. de Boer, H.C. de Vet. Linking measurement error to minimal important change of patient-reported outcomes. J. Clin. Epidemiol. 62 (2009) 1062-1067.

[82] D.J. Powell, C. Liossi, W. Schlotz, R. Moss-Morris. Tracking daily fatigue fluctuations in multiple sclerosis: ecological momentary assessment provides unique insights. J. Behav. Med. 40 (2017) 772-783.

[83] S.L. Kasser, A. Goldstein, P.K. Wood, J. Sibold. Symptom variability, affect and physical activity in ambulatory persons with multiple sclerosis: Understanding patterns and time-bound relationships. Disabil. Health J. 10 (2017) 207-213.

[84] J.F. Kurtzke. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33 (1983) 1444-1452.

[85] D.E. Goodkin, D. Cookfair, K. Wende, et al. Inter- and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke Expanded Disability Status Scale (EDSS). Multiple Sclerosis Collaborative Research Group. Neurology 42 (1992) 859-863.

[86] B. Sharrack, R.A. Hughes, S. Soudain, G. Dunn. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain 122 (1999) 141-159.

[87] J.C. Hobart, J. Freeman, A.J. Thompson. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 123 (2000) 1027-1040.

[88] J.A. Cohen, S.C. Reingold, C.H. Polman, J.S. Wolinsky. Disability outcomes measures in multiple sclerosis clinical trials: current status and future prospects. Lancet Neurol. 11 (2012) 467-476.

[89] B.C. Kieseier, C. Pozzilli. Assessing walking disability in multiple sclerosis. Mult. Scler. 18 (2012) 914-924.

[90] M. van Bloemendaal, A.T. van de Water, I.G. van de Port. Walking tests for stroke survivors: a systematic review of their measurement properties. Disabil. Rehabil. 34 (2012) 2207-2221.

[91] C. Ammann-Reiffer, C.H. Bastiaenen, R.A. de Bie, H.J. van Hedel. Measurement properties of gait-related outcomes in youth with neuromuscular diagnoses: a systematic review. Phys. Ther. 94 (2014) 1067-1082.

[92] A. Zanudin, M.L. van der Linden, K. Jagadamma, T.H. Mercer. Psychometric properties of gait quality and walking performance measures in adolescents and young adults with Cerebral Palsy: A systematic review. Gait Posture 58 (2017) 30-40.

Table 1 Quality criteria for measurement properties [25-26].

Measurement property	Rating*	Criteria
Reliability		
Internal consistency	+	At least limited evidence for unidimensionality or positive structural validity AND Cronbach's alpha(s) ≥0.70 and ≤0.95
	?	Not all information for '+' reported OR conflicting evidence for unidimensionality or structural validity OR evidence for lack of unidimensionality or negative structural validity
	-	Criteria for '+' not met
Reliability	+	ICC or weighted Kappa ≥0.70
	?	ICC or weighted Kappa not reported
	-	Criteria for '+' not met
Measurement error	+	SDC or LoA < MIC
	?	MIC not defined
	-	Criteria for '+' not met
Validity		
Construct validity	+	At least 75% of the results are in accordance with the hypotheses
(Hypothesis testing)	?	No correlations with instrument(s) measuring related construct(s) AND no differences between relevant groups reported
	-	Criteria for '+' not met
Criterion validity	+	Convincing arguments that gold standard is "gold" AND correlation with gold standard ≥0.70
	?	Not all information for '+' reported
	-	Criteria for '+' not met
Responsiveness		
Responsiveness	+	At least 75% of the results are in accordance with the hypotheses
	?	No correlations with changes in instrument(s) measuring related construct(s) AND no differences between changes in relevant groups reported
	-	Criteria for '+' not met

* + = positive rating; ? = indeterminate rating; - = negative rating

Table 2 Summary of the study characteristics, rating of the methodological quality using the COSMIN guidelines and rating of the quality of the results using the Terwee criteria [24-26].

