INSTITUTO POLITÉCNICO DE LISBOA

ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA

Instrumentos para a avaliação da marcha em indivíduos com diagnóstico de Esclerose Múltipla: revisão sistemática

MARIA MARGARIDA REALINHO DE JESUS

ORIENTADORA PROFESSORA LUÍSA PEDRO, PhD – ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA

Mestrado em Fisioterapia, ramo de especialização em Fisioterapia Neurológica

Lisboa, Julho de 2021

INSTITUTO POLITÉCNICO DE LISBOA

ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA

Instrumentos para a avaliação da marcha em indivíduos com diagnóstico de Esclerose Múltipla: revisão sistemática

MARIA MARGARIDA REALINHO DE JESUS

PROF. LUÍSA PEDRO, PhD – ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA

Mestrado em Fisioterapia, ramo de especialização em Fisioterapia Neurológica

Lisboa, Julho de 2021

Resumo

Introdução. O comprometimento da marcha é uma característica clínica da Esclerose Múltipla (EM), sendo uma das principais causas de incapacidade. Por esse motivo, a sua avaliação é essencial na gestão da condição clínica, no tratamento e na avaliação da eficácia da reabilitação.

Objetivo. O objetivo deste estudo é identificar quais são os instrumentos disponíveis, na literatura, para avaliar a marcha dos indivíduos com diagnóstico de EM.

Método. A pesquisa foi realizada na *MEDLINE, Web of Science, Embase e Scopus*, seguindo as diretrizes *Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA).* O protocolo do estudo foi registrado no *International Prospective Register of Systematic Reviews (PROSPERO).* Na seleção dos estudos foram incluídos artigos escritos em inglês e que representem a validação de instrumentos de avaliação da marcha, exclusivamente, para adultos com este diagnóstico. Dois autores extraíram, independentemente, os dados de confiabilidade, erro de medição e validade. E avaliaram a qualidade metodológica dos artigos, através *COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN).* A síntese dos dados foi realizada para determinar o nível de evidência de cada propriedade psicométrica em cada instrumento de medida.

Resultados. A pesquisa identificou 556 estudos. Após a remoção dos duplicados e artigos que não cumpriam os critérios de elegibilidade, foram selecionados 38 estudos. Nestes identificámos 25 instrumentos que se distribuem por 5 categorias comumente utilizadas (baseados no desempenho, no indivíduo, clínicos, no observador e biomarcadores).

Conclusão. São inúmeros os instrumentos que permitem avaliar a marcha de diferentes formas e contextos. É necessário conhecer quais são essas medidas para selecionar as que melhor se adequam ao ambiente e objetivo pretendido.

Este trabalho concluiu que os sensores inerciais, o T25FW, a 6MWT e MSWS-12 são medidas pertinentes na avaliação deste construto nestes indivíduos, visto que apresentam alta evidência quanto à confiabilidade.

Índice

Índice de tabelas

Índice de Figuras

Índice de apêndices

Lista de abreviaturas

10MTW - 10-metre Timed Walk

100MTW - 100-metre Timed Walk

2MWT - Two-Minute Walk Test

30MTW - 30-metre Timed Walk

5UTT - 5 U-Turn Test

6MWT - Six-Minute Walk Test

A

AI - Ambulation Index

C

ClinROMs - Clinician-Reported Outcome Measures

CNS - Central Nervous System

COSMIN - COnsensus-based Standards for the selection of health Measurement INstruments

D

DGI - Dynamic Gait Index

E

EDSS - Expanded Disability Status Scale

F

FGA - Functional Gait Assessment

G

GAIT - Gait Assessment and Intervention Tool

GRADE - Grading of Recommendations Assessment, Development and Evaluation

I

ICC - Intraclass correlation coefficient

ICCinter - Intraclass correlation coefficients between raters

ICCintra - Intraclass correlation coefficients within-rater

ICF - International Classification of Functioning, disability and health

L

LoA - Limits of Agreement;

M

MIC - Minimal Important Change

MS - Multiple sclerosis

MSSymS - MS Symptom Scores

MSWS-12 - Multiple Sclerosis Walking Scale-12

N

NPWT - Narrow Path Walking Test

O

ObsROMs - Observer-Reported Outcome Measures

P

PerFOMs - Performance-Based Outcome Measures

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PROMs - Patient-Reported Outcomes

PROSPERO - Prospective Register of Systematic Reviews

PwMS - Person with MS

R

RVGA - Rivermead Visual Gait Assessment

S

SDC - Smallest Detectable Change

SEM - Standard Error of Measurement;

SMSW - Short Maximum Speed Walk Test

SSST - The Six Spot Step Test;

T

T25FW - Timed 25-Foot Walk

TUG - Timed Up and Go Test

V

VPC - Visual Perceptive Computing

W

WA-VAS - Walking Ability Visual Analogue Scale

1. Introdução

A perturbação da marcha é uma característica clínica da Esclerose Múltipla (EM), resultante da combinação de vários sintomas e défices comuns, tais como fadiga, fraqueza muscular, espasticidade, ataxia e problemas de equilíbrio. A sua etiologia pode ser neurológica, sensorial, ou motora, sendo associada a lesões difusas do Sistema Nervoso Central (SNC). Portanto, é uma das manifestações mais visíveis desta patologia.

Todos estes sintomas podem ser alvo da intervenção por parte fisioterapeuta, de forma a melhorar a sintomatologia associada a esta condição clínica. Logo, a intervenção deste profissional é um meio para gerir as mudanças na marcha das pessoas que sofrem desta patologia. Assim confirma-se a relevância da avaliação regular da marcha nestes indivíduos para permitir uma melhor gestão, tratamento, monitorização da atividade clínica da doença e avaliação da eficácia da reabilitação.

Desta forma, o principal objetivo deste trabalho é identificar quais são os instrumentos disponíveis na literatura para avaliar a marcha em indivíduos diagnosticados com EM. Adicionalmente pretendemos: caracterizar esses instrumentos, analisar as propriedades psicométricas de cada ferramenta de avaliação e discutir sobre as suas vantagens e desvantagens de cada medida de avaliação.

O procedimento adotado, durante esta revisão sistemática, está de acordo com *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) e pelos *COnsensus-based Standards for the selection of health Measurement INstruments* (COSMIN).

Inicialmente foi realizada uma pesquisa nas seguintes bases de dados eletrónicas MEDLINE via PubMed, Web of Science, Embase e Scopus para identificar o maior número de ferramentas de medição disponíveis. Após a seleção dos estudos que contém as validações e informações dos instrumentos de medida, decorreu a extração dos dados que incluí: identificação dos instrumentos utilizados para avaliar a marcha nos indivíduos com este diagnóstico; caracterização de cada instrumento de avaliação; análise das suas propriedades psicométricas; e, finalmente, a discussão das vantagens e desvantagens de cada ferramenta utilizada para avaliar a marcha nestes indivíduos.

Posteriormente à identificação das ferramentas mais utilizadas, estas serão agrupadas consoante a sua caracterização, ou seja dados: obtidos por clínicos (ClinROMs); baseados no desempenho (PerFOMs); relatados pelo observador (ObsROMs); auto-reportados (PROMs); de biomarcadores.

Seguidamente, as propriedades de medição dos instrumentos identificados nos estudos incluídos foram analisadas quanto: aos coeficientes de correlação intraclasse (ICCs)

para estimar diferenças de medição intra-avaliador (ICCintra) e entre avaliadores (ICCinter) para a confiabilidade; SEM (*standard error of measurement*) para estimar o erro de medição; SDC (*smallest detectable change*) e validade de construto e critério.
Assim, e conforme recomendado, foram seguidas as instruções da ferramenta *COSMIN Risk of Bias* para avaliar a qualidade dos estudos incluídos quanto à confiabilidade, erro de medição e validade. Posteriormente, a qualidade da evidência foi categorizada como alta, moderada, baixa ou muito baixa, utilizando a abordagem modified *Grading of Recommendations Assessment, Development and Evaluation* (GRADE).
Esta dissertação contém elementos pré-texuais (Índice e Resumo) e elementos pós-textuais (Apêndices). Bem como elementos textuais que incluem o artigo, na língua inglesa, uma introdução e uma breve conclusão e considerações finais.
Por último, acreditamos que o nosso trabalho, tal como descrito neste manuscrito, será do interesse para o campo da reabilitação, pois é importante que os profissionais de saúde, incluindo fisioterapeutas, conheçam todas os instrumentos utilizados para avaliar a marcha nesta população, de modo a quantificar os seus défices e, assim, planear intervenções tendo em conta essas dificuldades.

2. Article

Abstract

Background. Gait impairment are considered a significant cause of disability and are a clinical feature of Multiple Sclerosis. For this reason, their regular assessment is essential for better management and treatment and to evaluate the effectiveness of rehabilitation.

Objective. The aim of this study is to know what instruments are available in the literature to assess gait in patients diagnosed with MS.

Data Sources and Searches. A complete literature search was conducted in MEDLINE, Web of Science, Embase, and Scopus, following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The study protocol was registered at the International Prospective Register of Systematic Reviews (PROSPERO).

Study Selection. The authors selected studies written in English and representing validation of gait assessment tools exclusively for adults with a diagnosis of MS.

Data Extraction and Quality Assessment. Two authors independently extracted the data for reliability, measurement error, validity and assessed methodological quality using the COnsensus-based standards to select health Measurement Instruments (COSMIN).

Data Synthesis and Analysis. All authors performed data synthesis to determine the level of evidence per measurement property per tool.

Results. The search identified 556 studies. After removing duplicates and articles that did not meet the criteria, 38 studies remain. In this studies were identified 25 instruments for measuring gait in people with MS. There are five categories of the most commonly used instruments (performance-based, patient-reported, clinician-reported observer-reported, and biomarker outcome).

Conclusion. There are countless instruments (PerFOMs, PROMs, ClinROMs, ObsROMs, and Biomarker Outcomes) that allow evaluating gait in different ways and contexts. It is necessary to know which measures exist and choose the best ones that suit the environment and the objective. This work concludes that the T25FW, 6MWT, MSWS-12, and wearable, wireless and inertial sensors appear to provide gait assessments as they have high-reliability evidence.

Introduction

Multiple sclerosis (MS) is a chronic, progressive inflammatory condition that primarily affects the central nervous system (CNS). It is the most common neurological disease in young adults, as it is typically diagnosed between the ages of 20 and 50, with the majority of affected individuals developing the symptoms in their 20s and 30s.[1,2] Because the onset of the disease occurs in early adulthood, individuals with MS often faced a lifetime of disability that significantly impacted the quality of life, career development, family life, and social integration.[1,2,3] One of the factors that most contributed to the socioeconomic burden and the status of a person with MS (PwMS) is walking impairment.[3,4]

Walking is among the most basic and ancient human species functions and is closely related to the population's autonomy and quality of life.[5] For this reason, walking is included in the International classification of functioning, disability and health (ICF), which defines as ''moving along a surface on foot, step by step, so that one foot is always on the ground, such as when strolling, walking forwards, backward, or side-ways''.[5]

Fifty percent of PwMS have difficulty walking in the first 15 years and require a walking aid.[2,3] According to LaRocca (2011), 85% of PwMS have difficulty walking and have ambulatory deficits. This study found that 70% of PwMS with difficulty walking agreed that this was the most challenging aspect of their MS.[3]

It may seem somewhat artificial to separate the management of symptoms that contribute to walking limitations from global gait disturbance.[6]

Gait disorders are perceived as the leading cause of disability. They are a clinical hallmark of MS, often resulting from the combination of multiple common symptoms and deficits such as fatigue, muscle weakness, spasticity, ataxia, and balance problems;[7,8] and can have various etiologies[6], including neurological, sensory, or motor impairments associated with diffuse lesions of CNS. However, cognitive and visual impairment may also play a role.[5]

So, gait pattern functions are defined as ''movement patterns associated with walking, running or other whole-body movements,'' and gait impairments may include ''spastic gait, hemiplegic gait, paraplegic gait, asymmetric gait, limping, and stiff gait patterns.[5]

These findings recognize that walking difficulty is a significant symptom of MS, documentation of gait impairment is essential to allow better management, treatment of mobility difficulties, monitoring the clinical activity of the disease, and evaluation of the effectiveness of rehabilitation.[3,5,6]

Since gait differs between patients according to their experience and disease severity, it cannot be studied with a single test alone.[5] The challenge lies in achieving a suitable compromise between recording a maximum of data to give a valuable overview of the gait and clinical feasibility in routine practice and investigation.[5]

To date, changes in gait have been studied in laboratory and clinical settings.[9,10] The majority of research on these construct dysfunction in PwMS has focused on subjective clinical assessments and spatiotemporal parameters.[10]

Examinations of spatiotemporal parameters have revealed that PwMS walk slower, taking shorter, slower steps, and spending more of their gait cycle in double-support than healthy controls, and these impairments scale with disability.[3,9,10] Also, performance walking tests are common in MS and routinely demonstrate that PwMS walk slower than their peers without MS.[10]

So, it is essential that healthcare professionals, including physiotherapists, are aware of all the tools used to assess gait in this population, quantify the deficits, and plan interventions with these difficulties in mind.[3,5,6,10]

To date, no systematic literature review has been performed on the gait assessment instruments available in patients diagnosed with MS. This study aimed to find the gait assessment measures in patients diagnosed with MS. It will also characterize and analyze each instrument's features and psychometric properties to assess gait in PwMS.

Materials and method

The study protocol was registered at the International Prospective Register of Systematic Reviews (PROSPERO; http://www.crd.york.ac.uk/PROSPERO; registration number: CRD42020216570). We followed the working procedure developed by the COnsensus-based Standards for the Selection of Health Measurement INstruments (COSMIN) for conducting a systematic review of measurement properties.[11] The systematic review was reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[12]

- Data sources and searches

The objective of our study was to highlight the measurement instruments to assess gait in MS.

A complete literature search was performed in MEDLINE via PubMed, Web of Science, Embase, and Scopus from 1 December 2020 to 31 March 2021. Full texts of all articles judged to be of possible interest based on title and abstract were retrieved by the authors. The combinations of keywords used were: “multiple sclerosis” with “gait” or “walking” and “assessment” or “test” or “scale” and “reliability” or “validity” or “measurement error.” The

authors conducted keyword searches in the title and abstract. Can find the search strategy in Appendix 1.

Based on the titles and abstracts of the articles, we checked the population and gait criteria. Authors independently selected papers for inclusion based on their titles and abstracts; removed duplicates; and analyzed copies of the full text for eligibility according to criteria established a priori. Furthermore, we searched the reference list of all identified articles for additional studies.

In selecting the measurement tool, we considered three main aspects: (i) appropriateness for the target population, (ii) practical aspects of test administration, and (iii) psychometric properties.

Study selection

The included articles selected by two authors are according to the following inclusion criteria: (1) studies validating instruments assessing qualitative or quantitative gait and its psychometric properties, or comparisons between 2 or more gait instruments (which may include scales or tests administered by health professionals, or self-report, performance test, laboratory motion analysis, and spatiotemporal parameters) exclusively for neurological patients diagnosed with MS; (2) articles written in English; (3) the population studied consists of adult patients (aged 18 years) with MS; (4) regardless of MS type, time since diagnosis, but with a degree of disability less than 7.0 in EDSS.[13]

This systematic review, according to the following exclusion criteria, excluded articles that: (1) include individuals with a diagnosis other than MS; (2) did not have gait assessment instruments validated for this population; (3) not describe instruments (4) which it was not possible to obtain the full text; (5) was systematic reviews; (6) which is the instrument used to measure other outcome (for example, in randomized controlled trials)[14] (7) which the instrument used in a validation study of another instrument; [14] and (8) do not include humans.

- Data extraction

After the studies were selected, data extraction took place. Each reviewer independently read, assessed, and extracted data from all included articles.

Data extraction from the included articles was performed by describing: the information in the research studies, the characteristics of the gait assessment instruments, and the psychometric properties of the gait assessment tools.

One of the measurement properties of the instruments identified in the included studies was reliability.[15] This property was measure by intraclass correlation coefficients (ICCs) for intra-rater estimation (repeatedly presenting the same observations to 1 observer) and between evaluators (when submitting the same observations to 2 or more observers).[15] The other properties were SEM to estimate measurement error, SDC to estimate which change in measurement result will be clinically relevant and construct and criterion validity.[15]

- Quality assessment

In this work, we use the COSMIN Risk of Bias tool checklist to assess the methodological quality of measurement properties studies.[14,15]

Considering the tools found in the research, it was the ideal instrument to assess the properties of the studies because they are complex tools with more sources of variation, which can potentially influence the scores.[14,15]

Thus, and as recommended, we follow the instructions in the COSMIN Risk of Bias tool user manual to assess study quality for reliability and measurement error.[16]

The COSMIN Risk of Bias checklist includes standards on design requirements and preferred statistical methods organized in boxes per measurement property.[15] So, we assessed using nine standards for reliability and 8 for a measurement error.[16] The Reliability box contains six standards on design requirements and three standards on the preferred statistical methods for studies for reliability. The Measurement Error box includes the same six standards about the design requirements and two standards about the preferred statistical methods.[16]

Also, the quality of studies is assessed for validity, follow the instructions in the PROMs measures user guide.[14]

To assess the quality of a study, should score each standard, and the worst-score-count method is applied to determine the risk of bias (Appendix 2).[15]

We determined an overall score for the methodological quality of each measurement property by taking the lowest rating of one of the items in a box. Two reviewers independently rated the methodological quality of each study, rating each item as very good, adequate, doubtful, and inadequate. Resolved disagreements through discussion.

- Data synthesis and analysis

During the data synthesis, each reviewer read, synthesized, and analyzed the data independently. The characterization of the instruments is divide according to the different types of measurement instruments present, such: patient-reported outcome measure (PROMs), observer-reported outcome measures (ObsROMs), clinician-reported outcome measurement (ClinROMs), performance-based outcome measurement (PerFOMs), and biomarker outcomes– also called laboratory values.

The authors summarized all evidence on psychometric properties, and outcomes were rated as “positive,” “negative,” or “indeterminate” according to result rating criteria accepted by consensus in an international Delphi study (Tab. 1).[17]

Although minimal important change is not considered a measurement property in the COSMIN taxonomy, it included it in the data extraction because this measure is necessary to determine the level of evidence for measurement error.

Finally, we categorize the quality of the evidence as high, moderate, low, or very low, using the modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach.[17] Based on the:

1) risk of bias (i.e., the methodological quality of the studies);

2) inconsistency (i.e., unexplained inconsistency of results across studies);

3) imprecision (i.e., the total sample size of the available studies);

4) indirectness (i.e., evidence from different populations than the population of interest in the review).

Following this review, we recommend the measurement protocol with the best evidence for a given measurement tool.[17]

Table 1. Criteria for Result Ratings of Measurement Properties[17]

Table 1. Criteria for Result Ratings of Measurement Properties[17]

Measurement Property	Rating	Criterion for Result Rating
Reliability	+	ICC or (weighted) Kappa ≥ 0.70
	?	ICC or (weighted) Kappa not reported
	−	ICC or (weighted) Kappa < 0.70
Measurement error	+	SDC or LoA or CV*√2*1.96< M(C)IC; % specific agreement > 80%
	?	MIC not defined
	−	SDC or LoA or CV*√2*1.96> M(C)IC; % specific agreement < 80%
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis
	?	No hypothesis defined (by the review team)
	−	The result is not in accordance with the hypothesis
Criterion validity	+	Correlation with gold standard ≥0.70 OR AUC≥0.70
	?	Not all information for '+' reported
	−	Correlation with gold standard <0.70 OR AUC<0.70
Internal consistency	+	At least low evidence for sufficient structural validity AND Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or subscale
	?	Criteria for "At least low evidence for sufficient structural validity" not met
	-	At least low evidence for sufficient structural validity AND Cronbach's alpha(s) < 0.70 for each unidimensional scale or subscale
A Minimal detectable change		

ICC = intraclass correlation coefficient; LoA = limits of agreement; MIC = minimal important change; SDC = smallest detectable change; CV = coefficient of variation + = positive rating;? = indeterminate rating; − = negative rating;

- Funding source

The funders played no role in the design, conduct, or reporting of this study.

Results

- Included studies and characteristics

Searches of electronic data sources identified 556 studies. After removing duplicates and articles where the title and abstract did not meet the screening criteria, we evaluated 84 full texts. Additional studies were excluded, resulting in 38 papers that met the inclusion criteria. Also, two studies were included through a manual search of the reference lists. The included articles identified 25 instruments for measuring gait in people with MS (Figure 1).

Using data from the articles found through the search, we elaborate Table 2 - Characteristics of Included Studies in alphabetical order of first authors' names.

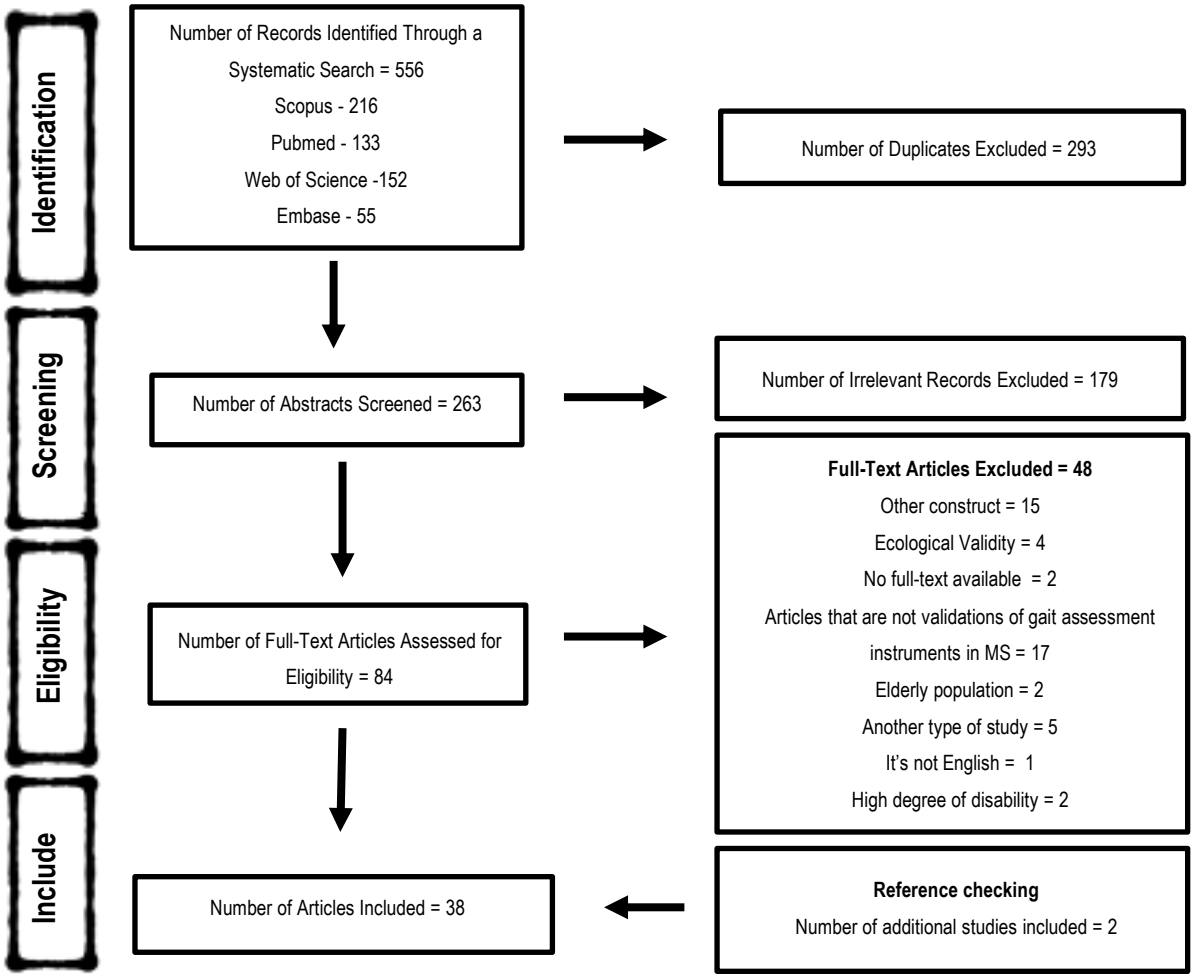

Figure 1.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)[12] flow diagram of search and study selection.

The most frequently cited gait assessment tools at MS in the selected articles were: Timed 25-Foot Walk (T25FW), Six-Minute Walk Test (6MWT) and Multiple Sclerosis Walking Scale-12 (MSWS-12).[21,27,35,39,42,45,49,48,50,51,54] Following were mentioned: The Timed Up and Go Test (TUG), The Six Spot Step Test (SSST) and Wearable, wireless and inertial sensors.[18,19,21,23,26,27,30,33,41,48,49,53]

When we analyzed the included articles, we found that the sample size ranged from 10 to 1034 participants. All studies included participants of both sexes. However, in each study, the percentage of women was always higher than that of men. The average age of the participants ranged between 33 and 56 years. And that the most common type of MS was relapsing-remitting. Most of the included participants were diagnosed with the disease at least 7 to 16 years ago.[18,20-23]

Of the included, 33 studies evaluated participants with EDSS. Their range of scores was between 0 and 6.5, meaning that individuals need a mobility aid to walk about 20m.[14] Thus, it appears that participants in most studies could still walk, even with a walking aid.

Table 2. Characteristics of Included Studies in Alphabetical Order of First Authors 'Names

Table 2. Characteristics of Included Studies in Alphabetical Order of First Authors 'Names

Study; Year of Publication/ Country	Sample Size (n PwMS total (n Female (F) / Male (M)))	Age (years), Mean (SD) [Range]	Disease duration since diagnosis (years) (SD)[Range]	EDSS score Mean (SD) [Range]	MS type (n or %)	Use of walking aid (n or %)	Instrument name
Ader et al (2020)[18] Ireland	37 (23F / 14M)	45.1 (9.9)	7.4 (7.7)	0: n= 13 1-1.5: n= 13 ≥2: n= 11			2 Inertial Sensors
Angelini et al (2020)[19] U.K	57	56.0 (9.3)		5.5 [3.0–6.5]		Unilateral: 25% Bilateral: 14%	Walked for 6 minutes wearing 3 tri-axial inertial sensors
Behrens et al (2014)[20] Germany	22 (13F/ 9M)	43 (9)		3.0 [0.0–6.0]			SMSW + VPC using Laboratory System with Kinect sensor
Bennett et al (2017)[21] USA	50 (30F / 20M)	53.2 (9.23) [32–73]	13.29 (10.8) [1–49]	5.02 (1.67) [1.5–6.75]			DGI; T25FW; 6MWT; 2MWT; TUG; MSWS-12
Bourke et al (2020)[22] Switzerland	76 (53F / 23M)	39.5 (7.9)	11.3 (7.0)	2.4 (1.4)	RRMS: 90.8% SPMS: 5.3% PPMS: 3.9%		Samsung Galaxy 7 smartphone + 2MWT
Callesen et al (2017)[23] Denmark	38 (26F / 12M)	53.6 (9.7) [no data]	15 [2–44]	≤3.5: n= 11 4-5.5: n= 12 6-7.5: n= 15	RRMS: 42% SPMS: 24% PPMS: 26%		SSST
Cheng et al (2021)[24] Switzerland, USA, Canada, Spain and Germany	76 (53F / 23M)	39.5 (7.9)	11.3 (7.0)	2.4 (1.4)	RRMS: 90.8% SPMS: 5.3% PPMS: 3.9%		5UTT
Coulter et al (2017)[25] U.K	20 (11F / 9M)	53.7 (7.4)		5.85		14	ActivPAL3 accelerometer
Craig et al (2017)[26] USA	15 (12F / 3M)	48.2 (8.7)	12.2 (5.9)	1.89 (0.98)	All RRMS		Wireless inertial sensor during TUG
Decavel et al (2018)[27] France	58 (37F / 21M)	50.7 (11.9)	14.1 (9.9)	5.2 (1.1)	RRMS: 12 SPMS: 26 PPMS: 20		T25WT; 6MWT; TUG
Filipovic´ Grc¨ic´ et al (2013)[28] Croatia	82 (61F / 21M)	37.8 (9.9)	9.0 (6.3)	2.5 [0–6.5]		13	WA-VAS; T25FW, SSST, 2MTW, MSWS-12
Fitzgerald et al (2019)[29] USA	102 (78F / 24M) 60 (45F / 15M)	47.1 (12.7) 48.3 (11.2)	13.1 (10.0) 13.3 (9.9)	3.0 [2.0–4.5]	RRMS: 62; 48 SPMS: 23; 7 PPMS: 9; 0		SymptoMScreen; TUG; T25FW
Flachenecker et al (2019)[30] Germany	102 (69 F / 33M)	43.0 (11.6)	10.1 (10.5)	4.0 [1.0–7.0]	RRMS: 68 SPMS: 22 PPMS: 12		A foot-worn sensor + T25FW
Forsberg et al (2013)[31] Sweden	81 (62F / 19M)	49 (11)	12 (8)	[1.0–6.0]	RRMS: 53 SPMS: 24 PPMS: 4	Indoors: 12 Outdoors: 38	DGI
Forsberg et al (2017)[32] Sweden	87 (69F / 18M)	53.8 (10.9) [28–75]	15.0 (9.8) [1–46]	3.5 (2)	RRMS: 40 SPMS: 37 PPMS: 10	Indoors: 12 Outdoors: 52	FGA
Fritz et al (2016)[33] USA	28 (11F / 17M)	20-50 years: n= 1 9.9 (5) 51-70 years: n=1 10.6 (3.1)	≤ 10 years: n= 13 9.8 (4.5) >10 years: n= 15 10.7 (3.5)	<4: n = 14 8.3 (2.3) ≥ 4: n= 14 12.3 (4.3)			SSST; TUG; 2MWT;T25FW
Gholami et al (2017)[34] Canada	10 (9F / 1 M)	61		5,05 [1- 6.5]		Most patients	Kinect camera for Windows
Goldman et al (2008)[35] USA	40 (31F / 9 M)	42.10 (8.1)	7.5 (6.5)	[0- 6.5]	RRMS: 29 SPMS: 10 PPMS: 1		6MWT
Gor-García-Fogeda et al (2020)[36] Spain	35 (23F/12M)	47.7 (11) [23-70]	11.2 (7.02) [1-28]	4.32 (1.4) [1-6]	RRMS: 62.85%	31	GAIT
Gor-García-Fogeda et al (2020)[37] Spain	35 (23F / 12M)	47.7 (11) [23-70]	11.2 (7.02) [1-28]	4.32 (1.4) [1-6]	RRMS: 62.85%	31	GAIT