Author/Year	Patient characteristics	COSMIN Measurement Property	Results	Rating	
				Methodological quality	Quality of the results
Choice reaction time & movement time (RT/MT)					
Apache et al (2002)[58]	n = 178 ,RR, SP,PP EDSS 0-6.5	Hypothesis testing	r_s = 0.84 with EDSS	**Fair** [a]	+
Apache et al (2005)[68]	n = 40, RR, SP EDSS median 4.5 3 sessions in 1-year	Responsiveness (no intervention)	RT/MT mean change =16.6% (.1)	**Poor** [b]	**?**
MSFC					
Cohen et al (2000)[45]	n = 10, SP EDSS mean 5.2 6 sessions (2 per day over 2 weeks)	Reliability	Intra-rater: ICC = 0.97 (session 4-5) Inter-rater: ICC = 0.96 (session 7-8)	**Poor** [c]	+
Cohen et al (2001)[43]	n = 436, SP EDSS mean 5.2 3 pre-baseline sessions over 28 days	Reliability Hypothesis testing	Intra-rater: ICC (over 4 sessions) = 0.87 r_s = -0.56 with EDSS	**Good** [d, e] **Fair** [o, q]	+ +
Cutter et al (1999)[41]	n = 378, RR, SP EDSS 0-6.5 3 annual sessions	Hypothesis testing Responsiveness (no intervention)	r_s = -0.22 with EDSS Average composite change Z-score: Baseline = -0.07 1-year = -0.07 2-year = -0.16	**Fair** [a, o ,q] **Fair** [q]	**?** **?**
Hobart et al (2004)[57]	n = 133, RR, SP,PP EDSS mean 3.1	Hypothesis testing	r = -0.64 with EDSS	**Good** [f, w]	+
Kalkers et al (2001)[62]	n = 131, RR, SP,PP EDSS mean 3.1	Hypothesis testing	r_s = -0.25 with T2 lesion load r_s = -0.24 with T1 lesion load	**Poor** [h]	**?**
Kragt et al (2008)[42]	n = 161, PP EDSS mean 5.0	Responsiveness (no intervention)	ES: EDSS = 0.23 MSFC = 0.16	**Poor** [b]	**?**
Miller et al (2000)[61]	n = 300 EDSS 0-8.5	Hypothesis testing	r_s = -0.80 with HRQoL	**Poor** [h]	**?**
MSWS-12					

Baert et al (2013)[66]	n = 284, RR, SP,PP EDDS mean 4.8 2 sessions (pre & post)	Responsiveness (physical rehabilitation)	AUC with Global Rating Scale: Whole group = 0.73 EDSS≤4 = 0.64 EDSS 4.5-6.5 = 0.77	**Fair** [a]	+
Filipovic et al (2011)[63]	n = 49, RR EDSS mean 3.0 2 sessions (pre & post)	Responsiveness (IVMP for 1month)	SRM = 1.05 ES = 1.02 RE (%) = 82.4	**Fair** [a]	**?**
Freeman et al (2013)[32]	n = 70, RR, SP, PP 3 annual sessions	Responsiveness (no intervention)	ES = -0.07 SEM = 5.66 r <0.35 with walking speed & RMI	**Fair** [j]	**?**
Hobart et al (2003)[33]	Community sample: n = 602 2 sessions (10 days apart)	Internal consistency	Community sample: Cronbach's α = 0.97 PPMS sample: Cronbach's α = 0.97 Steroids sample: Cronbach's α = 0.94	**Excellent**	+
		Reliability	Community sample: Test-retest ICC = 0.94	**Good** [d, e, k]	+
	Hospital-based sample: PP MS = 78 Steroids = 54 2 sessions (6 weeks apart)	Hypothesis testing	Steroids sample: r_s = 0.65 with EDSS	**Fair** [i, q]	+
		Responsiveness (steroid treatment)	With EDSS: ES = 0.45 SRM = 0.45 RE = 0.31	**Fair** [m, q, t]	**?**
Learmonth et al (2013)[18]	n = 82, RR, SP, PP EDSS mean 3.5 2 sessions (7 days apart)	Reliability	Test-retest: ICC(2,1) = 0.93	**Good** [k, t]	+
		Measurement error	SEM = 8; CV (5) = 27 MDC_{95} = 22; %MDC_{95} = 53%	**Fair** [i, m, n]	**?**
McGuigan et al (2004)[60]	Community sample = 149	Hypothesis testing	r_s = 0.84 with EDSS	**Fair** [a, o, q]	**?**
	Outpatient sample = 53 RR, SP, PP EDSS mean 4.0 2 sessions	Responsiveness (no intervention)	Z-score = -2.87	**Poor** [b]	**?**
Motl et al (2008)[55]	n = 133, RR, SP, PP EDSS mean 4.9 1 session	Internal consistency	Cronbach's α =.97	**Good** [p, x]	+
		Hypothesis testing	r_s = .77 with MSIS-29 (physical) r_s = .36 with MSIS-29 (psychological) r_s = .80 with EDSS	**Fair** [q]	+

Motl et al (2010)[56]	n = 24, RR PDDS median 1.0 1 session	Internal consistency	Cronbach's α = .95	**Poor** [c]	+
		Hypothesis testing	With O_2 cost of walking at: CWS, r = 0.64 FWS, r = 0.61 SWS, r = 0.64 With O_2 consumption: CWS, r = 0.24 FWS, r = 0.14 SWS, r = 0.44	**Poor** [c]	+
Motl et al (2011)[34]	n = 269, RR 3 sessions over a year	Internal consistency	Cronbach's α: Baseline = .96 6-month = .97 12-month = .97	**Excellent**	+
		Reliability	Test-retest ICC(2): Across 6-months = .86 Across 12-months = .87	**Fair** [i]	+
Pilutti et al (2013)[59]	n = 268, RR, SP,PP PDDS median 3.0 1 session	Hypothesis testing	r_s = .72 with T25FW r_s = -.75 with 6MWT	**Fair** [q]	+
Mokkink et al (2016)[53]	n = 625, RR, SP, PP,PR, CIS EDSS median 3.5	Internal consistency	RMSEA = 0.078 CFI = 1.000 TLI = 0.999 SRMR = 0.019 Guttman's lambda2 = 0.98	**Excellent**	+
Engelhard et al (2016)[54]	n = 293, RR, SP, PP,PR	Internal consistency	1D Rasch: BIC = 6112.5; AIC = 5933.7 3D GRM: BIC = 5972.7; AIC = 5677.3	**Excellent**	+
FAP/ Spatiotemporal parameters					
Sosnoff et al (2011)[37]	n = 13, RR, SP EDSS median 6.0 1 session	Hypothesis testing	FAP: r_s = -0.82 with T25FW r_s = -0.49 with MSWS-12 r_s = -0.81 with EDSS	**Poor** [c]	+
Pilutti et al (2013)[59]	n = 268, RR, SP,PP PDDS median 3.0	Hypothesis testing	Speed with T25FW: r = -.68 Cadence with T25FW: r = -.50	**Fair** [q]	+

	1 session		Speed with 6MWT: r = .67 Cadence with 6MWT: r = .52		
10mWT					
Feys et al (2014)[46]	n = 102, RR, SP,PP EDSS mean 4.6 3 sessions within a day	Reliability	Test-retest: Within-day variability (%) at usual speed: -Community walkers (CW) = 22.6 -Limited CW = 26.6 -Most limited CW = 43.3 Within-day variability (%) at fastest speed: -CW = 12.1 -Limited CW = 23.4 -Most limited CW = 38.4	**Poor** [b]	**?**
Freeman et al (2013)[32]	n = 70, RR, SP, PP 3 annual sessions	Responsiveness (no intervention)	ES = 0.001 r <0.35 with MSWS-12 & RMI	**Fair** [j]	**?**
Kempen et al (2011)[64]	n = 156 , RR EDSS mean 2.5 6 sessions in 6 years	Responsiveness (no intervention)	AUC = 0.79 with MFWC6 AUC = 0.86 with MFWC5 AUC = 0.74 with MFWC4 AUC = 0.82 with MFWC3	**Fair** [j]	**+**
Paltamaa et al (2005)[48]	Test-retest n = 19 Inter-rater n = 9	Reliability	Test-retest: ICC = 0.91 Inter-rater: ICC = 0.93	**Poor** [c]	**+**
	RR, SP, PP EDSS 0-6.5 2 sessions (1 week apart)	Measurement error	Test-retest: SEM = 0.09m/s Inter-rater: SEM = 0.10m/s	**Poor** [c]	**?**
Paltamaa et al (2008)[65]	Baseline n = 120 Follow-up n = 109 RR, PP EDSS median 2.0 3 sessions in 2years	Responsiveness (no intervention)	AUC = 0.76 with EDSS $MIC_{deterioration}$ = -0.19	**Fair** [a]	**+**
Stellman et al (2015)[16]	n = 28, RR, SP, PP EDSS mean 3.2 1 session	Hypothesis testing	r = 0.61 with accelerometry	**Poor** [c]	**+**
Vaney et al (1996)[38]	Reliability n = 25 Responsiveness n = 115 EDSS mean 6.6 5 sessions within-day	Reliability Responsiveness (physical & occupational therapy)	Test-retest: r_s = -0.8 with RMI Not adequate statistical information for responsiveness	**Poor** [c, u] **Poor** [y]	**?** **?**