Grobelny et al (2017)[38] Japan	83 (49F / 34M)	40.7 (14.2) [18-66]		2.8 [0.0-6]			SMSW + VPC using Laboratory System with Kinect sensor
Hobart et al (2003)[39] U.K	602	51 (12) [23-87]	19 (12) [1-56]	[0.0-6]			MSWS-12; T25FW
Kalron et al (2016)[40] Israel	229 (143F / 86M)	43.4 (12.4)	6.2 (7.1)	1.7 (0.7)	RRMS: 215 Progressive: 14		2MWT; 6MWT; TUG; T25FW; MSWS-12; walk ratio using the GAITRite
Kalron et al (2017)[41] Israel	285 (176F / 109M)	44.5 (13.4)	8.1 (8.1)	3.5 (1.6)	RRMS: 262		TUG
Learmonth et al (2013)[42] USA	82 (62F / 20M)	49.2 (9) [27-64]	11.8 (8.2) [0.5–32]	3.5 (2) [0–6.5]	RRMS: 65 SPMS: 10 PPMS: 7	Unilateral: 10 Bilateral: 10	T25FW; 6MWT; MSWS-12; accelerometry
Lord et al (1998)[43] U.K	3[a] 2[b] 7[b]						RVGA
McConvey & Bennett (2005)[44] USA	10			[2.0-6.0]			DGI
McGuigan & Hutchinson (2004)[45] Ireland	149 (109F / 40M)	46.6 (11.1) [19-77]	12.8 (8.8) [1-44]	4.0 (1.9) [0-7.0]	PPMS: 11.1% RRMS: 59.3% SPMS: 29.6%	44	MSWS-12
Molt et al (2010)[46] USA	26 (22F/ 4M)	43.1 (11.9)	11.6 (8.4)		RRMS: 26		Accelerometer; MSWS-12; 6MWD
Nieuwenhuis et al (2006)[47] Denmark	151 (106F / 45M)	42.2 (12.4)		4.3 (2.0) [0-6.5]		135	SSST
Nilsagard et al (2007)[49] Sweden	43 (30F / 13M)	52 (9)		≤ 4: 19 > 4: 24	RRMS: 22 PPMS: 8 SPMS: 13		10MWT; 30MWT; TUG
Paltamaa et al (2005)[50] Finland	19 (9F / 10M)[c] 9 (6F/ 3M)[b]	42.7 (9.2) [24-58][c] 48.9 (8.8) [34-58][b]	5.8 (5.7) [1-19][c] 9.3 (8.3) [1-24][b]	[0.0-6.5]	RRMS: 14[c]; 5[b] PPMS: 3[c]; 2[b] SPMS: 2[c] ; 2[b]	5[c] 4[b]	10MTW; 6MWT
Phan-Ba et al (2011)[51] USA and Belgium	141 (68.8% F)	40.0 (12.4) [14-74]		2.5 [0-5.5]	RR: 90.3% PP: 9.7%		T25FW; 100MTW
Rosenblum & Melzer (2017)[52] Israel	30 (18F / 12 M)	32.60 (5.67)	6.62 (6.61)	3.08 (1.16)	RRMS: 27 PPMS: 2		NPWT
Sandroff et al (2014)[53] USA	96 (77F / 19M)	52.7 (11.1)	11.8 (10.0)	4.5 [2-6.5]	RRMS: 79 Progressive: 13	34	SSST
Scalzitti et al (2018)[54] USA	28 (21F / 7M)	51.0 (8.7)	14.8 (13.4)	[0–6.5]	RRMS: 22 SPMS: 3	12	6MWT; 2MWT; TUG;
Sharrack et al (1999)[55] U.K	64 (42F / 22M)[b] 35 (20F /15M)[a] 50 (31F / 19M)[d]	40 [22-74][b] 38 [24-51][a] 36 [24-51][d]	13 [2-35][b] 11 [2-17][a] 12 [2-17][d]	4.5			AI
Zhang et al (2020)[56] Australia	1034 (77.9%F)	55.6 (11.4)	13				MSSymS

Persons with multiple sclerosis (PwMS); Expanded Disability Status Scale (EDSS); Relapsing–remitting MS (RRMS); Primary progressive MS (PPMS) Secondary Progressive MS (SPMS); Short Maximum Speed Walk test (SMSW); The Dynamic Gait Index (DGI); Timed 25-Foot Walk (T25FW); Six-Minute Walk Test (6MWT); Two-Minute Walk Test (2MWT); Multiple Sclerosis Walking Scale-12 (MSWS-12); The Timed Up and Go Test (TUG); The Six Spot Step Test (SSST); 5 U-Turn Test (5UTT); Walking Ability Visual Analogue Scale (WA-VAS); The Functional Gait Assessment (FGA); Gait Assessment and Intervention Tool (GAIT); Visual perceptive computing (VPC); Rivermead Visual Gait Assessment (RVGA); 10-metre Timed Walk (10MTW); 30-metre Timed Walk (30MTW); 100-metre Timed Walk (100MTW); Narrow Path Walking Test (NPWT); Ambulation Index (AI); MS Symptom Scores (MSSymS); *a*- intra-rater study; *b*- inter-rater study; *c*- Test-retest study; *d*- Validity study.

- General characteristics and clinical utility of the tools selected

A brief description of the objective of each instrument is the presented. In addition, Table 3 - Overview of the general characteristics of gait measurement instruments in MS - provides further information on each instrument.

- PerFOMs:

All tests require participants to follow standardized instructions and, for some tests, to walk as quickly and safely as possible. A walking aid may be used if necessary.

- T25FW: Assessment of walking speed (feet/s) and time to walk 25 feet (s) on a marked 25-foot path. [21,27,42,48,51]

- 10-metre Timed Walk (10MTW): Evaluate walking speed and record time to walk 10 m (sec) is recorded.[49,50]
- 30-metre Timed Walk (30MTW): Assess walking speed and record time to walk 30 m (sec) is recorded.[49]
- 100-metre Timed Walk (100MTW): Rate walking speed and record time to walk 100 m (sec) is recorded.[51]
- 6MWT: Assessment of the walking speed (m/min) and the total distance (m) covered in 6 minutes, in 22m hall.[21,27,35,42,50,54]
- Two-Minute Walk Test (2MWT): It's similar to 6mWT, but for 2 minutes.[21,54]
- TUG: The test analyze other functions besides walking, such as standing up from a chair or turning around an object.[21,27,41,49]
- SSST: Evaluates several factors that contribute to mobility, including coordination and balance.[23,33,48,53]
- The Dynamic Gait Index (DGI): Measure and document the patient's ability to respond to changing task demands during walking.[21,31,44]
- Functional Gait Assessment (FGA): This measure evaluates postural stability and an individual's ability to perform multiple motor tasks while walking. This test is a modification of the DGI.[32]
- Narrow Path Walking Test (NPWT): Designed to challenge balance control while walking on a narrow pathway in both single- and dual-task condition.[52]

- PROMs:

- MSWS-12: Measuring the impact of MS on the individual's walking ability.[21,39,42,45,49]
- Walking Ability Visual Analogue Scale (WA-VAS): An instrument for easy assessment of walking disability in the clinical setting.[28]
- MS Symptom Scores (MSSymS): In 2015, a measure was developed in which people with MS rate the 13 most common symptoms with a single item.[56]
- SymptoMScreen: Is a one-page instrument for rapid assessment of symptom severity in 12 neurologic areas commonly affected by MS.[29]

- ClinROMs:

- Ambulation Index (AI): This is a single-item ordinal scale that measures the speed of ambulation.[55]

- ObsROMs:

- Gait Assessment and Intervention Tool (GAIT): This is a measure of the components of coordinated gait movement and associated gait deficits. [24]
- Rivermead Visual Gait Assessment (RVGA): An objective validated measure designed, in 1998, to assess gait disturbances.[43]

- Biomarker outcomes:

- 5 U-Turn Test (5UTT): This is a quantitative, self-administered, smartphone-based test that can be incorporated into a daily routine. The goal is to examine U-turn ability while walking at a comfortable pace.[24]
- Accelerometry: Records daily steps and the average daily number of movements per day.[25,42,46]
- Walk ratio using GAITRite: Provides quantitative assessment of temporal and spatial parameters of gait.[40]
- Kinect Windows sensor: Offers the ability to analyze human gait through a camera and a depth sensor. It consists of an infrared projector and a camera that provide 3D motion capture of the entire body.[35]
- Short Maximum Speed Walk test (SMSW) + Visual perceptive computing (VPC) using Laboratory System with Kinect sensor: SMSW as a measure to analyze patients gait speed and degree of sway and VPC to evaluate walking parameters during SMSW.[20,37]
- Wearable, wireless and inertial Sensors: These devices typically include accelerometers, gyroscopes, magnetometers, or any combination of them, to objectively quantify movement patterns. They can be linked to performance tests.[18,19,26,30]
- Samsung Galaxy 7 smartphone + 2MWT: Instrumentation of the 2MWT performed in the person's home environment using a body-worn smartphone with inertial-sensor.[22]

Table 3. Overview of general characteristics of instruments to measure gait in MS

Table 3. Overview of general characteristics of instruments to measure gait in MS

Instrument Name	Instrument description	Scoring	Items	Equipment
		PerFOMs		
T25FW[21,42,51]	A subject is instructed to walk (i.e. maximal walking speed) across a clearly marked, linear 25-foot or 7.62-m course	Average (sec) of the 2 successive trials Higher numbers = slower gait speed	1	Stopwatch; Record form; Pen; Minimal space (25-foot, unobstructed corridor)
10-MTW[49,50]	A person is advised to walk in a linear line, in a 10 m course	The time (sec) to walk 10 m		Stopwatch; Record form; Pen; Minimal space (10m, unobstructed corridor)
30-MTW[49]	A person is advised to walk in a linear line, in a 30 m course	The time (sec) to walk 30 m		Stopwatch; Record form; Pen; 30m corridor
100-MTW[51]	A person is advised to walk 4 times in a 25 m course, turning 3 times	The time (sec) to walk 100 m		Stopwatch; Recording form; Pen; 25 m corridor
6MWT[21,35,42,50,54]	A person walks as fast and safe as possible for 6 min, in a 22m corridor	Distance walked, speed, the number and duration of rest during 6 minutes		Stopwatch; Record form; Pen; Trundle wheel; 2 cones; chair that can be moved along the walking course
2MWT[21,54]	The participant walks during 2 min	Distance walked, speed, the number and duration of rest during 2 minutes		Stopwatch; Record form; Pen; Trundle wheel; 2 cones
TUG[21,27,41,49]	The patient should sit on a standard arm chair, then walk to a line that is 3 meters away, turn around, walk back to the chair, and sit down	The time (sec) to complete the test		Stopwatch; Record form; Pen; Standard armchair; 3m corridor; 1 cone
SSST[23,33,48,53]	The patient is instructed to walk from one end to the other of a rectangular field (1× 5 m), while kicking five cylinder blocks	Average (sec) to complete the 4 runs (2 for each leg)		Stopwatch; Record form; Pen; 5m rectangular course; 5 blocks; Circles marked on the floor
DGI[21,31,44]	Observing of the degree of limitation, when performing specific tasks such as walking, climbing stairs and balance	4-point scale: 3 = No gait dysfunction to 0 = Severe impairment 24 points (No gait dysfunction)	8	Stopwatch; Record form; Pen; 6.1m walking are; Shoe box; 2 cones; Stairs with railing
FGA[32]	It was developed to overcome problematic issues related to the DGI. Focuses on changes in balance and gait patterns during walking tasks	3-point scale: 0 = severe impairment to 3 = normal performance 30 points (best performance)	10	Stopwatch; Record form; Pen; 6.1m walking area; Shoe box; 2 cones; Stairs with railing
NPWT[52]	The person walk within a 6 meter narrow path, without stepping out of the bounds. Doing in 3 time under single task (just walking), and 3 times under dual task (walking and performing a concurrent cognitive task)	The time (sec) to walk 6m, the number of steps and the cognitive task error	2	Stopwatch; Record form; Pen; A marked 6-m
		PROMs		
MSWS-12[21,42,45]	A self-administered questionnaire in which the participant classifies his limitations in his ability to walk because of MS, in the last two weeks	5-point scale: 1 = nothing to 5 = extremely Scores on the 12 items are summed. In a 0–100 scale: 12 is subtracted from the sum, and the result is divided by 48 and then multiplied by 100 60 points or 100% (extremely)	12	Record form; Pen
WA-VAS[28]	Patients rate their average level of walking ability over the previous 24 hours on 10 cm horizontal VAS, with a mark on that line	The distance from the left edge of the line to the mark placed by the patient	1	Record form; Pen
MSSymS[56]	Severities of 13 common MS symptoms are assess in the last 4 weeks	A Likert single-item 0–10 scale 0 = no symptom; 10 = worst symptom	1	Record form; Pen
SymptoMScreen[29]	A self-explanatory questionnaire which includes items for 12 domains commonly affected in MS. The wording of the 7 response options reflects functional changes that patients make in their daily lives	7- point Likert scales for each functional domain 0= not affected at all; 6 = total limitation/ unable to do most daily activities	12	Record form; Pen
		ClinROMs		
AI[55]	Evaluating the time and degree of assistance required to walk 25 feet	A Likert single-item 0–9 scale 9= restricted to wheelchair, unable to transfer independently	1	Stopwatch; Record form; Pen; Minimal space (25-foot, unobstructed corridor)
		ObsROMs		
GAIT[36,37]	GAIT is an observational scale that assesses kinematic parameters using video recordings	Section A - 7 points; section B - 32 points; and section C - 24 points. Ranging from 0 to 3 (maximum deviation) depending on the item. 62 points (gait deficits)	31	Video documents; Camera; 5m path; Record form; Pen
RVGA[43]	Comprises 2 observations of the arms covering both swing and stance of gait, and 18 observations of the trunk and lower limb (11 observations during the stance phase and 7 during the swing phase of gait)	4-point scale: 0 = normal - 3 = severe 59 points (grossly abnormal gait)	20	Record form; Pen
		Biomarker outcomes		
Accelerometry[42,46]	Participants use an accelerometer on their waist for days during waking hours. Can be associated with performance tests	Daily steps and the average daily movement count per day	No itens	Accelerometer; Velcro straps