Timed 25-Foot Walk					
Baert et al (2013)[66]	n = 284, RR, SP,PP EDDS mean 4.8 2 sessions (pre & post)	Responsiveness (physical rehabilitation)	AUC with Global Rating Scale: Whole group = 0.50 EDSS≤4 = 0.64 EDSS 4.5-6.5 = 0.45	**Fair** [a]	+
Coleman et al (2012)[39]	n = 296, RR, RP, SP, PP EDSS mean 5.8 4 sessions	Responsiveness (dalfampridine treatment)	r_s = -0.39 with CGI MICD = 0.35 m/s Relative improvement = 17.2%	**Fair** [a]	+
Filipovic et al (2011)[63]	n = 49, RR EDSS mean 3.0 2 sessions (pre & post)	Responsiveness (IVMP for 1 month)	SRM = 0.55 ES = 0.27 RE (%) = 68.3	**Fair** [a, t]	?
Hobart et al (2013)[40]	n = 533, RR, SP,PP EDSS mean 6.0 9 sessions	Reliability Hypothesis testing	Variability ranged from 10.03 – 11.44 r = -0.20 to -0.43 with MSWS-12	**Poor** [b] **Excellent**	? +
Kaufman et al (2000)[35]	n = 133, SP 3 sessions (6 month period)	Reliability Responsiveness (no intervention)	Not adequate statistical information for reliability Not adequate statistical information for responsiveness	**Poor** [b] **Poor** [b, h]	? ?
Larson et al (2013)[44]	n = 36, RR EDSS mean 3.5 2 sessions 1 week apart	Reliability	Test-retest ICC = 0.92	**Fair** [c]	+
Learmonth et al (2012)[17]	n = 24 EDSS mean 6.02 2 sessions 1 week apart	Reliability Measurement error	Test-retest ICC(2,3) = 0.94 SEM = 4.56s MDC_{95} = 12.6s	**Poor** [c] **Poor** [c]	+ ?
Learmonth et al (2013)[18]	n = 82, RR, SP, PP EDSS mean 3.5 2 sessions (7 days apart)	Reliability Measurement error	Test-retest ICC(2,1) = 0.991 SEM = 1s MDC_{95} = 2.7s % MDC_{95} = 36	**Good** [t] **Fair** [i, m, n]	+ ?
Schwid et al (2002)[49]	n = 63 EDSS 0-6.5 5 sessions	Reliability	Test-retest reliability: 95% CI: ± 16% of patients baseline score	**Poor** [b]	?
van Winsen et al (2010)[67]	n = 112, CIS, RR, SP, PP	Responsiveness	Sensitivity (%) = 25 Specificity (%) = 90	**Fair** [a]	?