5UTT[24]	Participants are instructed to walk and perform five successive U-turns (180º) going back and forth between two points, at least 4 m apart, within 60 s. A smartphone is carried in a belt bag or trouser pocket	Using a smartphone attached to the body, which measure of turn speed	Smartphone; flat level surface with 4 m; Belt bag or trouser pocket
Walk ratio using GAITRite[40]	Consists of an instrumented 4.6 m electronic walkway with compression-sensitive sensors arranged in a grid to identify the footprint contacts	Walk ratio = mean step length (mm)/cadence (step/min)	Walkway; Sensors
SMSW + VPC using Laboratory System with Kinect sensor[20,38]	Subjects started from a 5 m distance and were instructed to walk as fast as possible towards the camera. The Kinect system records live videos with a conventional camera and combines these with information from an infrared projector. The Software detects the human in the 3D video and models an artificial skeleton with 20 joints and their movement over time	Started automatically, when the subject entered the recording space, and ended at about 1.5m distance to the camera. Kinect sensor provide an approximate joint trajectory and subtle changes in the joint angles	A Microsoft Kinect sensor; Specialized cameras; Computer technology
Kinect windows sensor[34]			
Sensors[18,19,26,30]	Is a small unit that allow gait analysis, through the processing of trunk, lower back, wrist and ankle accelerations. The sensors can be attached by elastic strips, wireless sensors, or a camera and a depth sensor.	Obtains spatiotemporal gait parameters	Sensor; Velcro straps
Smartphone + 2MWT[22]	With a smartphone, the individual perform the 2MWT daily over a 24-week period	Gait characteristics (step/stride length, step/stride velocity, stance time, swing time and step/stride time)	Smartphone

-Quality assessment and data synthesis

To determine the overall quality of a study, the lowest rating of any standard in the box is taken (i.e., “the worst score counts” principle). For example, if for a reliability study one item in a box is rated as ‘inadequate, the overall methodological quality of that reliability study is rated as ‘inadequate.[16,17] And the same happens if was rated with “very good,” “adequate,” and “doubtful” and for measurement error and validity papers.

The methodological quality assessment process according to COSMIN guidelines is described in Appendix 2.

The authors extracted all information from the articles as shown in table 4 - Overview Table of quality and the results of the studies on the psychometric properties.

-Reliability

For most Perform tests, test-retest, intra-rater, and inter-rater reliability is excellent, with ICC values always greater than 0.70, except for the NPWT, because values range is from moderate to excellent (ICC = 0.49 to 0.94) and the FGA, because no data exist.

Regarding methodological quality, 12 articles examined the reliability of the Performs instruments.[21,23,27,35,42,44,49-52] We found an adequate level of reliability in five studies[21,27,49,50,52] and also five studies[35,42,44,48,51] showed a questionable level of reliability. One reliability study[23] was rated as inadequate due to methodological quality and was therefore excluded from synthesizing the best evidence. The poor scores were due to patients received treatment between the first and second applications of the tasks.

Next, for the results of the PROMs, the reliability data is found to be very good in all the studies presented (ICC or (weighted) Kappa ≥ 0.70).

Four articles examined the reliability of these instruments. One study[39] was classified as very good, one article with adequate[21] and two with Doubtful.[29,42]

AI instrument have excellent intra- and inter-examiner reliability. And the study on the instrument was rated as adequate.[55]

On the other hand, the observational instruments have the lowest reliability (ICC or (weighted) Kappa < 0.70), and the RVGA was not measured with the ICC. Methodological quality was investigated in 2 articles. One reliability study[43] was inadequate because the assessment time interval was not appropriate, and the ICC was not calculated. The other study was classified as doubtful.[36]

Finally, the biomarker results have excellent reliability values, most of them with values of ICC above 0.90.

Three studies were classified with adequate quality,[19,25,26] four as doubtful,[30,34,38,42] two as very good[22,24] and another as inadequate.[20] The latter was excluded from synthesizing the best evidence because the expert did not make the ratings blindly.

- Measurement error

The measurement error data were the most insufficient since all studies produced results above the MIC. Moreover, most studies do not present these results. Therefore more studies must be conducted to determine the MIC values for each instrument to evaluates the effectiveness of treatments and interventions.

Regarding methodological quality, six articles examined the measurement error of the Performs instruments.[23,27,42,50-52] An adequate level was found for three studies,[27,50,52] while 2 showed a questionable level of reliability.[43,51] One study[23] was classified as inadequate through methodological quality, such as the methodological reliability quality. For PROMs, only one study showed results for this property.[42] Its level of quality was classified as doubtful. On the other hand, for ClinROMs, the study found was classified as having an adequate level of evidence.[55] Finally, for biomarker outcomes, two studies were classified as very good[22,24] and the other two as Doubtful.[38,42]

- Validity

First, validity scores for Performs in a study with the TUG instrument[41] showed the lowest correlation with other instruments (T25FW and MSWS-12). For others instruments, the result is following the hypothesis.

The FGA instrument had the lowest validity data[32], with correlations with the TUG and MSWS-12 of - 0.74 and - 0.46, respectively.

Regarding methodological quality, 10 articles examined the validity of the performs instruments[21,31-33,35,41,48,51-54] Five studies have a very good level of methodological quality in criterion validity.[21,32,33,52,54] As for the construct validity, one article were classified as very good[51], 3 with adequate[31,48,54] and 2 as of doubtful quality.[41,53]

Table 4. Overview Table of quality and results of studies on psychometric properties

Table 4. Overview Table of quality and results of studies on psychometric properties										
Measurement Instrument	**Study**	**COSMIN Score**	**Reliability (+/?/-)**	**COSMIN Score**	**Measurement error (+/?/-)**	**COSMIN Score**	**Validity (Construct and Criterion) (+/?/-)**	**COSMIN Score**	**Internal Consistency (Cronbach's α) (+/?/-)**	**MIC**
PerFOMs										
T25FW	Learmonth et al (2013)[42]	Doubtful	Time(s): ICC[c] = 0.99 (+) Speed (ft/s): ICC[c] = 0.99 (+)	Doubtful	Time: SEM[f] = 1; SDC[f] = 2.8 (-) Speed: SEM[f] = 0.1; SDC[f] = 0.55 (-)				*	Time = 2.7 Speed =0.1s
	Bennett et al (2017)[21]	Adequate	ICC[c] = 0.86 (+)			Very good	TUG (ρ =−0.90)[h]; 2MWT (ρ = −0.90)[h]; 6MWT (ρ = −0.91)[h]; MSWS-12 (ρ = 0.73)[h]; DGI (ρ = −0.80)[h]; (+)		*	
	Phan-Ba et al (2011)[51]	Adequate	ICC[a] = 0.94 (+) ICC[b] =0.94 (+)	Doubtful	Overall CV[f] = 45% (?) Limited ambulation CV[f] = 46 % (?) Restricted ambulation CV[f] = 46% (?)	Very good	T100MW (r= 0.92)[i] (+) AUC = 0.88[h] (+)		*	
	Decavel et al (2019)[27]	Adequate	Normal speed: ICC= 0.77 - 0.99 (+) Fast speed: ICC= 0.79 - 0.98 (+) Dual task: ICC= 0.45 - 0.98 (+)	Adequate	Normal speed: SEM= 0.1 to 5.1 (?) Fast speed: SEM= 0.1 to 8.9 (?) Dual task: SEM= 0.1 to 7.4 (?)				*	Normal speed = 0.3 Fast speed = 0.4 Dual task = 0.3
	Nieuwenhuis et al (2006)[48]	Doubtful	ICC[c] = 0.96 (+)			Adequate	SSST (r= 0.92)[i] (+)		*	
10-MTW	Paltamaa et al (2005)[50]	Adequate	Normal speed: ICC[c] = 0.91 (+) Fast Speed: ICC[c] = 0.95 (+) Normal speed: ICC[b] = 0.93 (+) Fast Speed: ICC[b] = 0.96 (+)	Adequate	Normal speed (m/sec): CV[f] = 5.5 (?) Fast Speed (m/sec): CV[f] = 5.1 (?) Normal speed (m/sec): CV[e] = 8.6 (?) Fast Speed (m/sec): CV[e] = 4.4 (?)				*	
	Nilsagard et al (2007)[49]	Adequate	ICC[c] = 0.92 (+)						*	
30-MTW	Nilsagard et al (2007)[49]	Adequate	ICC[c] = 0.93 (+)						*	
100-MTW	Phan-Ba et al (2011)[51]	Doubtful	ICC[a] = 0.95 (+) ICC[b] =0.95 (+)	Doubtful	Overall CV[f] = 41% (?) Limited ambulation CV[f] = 41% (?) Restricted ambulation CV[f] = 40% (?)	Very good	T25FW (r= 0.92)[i] (+) AUC = 0.88[h] (+)		*	
6MWT	Goldman et al (2008)[35]	Very good	ICC[a] = 0.94 (+) ICC[b] =0.91 (+)				MSWS-12 (r= 0.81); T25FW (r = -0.83) (+)		*	
	Paltamaa et al (2005)[50]	Adequate	Distance: ICC[c] = 0.96 (+) Distance: ICC[b] = 0.93 (+)	Adequate	Distance: CV[f] = 3.9 m (-) Distance: CV[e] = 6.8 m (+)				*	
	Learmonth et al (2013)[42]	Doubtful	Distance (m): ICC[c] = 0.96 (+) Speed (m/min): ICC[c] = 0.96 (+)	Doubtful	Distance: SEM[f] = 32; SDC[f] = 88.4 (-) Speed: SEM[f] = 5.3; SDC[f] = 14.6 (-)				*	Distance = 88 Speed = 14.6
	Bennett et al (2017)[21]	Adequate	ICC[c] = 0.97(+)			Very good	TUG (ρ =−0.93)[h]; 2MWT (ρ =−0.96)[h]; T25FW (ρ =−0.91)[h]; MSWS-12 (ρ = -0.77)[h]; DGI (ρ = 0.82)[h] (+)		*	
	Decavel et al (2019)[27]	Adequate	ICC= 0.98 (+)	Adequate	SEM= 31.01; SDC= 85.7 (?)				*	
	Scalzitti et al (2018)[54]					Adequate	2MWT (r= 0.95)[i]; TUG test: (r= -0.92)[i] (+)		*	
2MWT	Bennett et al (2017)[21]	Adequate	ICC[c] = 0.96 (+)			very good	TUG (ρ =−0.90)[h]; 2MWT (ρ = 0.85)[h]; 6MWT (ρ = 0.96)[h]; MSWS-12 (ρ = -0.76)[h]; DGI (ρ = 0.85)[h] (+)		*	
	Scalzitti et al (2018)[54]					Adequate	6MWT (r= 0.95)[i]; TUG test: (r= -0.91)[i] (+)		*	

SSST	Nieuwenhuis et al (2006)[48]	Adequate	ICC[c] = 0.95 (+)			Adequate	T25FW (r=0.92)[i] (+)		*	
	Callesen et al (2019)[23]	Inadequate	Within day 1: ICC[b] = 0.98 (+) Within day 2: ICC[b] =0.95 (+) Day-to-day: ICC[b] = 0.97 (+)	Inadequate	Absolute LOA within day 1[f] = ±1.6 (-) Absolute LOA within day 2[f] = ± 2.2 (-) Absolute LOA day day-to-day[f] = ±2.0 (-) LOA[e] = 0.4 LOA[e] = 0.6				*	2
	Fritz et al (2016)[33]					Very good	T25FW (r = 0.731)[h]; TUG (r = 0.805)[h]; 2MWT (r = - 0.786)[h]; Walk velocity (r = - 0.749)[h] (+)		*	
	Sandroff et al (2015)[53]					Doubtful	TUG (ρ = 0.86)[i]; 6MWT (ρ = - 0.84)[i]; MSWS-12 (ρ = 0.67)[i]; T25FW (ρ = 0.90)[i]; Steps per day (ρ = - 0.65)[i] (+)		*	
DGI	(McConvey & Bennett, 2005)[44]	Adequate	Pearson bivariate analysis[a]: scores 0.76 to 0.99 for the 11 therapists (?) ICC[b] =0.983 (+) 8 DGI components: ICC[b] = 0.91/ 0.98 (+)						*	
	Bennett et al (2017)[21]	Adequate	ICC[c] = 0.96 (+)			Very good	TUG (ρ =−0.81)[h]; 2MWT (ρ = 0.85)[h]; 6MWT (ρ = 0.86)[h]; MSWS-12 (ρ = -0.70)[h]; T25FW (ρ = - 0.80)[h] (+)		*	
	Forsberg et al (2013)[31]					Adequate	TUG (ρ =−0.76)[i]; MSWS-12 (ρ = -0.72)[i]; T25FW (ρ = - 0.78)[i] (+)		*	
FGA	Forsberg et al (2017)[32]					Very good	TUG (ρ= -0.74) (+)[h]; MSWS-12 (ρ= -0.46)[h] (-)		*	
TUG	Bennett et al (2017)[21]	Adequate	ICC[c] = 0.97 (+)			Very good	T25FW (ρ =−0.90)[h]; 2MWT (ρ = -0.90)[h]; 6MWT (ρ = -0.93)[h]; MSWS-12 (ρ = -0.70)[h]; DGI (ρ = -0.81)[h]; (+)		*	
	Decavel et al (2019)[27]	Adequate	ICC = 0.97 (+)	Adequate	SEM= 2.8; SDC= 7.7 (?)				*	
	Karlon et al (2017)[41]					Doubtful	2MWT (r= 0.73)[i]; (+); T25FW (r= 0.59)[i]; MSWS-12 (r= 0.47)[i] (-)		*	
	Nilsagard et al (2007)[49]	Adequate	ICC[c] =0.91(+)						*	
NPWT	Rosenblum & Melzer (2017)[52]	Adequate	Single task: ICC= 0.46 to 0.94 (- to +) Dual task: ICC= 0.55 to 0.93 (- to +)	Adequate	SEM single task = 0.01 to 0.85 (?) SEM dual task= 0.01 to 1.23 (?)	Very good	Single and Dual task MSWS-12 (*r* = 0.73)[h]; 2MWT (*r* = - 0.79)[h] (+)		*	
PROMs										
MSWS-12	Bennett et al (2017)[21]	Adequate	ICC[c] = 0.86 (+)			Very good	T25FW (ρ =0.73)[h]; 2MWT (ρ = -0.76)[h]; 6MWT (ρ = -0.77)[h]; T25FW (ρ =0.73)[h]; DGI (ρ = -0.70)[h]; (+)			
	Hobart et al. (2003)[39]	Adequate	ICC[c] = 0.94 (+)					Very Good	0.94 to 0.97	
	Learmonth et al (2013)[42]	Doubtful	ICC[c] = 0.93 (+)	Doubtful	SEM[f] = 8; CV[f] = 27%; SDC[f] = 22.1 (-)					22
	McGuigan & Hutchinson (2004)[45]					Very good	T25FW (*r* = 0.65)[i] (-)			
WA-VAS	Filipovic´ Grc˘ic´ et al (2013)[28]					Very good	T25FW (ρ = 0.61)[i]; 2MWT (ρ = -0.64)[i]; daily step count (ρ = –0.51)[i] (-); SSST (ρ = 0.73)[i]; MSWS-12 (ρ = 0.75)[i]; (+)			
0–10 MS symptom	Zhang et al (2020)[56]					Very good	PDDS (r = 0.82)[h] (+)			
SymptoMScreen	Fitzgerald et al (2019)[29]	Doubtful	ICC[c] = ≥0.70 (+)			Very good Adequate	T25FW (r = 0.63)[h]; walking speed (r = 0.54)[h]; processing speed (r=−0.51)[h] (-)	Very Good	0.93	
ClinROMs										
AI	Sharrack et al. (1999)[55]	Adequate	ICC[a] = 0.93 (+) Kappa[a] = 0.59 (-) ICC[b] = 0.96 (+) Kappa[b] = 0.73 (+)	Adequate	Agreement: no difference[d] = 66% (-); difference of ≤1 point[d] = 94% (+); difference of ≤2 point[d] = 97% (+); difference of ≤3 point[d] =100% (+)	Adequate	EDSS (r= 0.68)[i] (-)		*	

					Agreement: no difference[e] = 77% (-); difference of ≤1 point[e] = 100% (+)					
ObsROMs										
RVGA	Lord et al (1998)[43]	Inadequate	LSD[a]: 10.5 (?) Kendall's coefficient of concordance[b]: *W* = 0.84 (?)			Very good	Walking time (r= 0.77)[h] (+); stride length (r= -0.61)[h] (-)		*	
GAIT	Gor-García-Fogeda et al (2020)[36,37]	Doubtful	all values: ICC[a] = 0.90 to 0.96 (+) all values: ICC[b] = 0.6 (-)			Adequate	RVGA (r= 0.97)[i] (+); 10MWT (r= 0.41)[i]; TUG (r= 0.33)[i]; FGA (r= -0.59)[i]; MSWS-12 (r= 0.57)[i] (-)		*	Right side: 1.19 Left side: 0.77
biomarker outcomes										
5UTT	Cheng et al (2020)[24]	Very Good	ICC = 0.87 (+)	Very Good	SEM= 5.54; SDC = 15.3 (-)	Very Good	T25FW (r= -0.51)[i] (-)		*	15.36
Accelerometer	Learmonth et al 2013[42]	Doubtful	count (/day): ICC[c] = 0.88 (+) steps(/day): ICC[c] = 0.91 (+)	Doubtful	count: SEM[f] = 28.4; SDC[f] = 78.4 (-) steps: SEM[f] = 726; SDC[f] = 2006.4 (+)				*	Count = 78860 Steps = 2011
	Molt et al 2010[46]					Very good	MSWS-12 (ρ=−0.68)[h]; 6MWD (ρ=0.52)[h]; oxygen cost of walking (ρ=−0.54)[h] (-)		*	
	Coulter et al (2017)[25]	Adequate	Steps: ICC[b] = 0.99 (+) walking duration: ICC[b] = 0.99 (+) upright duration: ICC[b] = 0.99 (+)			Inadequate	(no correlations or the area under curve calculated)[h] (?)		*	
walk ratio using the GAITRite	Kalron et al (2016)[40]					Doubtful	MSWS-12 (ρ = - 0.26)[h]; T25FW (ρ = - 0.31)[h]; 2MWT (ρ = 0.30)[h]; 6MWT (ρ = 0.35)[h]; TUG (ρ = - 0.24)[h] (-)		*	
Kinect Windows sensor	Gholami et al (2017)[34]	Doubtful	ICC[b] = 0.71 to 0.99 (+) ICC[b] =0.50 to 0. 61 (Step width; Knee range of motion) (-)						*	
Wearable, wireless and inertial Sensors	Ader et al (2020)[18]	Doubtful	ICC > 0.90 (+)						*	
	Angelini et al (2020)[19]	Adequate	MS moderate = 0.85 (+) MS severe = 0.90 (+)						*	
	Craig et al 2017[26]	Adequate	ICC[c] = 0.69 – 0.96 (+)						*	
	Flachenecker et al (2019)[30]	Doubtful	Self-selected speed (r[l]= 0.94 to 0.98) (+); As fast as possible (r[l]= 0.93 to 0.99) (+); Stride length (r[l]= 0.96) (+); Gait speed (r[l]= 0.97) (+)						*	
Samsung Galaxy 7 smartphone + 2MWT	Bourke et al (2020)[22]	Very god	ICC[c]:48 temporal gait features: 33 were > 0.75 (+); 44 spatial and spatiotemporal gait features: 25 were>0.75 (+)	Very god	Spatial and spatiotemporal gait: SEM = 4.16; SDC= 11.5 (-) Temporal gait: SEM = 2.35; SDC= 6.5 (-)				*	spatial and spatiotemporal = 11.53 Temporal = 6.51
SMSW + VPC using Laboratory System with Kinect sensor	Behrens et al (2014)[20]	Indequate	ICC[c] Average speed = 0.98 (+); Speed deviation = 0.90 (+); 3D deviation =0.54 (-); Left/right deviation = 0.60 (-); Up/down deviation=0.94 (+)			Adequate	T25FW (r= -0.45)[i] (-)		*	
	Grobelny et al (2017)[38]	Doubtful	ICC[c] Average speed = 0.99 (+); Speed deviation = 0.77 (+); 3D direction deviation=0.79 (+); Mediolateral deviation=0.51 (-); Vertical deviation = 0.92 (+)	Doubtful	Average speed: SEM[f] = 0.04; SDC[f] = 0.1 (?); Speed deviation: SEM[f] = 0.02; SDC[f] = 0.6 (?); 3D direction deviation: SEM[f] = 0.1;SDC[f] = 0.3 (?); Mediolateral deviation : SEM[f] = 1.1; SDC[f] = 3.0 (?); Vertical deviation: SEM[f] = 0.14; SDC[f] = 0.4 (?)	Adequate	T25FW speed (r = 0.78)[i] (+);MSWS-12 (r = -0.55)[i] (-)		*	

COSMIN = Consensus-Based Standards for the Selection of Health Measurement Instruments; ICC = intraclass correlation coefficient; LoA = limits of agreement; MIC = minimal important change; SEM = standard error of measurement; SDC = smallest detectable change; r = Pearson correlation coefficient; + = positive rating: ICC or weighted κ ≥ 0.70 (reliability), or SDC or LoA < MIC (measurement error), or the result is in accordance with the hypothesis (construct validity), or correlation with gold standard ≥ 0.70 or AUC≥ 0.70; ? = indeterminate rating: ICC or weighted κ not reported (reliability), or MIC not defined (measurement error) no hypothesis defined (construct validity), or not all information for + reporting (criterion validity); − = negative rating: criteria for "+" not met; * = not applicable

[a]Intrarater reliability = when presenting repeatedly the same observations to 1 observer; [b]Interrater reliability = when presenting the same observations to 2 or more observers; [c]Test-retest reliability = when presenting the same task to the same subjects 2 or more times.[57]

[d]Intrarater measurement error = when presenting repeatedly the same observations to 1 observer; [e]Interrater measurement error = when presenting the same observations to 2 or more observers; [f]Test-retest measurement error = when presenting the same task to the same subjects 2 or more times.[58]

[g]The SDC (defined as "the smallest change that can be detected by the measurement instrument, beyond measurement error") was calculated[18,19] by the authors of the present systematic review by √ multiplying the SEM presented in the original article by 1.96 × 2.

[h]Criterion Validity; [i]Construct Validity

[j]The MIC (defined as "the smallest change in score that is perceived as important by patients") was added for reference because this value is needed to determine the result ratings for measurement error.[19]

[l]r= averages measures of ICC.

Analyzing the results of the PROMs, we saw that all compared instruments have correlations very close to 0.70.[21,28,29,45,49,56] For many scales, the values are lower, representing moderate results in terms of criterion validity. The tool with the lowest values is SymptoMScreen.[56] Methodologically, the quality of the studies was rated as very good or adequate.

The same is true for ClinROMs and ObsROMs when the validity results are significant but moderate. The instrument GAIT has the lowest correlations except for the correlation with the RVGA.[37] The quality of the studies was classified as adequate, and one study[43] is very good.

The same happens for the biomarker outcomes. The validity is moderate, with the walk ratio instrument using the GAITRite having the lowest correlations.[41] In terms of methodological quality, two studies were classified as very good.[24,46] One is Doubtful[40] and another is inadequate because the correlations or areas under the curve are not calculated.[25]

- Internal consistency

For the PROMs instruments, internal consistency is excellent (Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or subscale) with values greater than Cronbach's α = 0.93.[29,39,49] All studies are classified as "very good" in the quality assessment.

- Summary of the level of evidence

In this paper, we summarized all information to provide an overview of the available evidence on the quality of gait measurement instruments in MS. Therefore, we used a modified GRADE approach to classify the quality of the evidence - Table 5 and Appendix 3.

Table 5. Summary of Modified GRADE approach for grading the quality of evidence

Table 5. Summary of Modified GRADE approach for grading the quality of evidence			
Measurement instrument	**Reliability**	**Measurement error**	**Validity**
T25FW	High	Low	Low
10MTW	Moderate	Low	Low
30MTW	Low	No data	No data
100MTW	Low	Low	Low
6MWT	High	Low	Low
2MWT	Moderate	No data	Moderate
SSST	Low	Very Low	Very Low
DGI	Moderate	No data	High
FGA	No data	No data	Moderate
TUG	Moderate	Moderate	Moderate
NPWT	Low	Low	Low
MSWS-12	High	Low	Moderate
WA-VAS	No data	No data	Moderate
0–10 MS symptom	No data	No data	Moderate
SymptoMScreen	Low	No data	Moderate
AI	Moderate	Moderate	Moderate
RVGA	Very Low	No data	Low
GAIT	Low	No data	Low
5UTT	Moderate	Moderate	Moderate
Accelerometer	Moderate	Low	Low
Kinect Windows sensor	Low	No data	No data
Wearable, wireless and inertial Sensors	High	No data	No data
Samsung Galaxy 7 smartphone + 2MWT	Moderate	Moderate	No data
SMSW + VPC using Laboratory System with Kinect sensor	Low	Low	Very Low
walk ratio using the GAITRite	No data	Very Low	Low

Discussion

This review aims to identify the tools used to assess the gait in MS. We conclude five categories of the most commonly used instruments (PerFOMs, PROMs, ClinROMs, ObsROMs, Biomarker Outcomes), each specific to a target. PerFORMs instruments are tools that assess walk performance by administering tests that are suggested and scored by an examiner. PROMs are subjective instruments as the patient does the assessment. While ClinROMs are performed only by clinicians, mainly neurologists. ObsROMs are observational instruments and depend on observation by an investigator. Finally, biomarker outcomes are instruments closer to assessing the actual gait or walking that the person performs in daily life. However, they can also be applicable in conjunction with performance instruments.

A large number of instruments available in the literature leads us to confirm the importance of gait assessment in this population, as it is considered one of the leading causes of disability and is a clinical feature of MS.[3]

Physiotherapist intervention is an advantage for improving gait and walking changes in MS. To enhance for better management, treatment, and assessment of rehabilitation effectiveness is essential to know the available instruments to measure these constructs in this condition.[3,5,6,10]

When deciding which instrument to use, clinicians and researchers should consider several characteristics, such as the scope, the population in which the instrument will be used, its dimensions, and the evidence presented when assessing each psychometric property.

The results obtained through evaluating the evidence, using the GRADE approach, show a high level of evidence for the reliability of the T25FW, 6MWT, MSWS-12, and wearable, wireless and inertial sensors.

Analyzing these tools, we verify that T25FW and the 6MWT have the inherent advantage of readily providing a quantitative measure of walking performance. The first considers walking speed, and the second evaluates walking distance. Also, these instruments are easily applicable in clinical practice because they are inexpensive and can be used on a wide range of patients. The T25FW is also of high practical value in the clinical setting, requiring minimal time and space. In comparison, one reasonable limitation of the 6MW test in the clinical context is the need for a walkway of sufficient length to enable comfortable walking while turns. An alternative is 2MWT, a shorter test with greater feasibility and moderate reliability and validity.

On the other hand, The MSWS-12 was designed as a disease-specific, patient-based instrument for clinical trials and practice to capture the complex impact of MS on walking

ability. Thus, it is an instrument that represents the user's perspective of their walking ability, and the results may help determine interventions that are better suited to their needs. Also easy to use, inexpensive, and can be applied in any context, whether in the clinic or at home.

In addition to these measurements, it is interesting to analyze, through direct observation, the gait performance of patients in their environment and for a prolonged period. Technological advances make it possible to fill this gap, as it is possible to assess gait outside the clinic or gait laboratory.

Wearable, wireless and inertial sensors have allowed the gait to be assessed in the real context of the user and are also easy to use as they are wearable. However, disadvantages include the cost of the devices, the need for calibration and the need for a high rate of patient compliance with the use of the device.

Finally, all the instruments evaluated are not so optimistic regarding measurement error and validity. Because for the first psychometric property, no tool ranked with high evidence, and for the second, only the DGI showed strong evidence. So, we quickly concluded that further research is needed regarding these psychometric properties to ensure that, when used, they measure the construct that proposed to be measured and that we can trust the measurement error.

The results of this review are in line with the existing literature.[5,6]

Our review presents several instruments that all have different objectives but that can be completed and complemented in gait analysis. So it is necessary to recognize the importance of regular gait assessment from both the clinical and patient perspectives.

- Strengths and limitations of this review

This article is the only systematic review that identifies, evaluates, and synthesizes the evidence on instruments used to measure gait in this population. This work was conducted according to COSMIN and PRISMA standards, using an appropriate method following the expert recommendations.[14,16]

A strength of this review was the use of the COSMIN checklist to assess the methodological quality of the articles included.[30] As far as we know, these checklist was the ideal tool to assess these studies, because is the widely used consensus-based quality assessment tool specifically designed for studies on measurement properties.