	EDSS mean 4.5 2 sessions (pre & post)	(IVMP for 6 weeks)	LR+ = 2.50 LR- = 0.83		
Jensen et al (2016)[69]	n = 105 EDSS mean 5.6 2 sessions	Responsiveness (SR-Fampridine treatment)	MCID = 1.3s %MCID = 14.2	**Poor** [h, y]	**?**
2-Minute Walk Test					
Baert et al (2013)[66]	n = 284, RR, SP, PP EDDS mean 4.8 2 sessions (pre & post)	Responsiveness (physical rehabilitation)	AUC with Global Rating Scale: Whole group = 0.64 EDSS≤4 = 0.74 EDSS 4.5-6.5 = 0.60	**Fair** [a]	**+**
Feys et al (2014)[46]	n = 102, RR, SP,PP EDSS mean 4.6 3 sessions within a day	Reliability	Within-day variability (%): CW = 12.0 Limited CW = 13.8 Most limited CW = 26.3	**Poor** [b]	**?**
Filipovic et al (2011)[63]	n = 49, RR EDSS mean 3.0 2 sessions (pre & post)	Responsiveness (IVMP for 1 month)	SRM = 0.89 ES = 0.54 RE (%) = 95.1	**Fair** [a]	**?**
Stellman et al (2015)[16]	n = 28, RR, SP, PP EDSS mean 3.2 1 session	Hypothesis testing	r = 0.79 with accelerometry	**Poor** [c]	**+**
6-Minute Walk Test					
Baert et al (2013)[66]	n = 284, RR, SP,PP EDDS mean 4.8 2 sessions (pre & post)	Responsiveness (physical rehabilitation)	AUC with Global Rating Scale: Whole group = 0.68 EDSS≤4 = 0.77 EDSS 4.5-6.5 = 0.65	**Fair** [a]	**+**
Feys et al (2014)[46]	n = 102, RR, SP,PP EDSS mean 4.6 3 sessions within a day	Reliability	Within-day variability (%): CW = 10.1 Limited CW = 15.7 Most limited CW = 28.7	**Poor** [b]	**?**
Freeman et al (2013)[32]	n = 70, RR, SP, PP 3 annual sessions	Responsiveness (no intervention)	ES = 0.03 'general mobility': r = 0.499	**Fair** [j]	**?**
Fry et al (2006)[47]	n = 12, RR, SP, PP EDSS mean 3.6 2 sessions (1 week apart)	Reliability	Test-retest ICC = 0.96	**Poor** [c]	**+**

Goldman et al (2008)[15]	n = 40, RR, SP, PP EDSS 0-6.5 3 sessions (in 4 hours)	Reliability	Test-retest: ICC = 0.94 Inter-rater: ICC = 0.91	**Fair** [t]	+
Learmonth et al (2012)[17]	n = 24 EDSS mean 6.02 2 sessions 1 week apart	Reliability Measurement error	Test-retest: ICC (2,1) = 0.96 SEM = 27.48m MDC_{95} = 76.2m	**Poor** [c] **Poor** [c]	+ **?**
Learmonth et al (2013)[18]	n = 82, RR, SP, PP EDSS mean 3.5 2 sessions (7 days apart)	Reliability Measurement error	Test-retest: ICC(2,1) = 0.96 SEM = 32m MDC_{95} = 88m % MDC_{95} = 20	**Good** [t] **Fair** [i, m, n]	+ **?**
Paltamaa et al (2005)[48]	Test-retest n = 19 Inter-rater n = 9 RR, SP, PP EDSS 0-6.5 2 sessions (1 week apart)	Reliability Measurement error	Test-retest: ICC = 0.96 Inter-rater: ICC = 0.93 Test-retest: SEM = 30.65 m Inter-rater: SEM = 35.85 m	**Poor** [c] **Poor** [c]	+ **?**
Paltamaa et al (2008)[65]	Baseline n = 120 Follow-up n = 109 RR, PP EDSS median 2.0 3 sessions in 2 years	Responsiveness (no intervention)	AUC = 0.76 with EDSS $MIC_{deterioration}$ = -55.06	**Fair** [a]	+
Stellman et al (2015)[16]	n = 28, RR, SP, PP EDSS mean 3.2 1 session	Hypothesis testing	r = 0.68 with accelerometry	**Poor** [c]	+
Toomey et al (2013)[36]	n = 8 1 session(4assessors)	Reliability	Inter-rater: ICC = 0.984	**Poor** [c]	+
RPE					
Heine et al (2015)[50]	n = 31 RR, SP, PP EDSS mean 2.5 2 sessions (1-3 weeks apart)	Reliability Measurement error	Test-retest: ICC = 0.706 SEM = 1.1 $SDC_{individual}$ = 2.9 SDC_{group} = 0.52 LoA = -2.9-2.9	**Poor** [z] **Poor** [z]	+ **?**
Cleland et al (2016)[51]	n = 16 RR, SP, PP EDSS median 1.75	Reliability Hypothesis testing	Test-retest: ICC = 0.870 r = .691 with VO_2 (L/min) r = .507 with VO_2 (mL/kg/min)	**Poor** [c] **Poor** [c, h]	+ **?**

	2 sessions (6-10 days apart)				
VO_2 peak					
Heine et al (2015)[50]	n = 31 RR, SP, PP EDSS mean 2.5	Reliability	Test-retest: ICC = 0.933 for VO_2 peak ($mL \cdot kg^{-1} \cdot min^{-1}$)	**Poor** [z]	+
	2 sessions (1-3 weeks apart)	Measurement error	VO_2 peak ($mL \cdot kg^{-1} \cdot min^{-1}$): SEM = 1.7 $SDC_{individual}$ = 4.6 SDC_{group} = 0.82 LoA = -5.0-4.3	**Poor** [z]	**?**

COSMIN item rating: a: hypothesis vague or not formulated, possible to deduce; b: not appropriate statistical methods; c: small sample size; d: no description of ICC model used; e: assume that patients were stable in the interim period; f: expected magnitude of the correlations not stated; g: assumable that statistical methods were appropriate; h: no information about the psychometric properties of the comparator instruments; i: not clear how missing items were handled; j: unclear or not described what occurred in the interim period; k: assumable that measurements were independent; l: AUC or correlations not calculated; m: unclear if patients were stable; n: doubtful whether time interval was appropriate; o: poor description of the comparator instrument; p: no description of the % of missing data; q: some information on measurement properties or a reference; r: internal consistency not calculated for each subscale separately; s: no ICC, Spearman or Pearson's correlations calculated; t: due to sample size; u: only correlations, not ICC calculated; v: minimal number of hypothesis formulated a priori; w: expected direction of the correlations or differences not stated; x: not described but can be deduced how missing items were handled; y: unclear what was expected; z: test conditions were not similar

Abbreviations: AIC: Akaike information criterion; AUC: Area under the curve; BIC: Bayesian information criterion; CFI: comparative fit index; CGI: Clinician Global Impression; CIS: Clinically Isolated Syndrome; CWS: comfortable walking speed; ES: effect size; FWS: faster walking speed; GPCM: generalized partial credit model; GRM: graded response model; HRQoL: Health-related Quality of Life; ICC: Intraclass Correlation Coefficient; IVMP: intravenous methylprednisolone therapy; LoA: limits of agreement; LR: Likelihood ratio; MDC: minimum detectable change; MIC: minimal important change; MICD: Minimally important clinical difference; MFWC: Modified Functional Walking Categories; PDDS: Patient Determined Disease Steps; PP: Primary Progressive; r: Pearson's correlations; r_s: Spearman coefficient; RE: relative efficiency; RMI: Rivermead Mobility Index; RMSEA: root mean square error of approximation; RR: Relapsing Remitting; SDC: smallest detectable change; SEM: standard error of mean; SP: Secondary Progressive; SRM: standardized response mean; SRMR: root mean square residual; SWS: slower walking speed; TLI: Tucker-Lewis index.

Table 3 Level of evidence for each outcome measure identified in the principal search.

Outcome measure	Internal consistency	Reliability			Measurement error	Hypothesis testing	Responsiveness
		Intra-rater	Inter-rater	Test-retest			
RT/MT	n/a	n/a	n/a	n/a	n/a	+ Limited	? Unknown
MSWS-12	+++ Strong	n/a	n/a	+++ Strong	? Unknown	++ Moderate	+ Limited
MSFC	n/a	++ Moderate	? Unknown	n/a	n/a	++ Moderate	? Unknown
Spatiotemporal parameters	n/a	n/a	n/a	n/a	n/a	+ Limited	n/a
10mWT	n/a	n/a	? Unknown	? Unknown	? Unknown	? Unknown	++ Moderate
T25FW	n/a	n/a	n/a	++ Moderate	? Unknown	+++ Strong	++ Moderate
2MWT	n/a	n/a	n/a	? Unknown	n/a	? Unknown	+ Limited
6MWT	n/a	n/a	+ Limited	++ Moderate	? Unknown	? Unknown	++ Moderate
VO_2 peak	n/a	n/a	n/a	? Unknown	? Unknown	n/a	n/a
RPE	n/a	n/a	n/a	? Unknown	? Unknown	? Unknown	n/a