One of the limitations of this study was the including articles with participants with an EDSS score greater than 6 (need to use a walking aid). Although the quality of gait assessment can be changed using an auxiliary device (through gait pattern adaptations),

we decided to include articles with EDSS up to 7.0. Because 50% of PwMS have difficulty walking in the first 15 years and require a walking aid.[2,3] In addition, in the literature, articles on gait impairment include participants with EDSS values up to 7.0.

The small sample sizes found in most of the studies might have influenced the strength of the evidence supporting the instruments during the assessment through the COSMIN.

Due to the limitation in the search strategy of only assessing articles published in English, studies on instruments developed and used in other languages may not have been identified.

Also, the lower ratings of the results and quality of studies on measurement error may influence the conclusions on this characteristic.

Conclusion

Based on the results, we found many instruments to assess gait, so we reviewed the importance of these instruments to assess this construct.

Due to the complexity of these constructs, there are a variety of these instruments (PerFOMs, PROMs, ClinROMs, ObsROMs, and Biomarker Outcomes) that allow to assessment in different ways in different contexts.

No one tool is critical or exclusive to this assessment in this population. But it is necessary to know which ones exist and choose the best ones that suit the context and the objective. Or use the various to complement the assessment and collect subjective, objective, observational, or spatiotemporal data.

After reviewing all the instruments discussed above, the T25FW, 6MWT, MSWS-12, and wearable, wireless and inertial sensors appear to provide gait assessments as they have high-reliability results.

However, further studies on the measurement error and validity of these tests are needed to thoroughly recommend them as measurement tools in research and clinical practice for patients with this condition.

Systematic review registration

This review was registered in the International Prospective Register of Systematic Reviews (PROSPERO)(CRD4202021657).

Disclosure

The authors completed the ICJME Form for Disclosure of Potential Conflicts of Interest. The authors declared no conflicts of interest.

References

1 Compston A, Coles A. Multiple sclerosis. Lancet. 2008;372:1502–1517. https://doi.org/10.1016/S0140-6736(08)61620-7

2 Goverover Y., Genova H. M., DeLuca J., Chiaravalloti N.D. Impact of Multiple Sclerosis on Daily Life.2017.https://doi.org/10.1007/978-0-387-98188-8

3 LaRocca N.G. Impact of walking impairment in multiple sclerosis. The Patient: Patient-Centered Outcomes Research. 2011;4:189-201 https://doi.org/10.2165/11591150-000000000-00000

4 Heesen C, Böhm J, Reich C, Kasper J, Goebel M, Gold SM. Patient perception of bodily functions in multiple sclerosis: gait and visual function are the most valuable. Mult Scler. 2008;14:988—91. https://doi.org/10.1177/1352458508088916

5 Decavel, P., & Sagawa, Y. Gait quantification in multiple sclerosis: A single-centre experience of systematic evaluation. Neurophysiologie Clinique. 2019; 49(2):165–171. https://doi.org/10.1016/j.neucli.2019.01.004

6 Bethoux F. Gait disorders in multiple sclerosis. Contin Minneap Minn. 2013;19: 1007—22. https://doi.org/10.1212/01.CON.0000433286.92596.d5

7 Kelleher KJ, Spence W, Solomonidis S, et al. Ambulatory rehabilitation in multiple sclerosis. Disabil Rehabil. 2009; 31: 1625-32. https://doi.org/10.1080/09638280902751931

8 Pearson O. R., Busse M.E., van Deursen R.W.M., Wiles C.M. Quantification of walking mobility in neurological disorders. QJM. 2004;97: 463–475. https://doi.org/10.1093/qjmed/hch084

9 Shema, S., Inbar, S., Anat, H., Regev, K., Hsieh, K. L., & Karni, A. A wearable sensor identifies alterations in community ambulation in multiple sclerosis: contributors to real-world gait quality and physical activity. Journal of Neurology. 2020. https://doi.org/10.1007/s00415-020-09759-7

10 Socie M. J., Sosnoff J.J. Gait variability and multiple sclerosis. Mult Scler Int. 2013; 645197. https://doi.org/10.1155/2013/645197

11 Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010;10:22. https://doi.org/10.1186/1471-2288 -10-22

12 Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009. https://doi.org/10.1371/journal.pmed.1000097

13 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an Expanded Disability Status Scale (EDSS). Neurology. 1983; 33:1444–1452. https:// doi.org/10.1212/wnl.33.11.1444

14 Mokkink, L. B., Prinsen, C. A. C., M., Patrick, D. L., Alonso, J., C. P. M., Bouter, de Vet, H. C. W., Terwee, C B. COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) - user manual. BMC Medical Research Methodology. 2018;20(293):1–13. https://doi.org/10.21203/rs.3.rs-40864/v1

15 Mokkink, L. B., Boers, M., van der Vleuten, C. P. M., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., Terwee, C B. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Medical Research Methodology. 2020;20(293):1–13. https://doi.org/10.21203/rs.3.rs-40864/v1

16 Terwee, C. B., Prinsen, C. A. C., Ricci Garotti, M. G., Suman, A., de Vet, H. C. W., & Mokkink, L. B. The quality of systematic reviews of health-related outcome measurement instruments. Quality of Life Research. 2016; 25(4): 767–779. https://doi.org/10.1007/s11136-015-1122-4

17 Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & De Vet, H. C. W. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Quality of Life Research. 2012; 21(4): 651–657. https://doi.org/10.1007/s11136-011-9960-1

18 Ader, L. G. M., Greene, B. R., McManus, K., Tubridy, N., & Caulfield, B. Short bouts of gait data and body-worn inertial sensors can provide reliable measures of spatiotemporal gait parameters from bilateral gait data for persons with multiple sclerosis. Biosensors. 2020; 10(9), 1–10. https://doi.org/10.3390/BIOS10090128

19 Angelini, L., Hodgkinson, W., Smith, C., Dodd, J. M., Sharrack, B., Mazzà, C., & Paling, D. Wearable sensors can reliably quantify gait alterations associated with disability in people with progressive multiple sclerosis in a clinical setting. Journal of Neurology. 2020; 267(10), 2897–2909. https://doi.org/10.1007/s00415-020-09928-8

20 Behrens, J., Pfüller, C., Mansow-Model, S., Otte, K., Paul, F., & Brandt, A. U. Using perceptive computing in multiple sclerosis - The Short Maximum Speed Walk test.

Journal of NeuroEngineering and Rehabilitation. 2014; 11(1), 1–10. https://doi.org/10.1186/1743-0003-11-89

21 Bennett, S. E., Bromley, L. E., Fisher, N. M., Tomita, M. R., & Niewczyk, P. Validity and reliability of four clinical gait measures in patients with multiple sclerosis. International Journal of MS Care. 2017; 19(5), 247–252. https://doi.org/10.7224/1537-2073.2015-006

22 Bourke, A. K., Scotland, A., Lipsmeier, F., Gossens, C., & Lindemann, M. Gait characteristics harvested during a smartphone-based self-administered 2-minute walk test in people with multiple sclerosis: Test-retest reliability and minimum detectable change. Sensors (Switzerland). 2020; 20(20), 1–16. https://doi.org/10.3390/s20205906

23 Callesen, J., Richter, C., Kristensen, C., Sunesen, I., Næsby, M., Dalgas, U., Skjerbæk, A, G,. Test–retest agreement and reliability of the Six Spot Step Test in persons with multiple sclerosis. Multiple Sclerosis Journal. 2019; 25(2) 286-294. https://doi.org/10.1177/1352458517745725

24 Cheng, W. Y., Bourke, A. K., Lipsmeier, F., Bernasconi, C., Belachew, S., Gossens, C., Lindemann, M. U-turn speed is a valid and reliable smartphone-based measure of multiple sclerosis-related gait and balance impairment. Gait and Posture. 2021; 84, 120–126. https://doi.org/10.1016/j.gaitpost.2020.11.025

25 Coulter, E. H., Miller, L., McCorkell, S., McGuire, C., Algie, K., Freeman, J., … Paul, L. Validity of the activPAL3 activity monitor in people moderately affected by Multiple Sclerosis. Medical Engineering and Physics. 2017; 45, 78–82. https://doi.org/10.1016/j.medengphy.2017.03.008.

26 Craig, J. J., Bruetsch, A. P., Lynch, S. G., Horak, F. B., & Huisinga, J. M. Instrumented balance and walking assessments in persons with multiple sclerosis show strong test-retest reliability. Journal of NeuroEngineering and Rehabilitation. 2017; 14(1), 1–9. https://doi.org/10.1186/s12984-017-0251-0

27 Decavel, P., Moulin, T., & Sagawa, Y. Gait tests in multiple sclerosis: Reliability and cut-off values. Gait and Posture. 2019; 67, 37–42. https://doi.org/10.1016/j.gaitpost.2018.09.020

28 Filipović Grčić, P., Matijaca, M., Bilić, I., Džamonja, G., Lušić, I., Čaljkušić, K., & Čapkun, V. Correlation analysis of visual analogue scale and measures of walking ability in multiple sclerosis patients. Acta Neurologica Belgica. 2013; 113(4), 397–402. https://doi.org/10.1007/s13760-013-0187-5

29 Fitzgerald, K. C., Salter, A., Tyry, T., Fox, R. J., Cutter, G., Mowry, E. M., & Marrie, R. A. Validation of the SymptoMScreen with performance-based or clinician-assessed outcomes. Multiple Sclerosis and Related Disorders. 2019; 29, 86–93. https://doi.org/10.1016/j.msard.2019.01.031

30 Flachenecker, F., Heiko, Gaßner., Julius, H., De-Hyung, L., Flachenecker, P., Jürgen, W., Bjoern, E., Ralf A, L., Jochen, K. Objective sensor-based gait measures reflect motor impairment in multiple sclerosis patients: reliability and clinical validation of a wearable sensor device. Multiple Sclerosis and Related Disorders. 2019. https://doi.org/10.1016/j.msard.2019.101903

31 Forsberg, A., Andreasson, M., & Nilsagård, Y. E. Validity of the Dynamic Gait Index in people with multiple sclerosis. Physical Therapy. 2013; 93(10), 1369–1376. https://doi.org/10.2522/ptj.20120284

32 Forsberg, A., Andreasson, M., & Nilsagård, Y. The functional gait assessment in people with multiple sclerosis: Validity and sensitivity to change. International Journal of MS Care. 2017; 19(2), 66–72. https://doi.org/10.7224/1537-2073.2015-061

33 Fritz, N. E., Jiang, A., Keller, J., & Zackowski, K. M. Utility of the Six-Spot Step Test as a Measure of Walking Performance in Ambulatory Individuals with Multiple Sclerosis Presented to the Consortium for Multiple Sclerosis Centers, May 30, 2014, Dallas, TX. Archives of Physical Medicine and Rehabilitation. 2016; 97(4), 507–512. https://doi.org/10.1016/j.apmr.2015.10.100

34 Gholami, F., Trojan, D. A., Kovecses, J., Haddad, W. M., & Gholami, B. A Microsoft Kinect-Based Point-of-Care Gait Assessment Framework for Multiple Sclerosis Patients. IEEE Journal of Biomedical and Health Informatics. 2017; 21(5), 1376–1385. https://doi.org/10.1109/JBHI.2016.2593692

35 Goldman, M. D., Marrie, R. A., & Cohen, J. A. Evaluation of the six-minute walk in multiple sclerosis subjects and healthy controls. Multiple Sclerosis. 2008; 14(3), 383–390. https://doi.org/10.1177/1352458507082607

36 Gor-García-Fogeda, M. D., Tomé-Redondo, S., Simón-Hidalgo, C., Daly, J. J., Molina-Rueda, F., Cano-de-la-Cuerda, R. Reliability and Minimal Detectable Change in the Gait Assessment and Intervention Tool in Patients With Multiple Sclerosis. American Academy of Physical Medicine and Rehabilitation. 2019; 685-691. https://doi.org/10.1002/pmrj.12264

37 Gor-García-Fogeda, M. D., Daly, J. J., Molina-Rueda, F., Cano-de-la-Cuerda, R. Construct validity of the Gait Assessment and Intervention Tool (GAIT) in people with Multiple Sclerosis. American Academy of Physical Medicine and Rehabilitation. 2021; 13(3), 223-339. https://doi.org/10.1002/pmrj.12423

38 Grobelny, A., Behrens, J. R., Mertens, S., Otte, K., Mansow-Model, S., Krüger, T., Schmitz-Hübsch, T. Maximum walking speed in multiple sclerosis assessed with visual perceptive computing. PLoS ONE. 2017; 12(12), 1–13. https://doi.org/10.1371/journal.pone.0189281

39 Hobart, J. C., Riazi, A., Lamping, D. L., Fitzpatrick, R., & Thompson, A. J. Measuring the impact of MS on walking ability: The 12-item MS Walking Scale (MSWS-12). Neurology. 2003; 60(1), 31–36. https://doi.org/10.1212/WNL.60.1.31

40 Kalron, A. Construct validity of the walk ratio as a measure of gait control in people with multiple sclerosis without mobility aids. Gait and Posture. 2016; 47, 103–107. https://doi.org/10.1016/j.gaitpost.2016.04.015

41 Kalron, A., Dolev, M., & Givon, U. Further construct validity of the Timed Up-and-Go Test as a measure of ambulation in multiple sclerosis patients. European Journal of Physical and Rehabilitation Medicine. 2017; 53(6), 841–847. https://doi.org/10.23736/S1973-9087.17.04599-3

42 Learmonth, Y. C., Dlugonski, D. D., Pilutti, L. A., Sandroff, B. M., & Motl, R. W. The reliability, precision and clinically meaningful change of walking assessments in multiple sclerosis. Multiple Sclerosis Journal. 2013; 19(13), 1784–1791. https://doi.org/10.1177/1352458513483890

43 Lord, S. E., Halligan, P. W., & Wade, D. T. Visual gait analysis: The development of a clinical assessment and scale. Clinical Rehabilitation. 1998; 12(2), 107–119. https://doi.org/10.1191/026921598666182531

44 McConvey, J., & Bennett, S. E. Reliability of the dynamic gait index in individuals with multiple sclerosis. Archives of Physical Medicine and Rehabilitation. 2005; 86(1), 130–133. https://doi.org/10.1016/j.apmr.2003.11.033

45 McGuigan, C., & Hutchinson, M. Confirming the validity and responsiveness of the Multiple Sclerosis Walking Scale-12 (MSWS-12). Neurology. 2004; 62(11), 2103–2105. https://doi.org/10.1212/01.WNL.0000127604.84575.0D

46 Molt, R., Dlugonski, D., Suh, Y., Weikert, M., Fernhall, B., Goldman, M. Accelerometry and Its Association With Objective Markers of Walking Limitations in Ambulatory Adults With Multiple Sclerosis. Arch Phys Med Rehabil. 2010; 91(12: 1942-1947). https://doi.org/10.1016/j.apmr.2010.08.011

47 Motl, R. W., & Snook, E. M. Confirmation and extension of the validity of the Multiple Sclerosis Walking Scale-12 (MSWS-12). Journal of the Neurological Sciences. 2008; 268(1–2), 69–73. https://doi.org/10.1016/j.jns.2007.11.003

48 Nieuwenhuis, M. M., Van Tongeren, H., Sørensen, P. S., & Ravnborg, M. The Six Spot Step Test: A new measurement for walking ability in multiple sclerosis. Multiple Sclerosis. 2006; 12(4), 495–500. https://doi.org/10.1191/1352458506ms1293oa

49 Nilsagard, Y., Lundholm, C., Gunnarsson, L., Denison., E. Clinical relevance using timed walk tests and 'timed up and go' testing in persons with Multiple Sclerosis.Physiotherapy Research International. 2007; 12(2) 105-114. https://doi.org/10.1002/pri.358

50 Paltamaa, J., West, H., Sarasoja, T., Wikström, J., & Mälkiä, E. Reliability of physical functioning measures in ambulatory subjects with MS. Physiotherapy Research International : The Journal for Researchers and Clinicians in Physical Therapy. 2005; 10(2), 93–109. https://doi.org/10.1002/pri.30

51 Phan-Ba, R., Pace, A., Calay, P., Grodent, P., Douchamps, F., Hyde, R., Hotermans, C., Delvaux, V., Hansen, I., Moonen, G., Belachew, S. Comparison of the Timed 25-Foot and the 100-Meter Walk as Performance Measures in Multiple Sclerosis. Neurorehabilitation and Neural Repair. 2011; 25(7), 672-679. https://doi.org/10.1177/1545968310397204

52 Rosenblum, U., & Melzer, I. Reliability and concurrent validity of the narrow path walking test in persons with multiple sclerosis. Journal of Neurologic Physical Therapy. 2017; 41(1), 43–51. https://doi.org/10.1097/NPT.0000000000000161

53 Sandroff, B. M., Motl, R. W., Sosnoff, J. J., & Pula, J. H. Further validation of the Six-Spot Step Test as a measure of ambulation in multiple sclerosis. Gait and Posture. 2015; 41(1), 222–227. https://doi.org/10.1016/j.gaitpost.2014.10.011

54 Scalzitti, D. A., Harwood, K. J., Maring, J. R., Leach, S. J., Ruckert, E. A., & Costello, E. Validation of the 2-minute walk test with the 6-minute walk test and other functional measures in persons with multiple sclerosis. International Journal of MS Care. 2018; 20(4), 158–163. https://doi.org/10.7224/1537-2073.2017-046

55 Sharrack, B., Hughes, R. A. C., Soudain, S., & Dunn, G. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain. 1999; 122(1), 141–159. https://doi.org/10.1093/brain/122.1.141

56 Zhang, Y., Taylor, B. V., Simpson, S., Blizzard, L., Palmer, A. J., & van der Mei, I. Validation of 0–10 MS symptom scores in the Australian multiple sclerosis longitudinal study. Multiple Sclerosis and Related Disorders. 2020). https://doi.org/10.1016/j.msard.2019.101895

57 Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. Stat Med. 2002;21:3431–3446. https://doi.org/10.1002/sim.1253

58 de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide. Cambridge, England: Cambridge University Press; 2011. https://doi.org/10.1017/CBO9780511996214

Appendix 1- Search strategy

Search strategy Pubmed

1) the construct: (gait[Title/Abstract]) OR (walking[Title/Abstract])
2) the population(s): multiple sclerosis[Title/Abstract]
3) the type of instrument(s): ((assessment[Title/Abstract]) OR (test[Title/Abstract])) OR (scale[Title/Abstract])
4) measurement properties of interest: (((reliability[Title/Abstract]) OR (validity[Title/Abstract])) OR (measurement error[Title/Abstract]))
5) #1 AND #2 AND #3 AND #4
6) FILTER: Language (English)
7) (“multiple sclerosis”) AND (“gait" OR “walking”) AND (“assessment” OR “test” OR “scale”) AND (“reliability” OR “validity” OR “measurement error”)

Results: 133 documents

Search strategy Scopus

1) the construct: TITLE-ABS-KEY (gait OR walking)
2) the population(s): TITLE-ABS-KEY ("multiple sclerosis")
3) the type of instrument(s): TITLE-ABS-KEY (assessment OR test OR scale))
4) measurement properties of interest: (TITLE-ABS-KEY (reliability OR validity OR "measurement error")
5) #1 AND #2 AND #3 AND #4
6) FILTER: Language (English)
7) (TITLE-ABS-KEY (assessment OR test OR scale) AND TITLE-ABS-KEY ("multiple sclerosis") AND TITLE-ABS-KEY (assessment OR test OR scale) AND TITLE-ABS-KEY (reliability OR validity OR "measurement error"))

Results: 216 documents

Search strategy Embase

1) the construct: Title, abstract or author-specified keywords (“gait" OR “walking”)
2) the population(s): Title, abstract or author-specified keywords (“multiple sclerosis”)

3) the type of instrument(s): Title, abstract or author-specified keywords (“assessment” OR “test” OR “scale”)
4) measurement properties of interest: Title, abstract or author-specified keywords (“reliability” OR “validity” OR “measurement error”)
5) #1 AND #2 AND #3 AND #4
6) FILTER: Language (English)
7) (Title, abstract or author-specified keywords (“multiple sclerosis”) AND (“gait" OR “walking”) AND (“assessment” OR “test” OR “scale”) AND (“reliability” OR “validity” OR “measurement error”))

Results: 55 documents

Search strategy Web of Science

1) the construct: AB= (gait OR walking)
2) the population(s): AB= (multiple sclerosis)
3) the type of instrument(s): AB= (assessment OR test OR scale)
4) measurement properties of interest: AB= (reliability OR validity OR measurement error)
5) #1 AND #2 AND #3 AND #4
6) FILTER: Language (English)
7) AB= (multiple sclerosis) AND (gait OR walking) AND (assessment OR test OR scale) AND (reliability OR validity OR measurement error)

Results: 149 documents

Appendix 2 - COSMIN risk of bias checklist

Risk of Bias ratings per standard per studies on reliability												
Study, year	**Ader et al (2020)**	**Angelni et al (2020)**	**Behrens et al (2014)**	**Bennett et al (2017)**	**Bourke et al (2020)**	**Callesen et al (2019)**	**Cheng et al (2021)**	**Coulter at al (2017)**	**Craig et al (2017)**	**Decavel et al (2019)**	**Fitzgerald et al (2019)**	**Flachenecker et al (2019)**
Design requirements	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**
Question 1: Were patients stable in the time between the repeated measurements on the construct to be measured?	Not applicable	Very good	Not applicable	Very good	Very good	Inadequate	Very good	Not applicable	Very good	Very good	Adequate	Not applicable
Question 2: Was the time interval between the repeated measurements appropriate?	Doubtful	Very good	Not applicable	Very good	Very good	Doubtful	Very good	Not applicable	Very good	Very good	Doubtful	Doubtful
Question 3: Were the measurement conditions similar for the repeated measurements – except for the condition being evaluated as a source of variation?	Adequate	Very good	Very good	very good	Very good	Very good	Very good	Adequate	Very good	Very good	Adequate	Adequate
Question 4: Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?	Adequate	Adequate	Inadequate	Adequate	Very good	Adequate	Very good	Adequate	Adequate	Adequate	Adequate	Not applicable
Question 5: Did the professional(s) assign scores or determine values without knowledge of the scores or values of other repeated measurement(s) in the same patients?	Not applicable	Adequate	Inadequate	Not applicable	Very good	Adequate	Very good	Adequate	Not applicable	Adequate	Adequate	Not applicable
Question 6: Were there any other important flaws in the design or statistical methods of the study?	Very good	Very good	Very good	very good	Very good	very good	Very good	Very good	very good	Very good	Very good	Doubtful
Question 7: For continuous scores: was an intraclass correlation coefficient (ICC) calculated?	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Adequate	Very good	Very good	Adequate	Adequate
Question 8: For ordinal scores: was a (weighted) kappa calculated?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
Question 9:	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable

For dichotomous/nominal scores: was Kappa calculated for each category against the other categories combined?												
Final Risk of Bias rating Reliability studies	**Doubtful**	**Adequate**	**Inadequate**	**Adequate**	**Very good**	**Inadequate**	**Very good**	**Adequate**	**Adequate**	**Adequate**	**Doubtful**	**Doubtful**
Standards on preferred statistical methods for Measurement error												
Question 7: For continuous scores: was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC), Limits of Agreement (LoA) or Coefficient of Variation (CV) calculated?	Not applicable	Not applicable	Not applicable	Not applicable	Very good	Very good	Very good	Not applicable	Not applicable	Very good	Not applicable	Not applicable
Question 8: For dichotomous/nominal/ordinal scores: Was the percentage specific (e.g. positive and negative) agreement calculated?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
Final Risk of Bias rating study on Measurement error	**Not applicable**	**Not applicable**	**Not applicable**	**Not applicable**	**Very good**	**Inadequate**	**Very good**	**Not applicable**	**Not applicable**	**Adequate**	**Not applicable**	**Not applicable**

Risk of Bias ratings per standard per studies on reliability									
Study, year	**Gholami et al (2017)**	**Goldman et al (2008)**	**Gor-García-Fogeda et al (2020)**	**Grobelny et al (2017)**	**Hobart et al (2003)**	**Learmonth et al (2013)**	**Lord et al (1998)**	**McConvey & Bennett, (2005)**	**Nieuwenhuis et al (2006)**
Design requirements	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**
Question 1: Were patients stable in the time between the repeated measurements on the construct to be measured?	Not applicable	Not applicable	Not applicable	Not applicable	very good	very good	very good	Not applicable	Adequate
Question 2: Was the time interval between the repeated measurements appropriate?	Doubtful	Very good	Doubtful	Doubtful	very good	Doubtful	Inadequate	very good	very good
Question 3: Were the measurement conditions similar for the repeated measurements – except for	doubtful	Very good	Doubtful	Very good	Very good	Doubtful	Very good	very good	Very good

the condition being evaluated as a source of variation?									
Question 4: Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?	Adequate	Very good	Adequate	Adequate	Adequate	Adequate	Very good	Not applicable	adequate
Question 5: Did the professional(s) assign scores or determine values without knowledge of the scores or values of other repeated measurement(s) in the same patients?	Adequate	Very good	Adequate	Not applicable	Adequate	Adequate	Very good	very good	adequate
Question 6: Were there any other important flaws in the design or statistical methods of the study?	very good	Very good	very good	Very good	Very good	very good	Very good	very good	very good
Question 7: For continuous scores: was an intraclass correlation coefficient (ICC) calculated?	very good	Very good	very good	very good	very good	very good	Inadequate	Adequate	adequate
Question 8: For ordinal scores: was a (weighted) kappa calculated?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Very good	Not applicable	Not applicable
Question 9: For dichotomous/nominal scores: was Kappa calculated for each category against the other categories combined?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
Final Risk of Bias rating Reliability studies	**Doubtful**	**Very good**	**Doubtful**	**Doubtful**	**Adequate**	**Doubtful**	**Inadequate**	**Adequate**	**Adequate**
Standards on preferred statistical methods for Measurement error									
Question 7: For continuous scores: was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC), Limits of Agreement (LoA) or Coefficient of Variation (CV) calculated?	Not applicable	Not applicable	Not applicable	Very good	Not applicable	very good	Not applicable	Not applicable	Not applicable
Question 8:	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable

For dichotomous/nominal/ordinal scores: Was the percentage specific (e.g. positive and negative) agreement calculated?									
Final Risk of Bias rating study on Measurement error	**Not applicable**	**Not applicable**	**Not applicable**	**Doubtful**	**Not applicable**	**Doubtful**	**Not applicable**	**Not applicable**	**Not applicable**

Risk of Bias ratings per standard per studies on reliability					
Study, year	**Nilsagard et al (2007)**	**Paltamaa et al (2005)**	**Phan-Ba et al (2011)**	**Rosenblum & Melzer (2017)**	**Sharrack et al (1999)**
Design requirements	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**
Question 1: Were patients stable in the time between the repeated measurements on the construct to be measured?	Adequate	very good	Adequate	Not applicable	Adequate
Question 2: Was the time interval between the repeated measurements appropriate?	Adequate	very good	Very good	Very good	Adequate
Question 3: Were the measurement conditions similar for the repeated measurements – except for the condition being evaluated as a source of variation?	Very good	Very good	Very good	Very good	Adequate
Question 4: Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?	Adequate	Adequate	Adequate	Adequate	Adequate
Question 5: Did the professional(s) assign scores or determine values without knowledge of the scores or values of other repeated measurement(s) in the same patients?	Adequate	Adequate	Adequate	Adequate	Adequate
Question 6: Were there any other important flaws in the design or statistical methods of the study?	Very good	Very good	very good	Very good	Very good
Question 7: For continuous scores: was an intraclass correlation coefficient (ICC) calculated?	very good	very good	very good	very good	adequate
Question 8: For ordinal scores: was a (weighted) kappa calculated?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable

Question 9: For dichotomous/nominal scores: was Kappa calculated for each category against the other categories combined?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
Final Risk of Bias rating Reliability studies	**Adequate**	**Adequate**	**Adequate**	**Adequate**	**Adequate**
Standards on preferred statistical methods for Measurement error					
Question 7: For continuous scores: was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC), Limits of Agreement (LoA) or Coefficient of Variation (CV) calculated?	Not applicable	Very good	Very good	Very good	Not applicable
Question 8: For dichotomous/nominal/ordinal scores: Was the percentage specific (e.g. positive and negative) agreement calculated?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
Final Risk of Bias rating study on Measurement error	**Not applicable**	**Adequate**	**Doubtful**	**Adequate**	**Not applicable**

Study, year	**Behrens et al (2014)**	**Filipovic´ Grc¯ic´ et al (2013)**	**Fitzgerald et al (2019)**	**Forsberg et al (2013)**	**Forsberg et al (2017)**	**Gor-García-Fogeda et al (2020)**	**Grobelny et al (2017)**	**Kalron et al (2016)**	**Kalron et al (2017)**	**McGuigan & Hutchinson (2004)**
Design requirements	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**
Question 1: Is it clear what the comparator instrument(s) measure(s)?	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good
Question 2: Were the measurement properties of the comparator instrument(s) sufficient?	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good
Question 3: Were design and statistical methods adequate for the hypotheses to be tested?	Adequate	Very good	Adequate	Adequate	Very good	Adequate	Adequate	Very good	Very good	Very good
Question 5: Was an adequate description provided of important characteristics of the subgroups?	Not applicable	Not applicable	Not applicable	Not applicable	Very good	Not applicable	Not applicable	Doubtful	doubtful	Very good
Question 6: Were design and statistical methods adequate for the hypotheses to be tested?	Not applicable	Not applicable	Not applicable	Not applicable	Very good	Not applicable	Not applicable	Adequate	Adequate	Very good