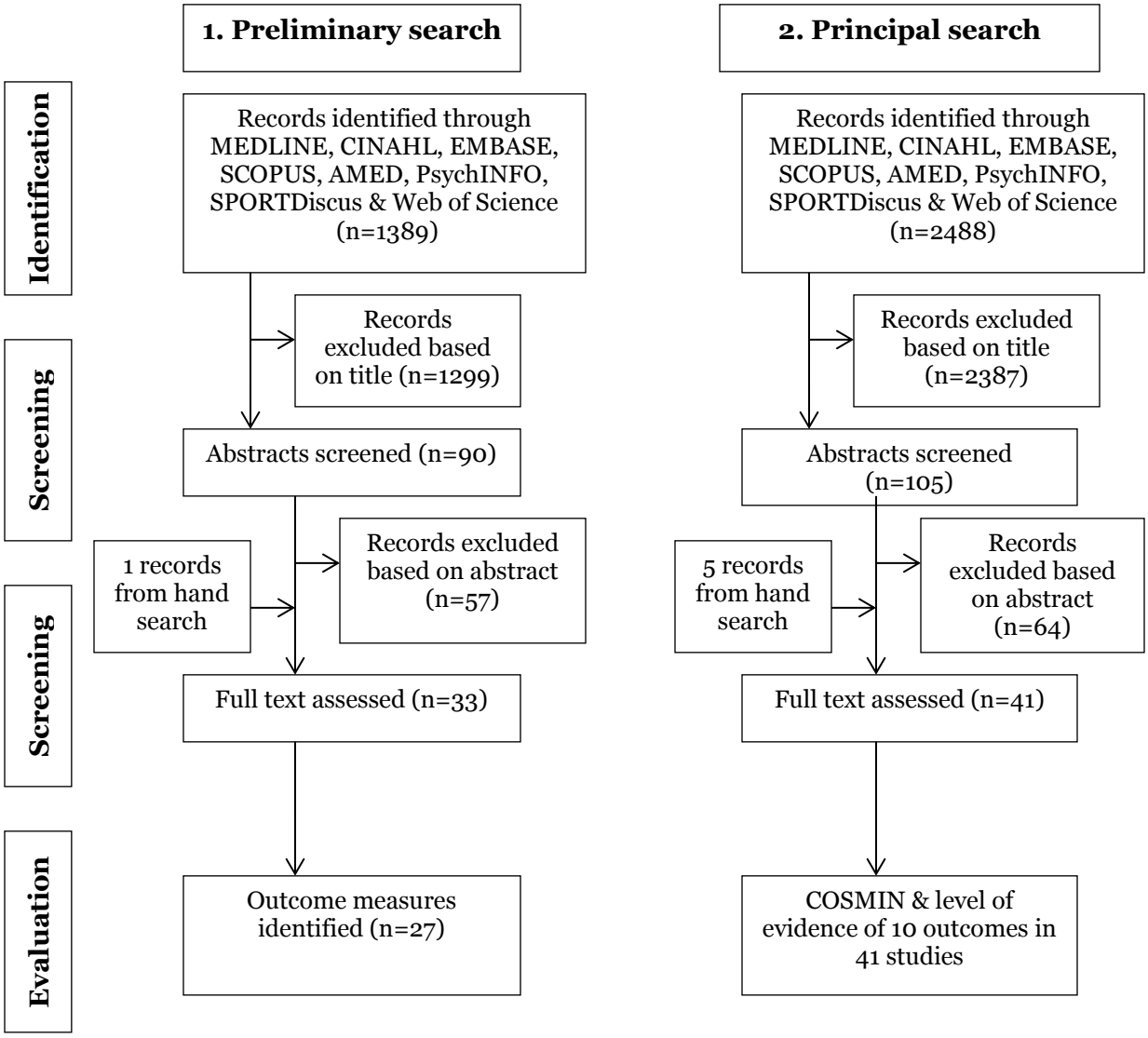

Figure 1 1. Preliminary search: identification of the outcomes measures that have been used to assess the effect of assistive technology for foot drop; 2. Principal search: studies evaluating the psychometric properties of outcome measures of walking performance, effort of walking and lower limb function.