Final Risk of Bias rating construct validity studies	Adequate	Very good	Adequate	Adequate	Very good	Adequate	Adequate	Doubtful	Doubtful	Very good

Risk of Bias ratings per standard per studies on construct validity								
Study, year	**Molt et al (2010)**	**Molt et al (2008)**	**Nieuwenhuis et al (2006)**	**Phan-Ba et al (2011)**	**Sandroff et al (2015)**	**Scalzatti et al (2018)**	**Sharrack et al (1999)**	**Vaney et al (2004)**
Design requirements	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**
Question 1: Is it clear what the comparator instrument(s) measure(s)?	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good
Question 2: Were the measurement properties of the comparator instrument(s) sufficient?	Very good	Very good	Very good	Very good	Very good	Very good	Adequate	Very good
Question 3: Were design and statistical methods adequate for the hypotheses to be tested?	Very good	Adequate	Adequate	Very good	Adequate	Adequate	Adequate	Adequate
Question 5: Was an adequate description provided of important characteristics of the subgroups?	Not applicable	Adequate	Not applicable	Very good	Doubtful	Adequate	Not applicable	Not applicable
Question 6: Were design and statistical methods adequate for the hypotheses to be tested?	Not applicable	Adequate	Not applicable	Very good	Adequate	Adequate	Not applicable	Not applicable
Final Risk of Bias rating construct validity studies	**Very good**	**Adequate**	**Adequate**	**Very good**	**Doubtful**	**Adequate**	**Adequate**	**Adequate**

Study, year	Bennet et al (2017)	Cheng et al (2021)	Coulter et al (2017)	Fitzgerald et al (2019)	Forsberg et al (2017)	Fritz et al (2016)	Lord et al (1998)	Molt et al 2008	Phan-Ba et al (2011)	Rosenblum et al (2017)	Zhang et al (2020)
Design requirements	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**	**Consensus**
Question 1: For continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Very good	Very good	Inadequate	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good
Question 2:	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable

For dichotomous scores: Were sensitivity and specificity determined?											
Question 3: Were there any other important flaws in the design or statistical methods of the study?	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good	Very good
Final Risk of Bias rating criterion validity studies	**Very good**	**Very good**	**Inadequate**	**Very good**	**Very good**	**Very good**	**Very good**	**Very good**	Very good	**Very good**	**Very good**

Risk of Bias ratings per standard per studies on internal consistency			
Study, year	**Hobart et al (2003)**	**Fitzgerald et al (2019)**	**Molt et al (2008)**
Design requirements	**Consensus**	**Consensus**	**Consensus**
Question 1: Was an internal consistency statistic calculated for each unidimensional scale or subscale separately?	Very good	Very good	Very good
Question 2: For continuous scores: Was Cronbach's alpha or omega calculated?	Very good	Very good	Very good
Question 3: For dichotomous scores: Was Cronbach's alpha or KR- 20 calculated?	Not applicable	Not applicable	Not applicable
Question 4: For IRT-based scores: Was standard error of the theta (SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	Not applicable	Not applicable	Not applicable
Final Risk of Bias rating construct validity studies	**Very good**	**Very good**	**Very good**

Appendix 3. Summary of findings GRADE

Appendix 3. Summary of Findings GRADE Reliability					
Measurement instrument	**Risk of bias**	**Inconsistency**	**Imprecision**	**Indirectness**	**Quality of evidence**
T25FW	*	*	*	*	High
10MTW	*	*	-1 Moderate	*	Moderate
30MTW	-1 Serious	*	-2 Low	*	Low
100MTW	-2 Very Serious	*	*	*	Low
6MWT	*	*	*	*	High
2MWT	-1 Serious	*	-1 Moderate	*	Moderate
SSST	-2 Very Serious	*	-2 Low	*	Low
DGI	*	*	-1 Moderate	*	Moderate
FGA	No data	No data	No data	No data	No data
TUG	*	*	*	-1 Serious	Moderate
NPWT	-1 Serious	-1 Serious	-2 Low	*	Low
MSWS-12	*	*	*	*	High
WA-VAS	No data	No data	No data	No data	No data
0–10 MS symptom	No data	No data	No data	No data	No data
SymptoMScreen	-2 Very Serious	*	*	-1 Serious	Low
AI	-1 Serious	*	-1 Moderate	-1 Serious	Moderate
RVGA	- 3 Extremely serious	-2 Very Serious	-2 Low	-1 Serious	Very Low
GAIT	-2 Very Serious	-1 Serious	-2 Low	-1 Serious	Low
5UTT	*	*	-1 Moderate	-1 Serious	Moderate
Accelerometer	-1 Serious	*	*	*	Moderate
Kinect Windows sensor	-2 Very Serious	*	-2 Low	*	Low
Wearable, wireless and inertial Sensors	*	*	*	*	High
Samsung Galaxy 7 smartphone + 2MWT	*	*	-1 Moderate	*	Moderate
SMSW + VPC using Laboratory System with Kinect sensor	-2 Very Serious	*	-1 Moderate	-1 Serious	Low
walk ratio using the GAITRite™	No data	No data	No data	No data	No data
*No risk of bias, inconsistency, imprecision and indirectness					

Appendix 3. Summary of Findings GRADE Measurement error					
Measurement instrument	**Risk of bias**	**Inconsistency**	**Imprecision**	**Indirectness**	**Quality of evidence**
T25FW	-2 Very Serious	*	*	*	Low
10MTW	-1 Serious	*	-2 Low	*	Low
30MTW	No data	No data	No data	No data	No data
100MTW	-2 Very Serious	*	*	*	Low
6MWT	-2 Very Serious	*	*	*	Low
2MWT	No data	No data	No data	*	No data
SSST	- 3 Extremely serious	*	-2 Low	*	Very Low
DGI	No data	No data	No data	No data	No data
FGA	No data	No data	No data	No data	No data
TUG	-1 Serious	*	-1 Moderate	-1 Serious	Moderate
NPWT	-1 Serious	-1 Serious	-2 Low	*	Low
MSWS-12	-2 Very Serious	*	-1 Moderate	*	Low
WA-VAS	No data	No data	No data	No data	No data
0–10 MS symptom	No data	No data	No data	No data	No data
SymptoMScreen	No data	No data	No data	No data	No data
AI	-1 Serious	*	-1 Moderate	-1 Serious	Moderate
RVGA	No data	No data	No data	No data	No data
GAIT	No data	No data	No data	No data	No data
5UTT	*	*	-1 Moderate	-1 Serious	Moderate
Accelerometer	-2 Very Serious	*	-1 Moderate	*	Low
Kinect Windows sensor	No data	No data	No data	No data	No data
Wearable, wireless and inertial Sensors	No data	No data	No data	No data	No data
Samsung Galaxy 7 smartphone + 2MWT	*	*	-1 Moderate	*	Moderate
SMSW + VPC using Laboratory System with Kinect sensor	-2 Very Serious	-2 Very Serious	-1 Moderate	-1 Serious	Low
walk ratio using the GAITRite™	-2 Very Serious	-2 Very Serious	-1 Moderate	-2 Very Serious	-2 Very Serious
*No risk of bias, inconsistency, imprecision and indirectness					

Appendix 3. Summary of Findings GRADE Validity					
Measurement instrument	**Risk of bias**	**Inconsistency**	**Imprecision**	**Indirectness**	**Quality of evidence**
T25FW	-2 Very Serious	*	*	*	Low
10MTW	-1 Serious	*	-2 Low	*	Low
30MTW	No data	No data	No data	No data	No data
100MTW	-2 Very Serious	*	*	*	Low
6MWT	-2 Very Serious	*	*	*	Low
2MWT	*	*	-1 Moderate	*	Moderate
SSST	- 3 Extremely serious	*	-2 Low	*	Very Low
DGI	*	*	*	*	High
FGA	*	-1 Serious	-1 Moderate	-1 Serious	Moderate
TUG	-1 Serious	*	-1 Moderate	-1 Serious	Moderate
NPWT	-1 Serious	-1 Serious	-2 Low	*	Low
MSWS-12	*	*	-1 Moderate	*	Moderate
WA-VAS	*	-1 Serious	-1 Moderate	-1 Serious	Moderate
0–10 MS symptom	*	-1 Serious	*	*	Moderate
SymptoMScreen	*	-1 Serious	*	-1 Serious	Moderate
AI	-1 Serious	*	-1 Moderate	-1 Serious	Moderate
RVGA	*	-1 Serious	-2 Low	-1 Serious	Low
GAIT	-1 Serious	-1 Serious	-2 Low	-1 Serious	Low
5UTT	*	*	-1 Moderate	-1 Serious	Moderate
Accelerometer	-2 Very Serious	*	-1 Moderate	*	Low
Kinect Windows sensor	No data	No data	No data	No data	No data
Wearable, wireless and inertial Sensors	No data	No data	No data	No data	No data
Samsung Galaxy 7 smartphone + 2MWT	No data	No data	No data	No data	No data
SMSW + VPC using Laboratory System with Kinect sensor	*	-2 Very Serious	-1 Moderate	-1 Serious	Very Low
walk ratio using the GAITRite	-2 Very Serious	-2 Very Serious	*	-2 Very Serious	Low
*No risk of bias, inconsistency, imprecision and indirectness					

3. Conclusões e considerações finais

Esta tese de mestrado pretende ser um contributo para a construção de uma comunidade que promova a profissão da fisioterapia para melhorar a saúde da sociedade. Isto porque é importante que o fisioterapeuta baseie a sua prática diária na melhor evidência disponível. Para isso, tem de procurar conhecimento e adaptar os seus paradigmas de intervenção às mudanças constantes, focando-se na literatura disponível.

Este trabalho visa contribuir para uma melhoria do desempenho dos fisioterapeutas na avaliação dos défices de marcha em indivíduos com diagnóstico de Esclerose Múltipla. Por isso foi definido a elaboração de um trabalho com o objetivo de identificar quais são os instrumentos de medida, disponíveis na literatura, para avaliar este construto nesta população específica. E igualmente fornecer uma visão geral de todos as ferramentas utilizadas e da qualidade das mesmas na medição de construto nesta população específica.

Com base nos resultados foram encontrados 25 instrumentos de avaliação que se dividem em cinco categorias (*PerFOMs, PROMs, ClinROMs, ObsROMs* e resultados de biomarcador (*biomarker outcomes)*).

Os instrumentos *PerFOMs* são ferramentas que avaliam o desempenho da marcha, através da aplicação de testes que são pontuados por um examinador. Os *PROMs* são instrumentos subjetivos, pois a avaliação das dificuldades da marcha é feita pelo indivíduo. Já os *ClinROMs* são avaliações realizadas apenas por médicos, principalmente neurologistas. *ObsROMs* são instrumentos de observação e dependem da observação de um examinador. Por fim, os resultados dos biomarcadores pretendem ser instrumentos mais próximos da avaliação real da marcha, pois são utilizados quando o indivíduo realiza a sua vida diária. Assim, todas estas medidas de avaliação de resultados permitem quantificar a marcha de diferentes formas e contextos.

O grande número de instrumentos disponíveis na literatura permite corroborar a importância da avaliação deste construto nesta população, visto que é considerada uma das principais causas de incapacidade da EM.

Durante o artigo foram utilizadas as diretrizes da *COSMIN*, que é uma iniciativa de uma equipa internacional e multidisciplinar com experiência no desenvolvimento e avaliação de ferramentas de medição. A checklist *COSMIN* permitiu avaliar a qualidade metodológica dos artigos incluídos. Sendo que é a ferramenta de avaliação de qualidade mais amplamente utilizada nos estudos sobre propriedades de medição. Além disso, considerando os instrumentos encontrados na pesquisa, foi o instrumento ideal para

avaliar as propriedades dos estudos. Por isso, a utilização da *COSMIN* permitiu enriquecer o manuscrito e melhorar a seleção dos instrumentos mais adequados.

Ao avaliar a evidência, através da abordagem *GRADE*, concluímos que *T25FW, 6MWT, MSWS-12* e os sensores vestíveis e sem fio têm um forte nível de evidência para a confiabilidade.

Analisando essas ferramentas, verificamos que o T25FW e o 6MWT têm a vantagem de fornecer prontamente uma medida quantitativa do desempenho da caminhada. O primeiro considera a velocidade de caminhada e o segundo avalia a distância percorrida. Além disso, esses instrumentos são facilmente aplicáveis na prática clínica porque são baratos e podem ser usados em uma ampla gama de pacientes.

Por outro lado, o MSWS-12 foi desenhado como um instrumento específico para avaliar o impacto da EM na capacidade de locomoção. Assim, é um instrumento que representa a perspectiva do indivíduo quanto à sua capacidade de marcha, e os seus resultados podem ajudar a determinar intervenções mais adequadas às suas necessidades. Também é fácil de utilizar, barato e pode ser aplicado na prática clínica ou em casa.

Além dessas medidas, é interessante analisar o desempenho da marcha por um período prolongado e no ambiente do indivíduo. Os avanços tecnológicos permitem preencher essa lacuna, pois é possível avaliar a marcha no contexto real do usuário, através dos sensores vestíveis, sem fio e inerciais.

Os resultados de todos os instrumentos avaliados, incluindo os citados em cima, não são tão favoráveis para o erro de medição e validade. Isso porque, para a primeira propriedade psicométrica nenhum instrumento foi classificado com evidência alta e para a segunda apenas o DGI apresentou forte evidência.

Nesta revisão, concluímos que existem inúmeros instrumentos que permitem avaliar a marcha em diferentes formas e contextos. É necessário conhecer quais são essas medidas para selecionar as que melhor se adequam ao contexto e objetivo pretendido. Neste trabalho verificámos que os sensores inerciais, o T25FW, a 6MWT e MSWS-12 são medidas promissoras na avaliação deste construto nestes indivíduos, visto que apresentam altos resultados quanto à confiabilidade. No entanto, mais estudos sobre o erro de medição e a validade desses testes são necessários para recomendá-los totalmente como ferramentas de medição da marcha na EM.

A preparação desta revisão sistemática foi um processo de aprendizagem contínua. Mas, acima de tudo, foi um desafio que exigiu capacidade de perseverança, empenho resiliência e determinação para contribuir com evidência e conhecimento relevantes para a prática da fisioterapia. Mas, acima de tudo, gostaríamos que este trabalho contribua para o crescimento da profissão e a sua relevância na sociedade e na vida das pessoas, particularmente das que têm diagnóstico de Esclerose Múltipla.