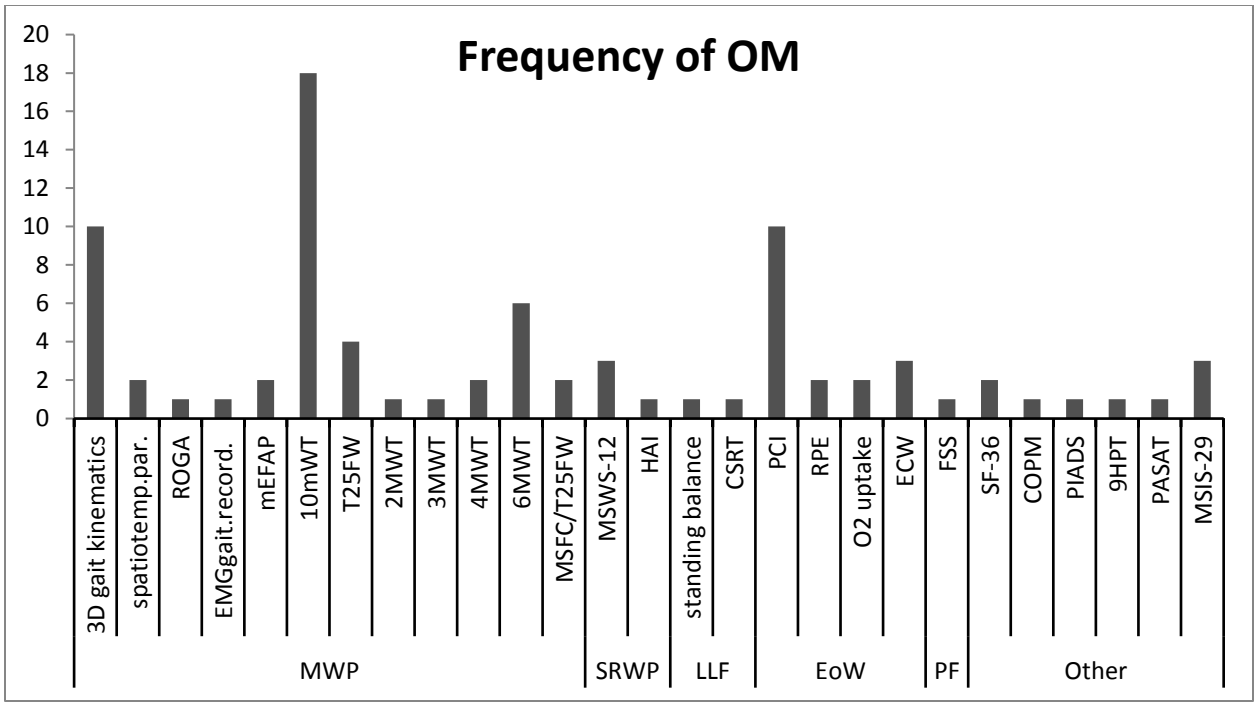

Figure 2 Outcome measures identified in the preliminary search and the reported frequency of use.
Abbreviations: MWP: measured walking performance; SRWP: self-reported walking performance; LLF: lower limb function; EoW: effort of walking; PF: perceived fatigue; spatiotemp. par.: spatiotemporal parameters; ROGA: Rivermead Observational Gait Analysis; EMG gait record: electromyography gait recording; mEFAP: modified Emory Functional Ambulation Profile; 10mWT: 10 meter Walk Test; T25FW: Timed 25 Foot Walk; 2MWT: 2 Minute Walk Test; 3MWT: 3 Minute Walk Test; 4MWT: 4 Minute Walk Test; 6MWT: 6 Minute Walk Test; MSFC: Multiple Sclerosis Functional Composite; MSWS-12: Multiple Sclerosis Walking Scale-12; HAI: Hauser Ambulation Index; CSRT: choice stepping reaction time; PCI: Physiological Cost Index; RPE: Rate of Perceived Exertion; ECW: energy cost of walking; FSS: Fatigue Severity Scale; SF-36: 36-Item Short Form Health Survey; COPM: Canadian Occupational Performance Measure; PIADS: Psychosocial Impact of Assistive Devices Scale; 9HPT: 9-Hole Peg Test; PASAT: Paced Serial Addition Test; MSIS-29: Multiple Sclerosis Impact Scale.