Outcome measures for Multiple Sclerosis: A review

Lisa A. Osborne¹, J. Gareth Noble¹, Inocencio D.C. Maramba¹, Kerina H. Jones¹, Rodden M. Middleton¹, Ronan A. Lyons¹, David V. Ford¹, & Phil Reed²

¹ College of Medicine, Swansea University, U.K.
² Department of Psychology, Swansea University, U.K.

Short Title: Outcome measures for MS.

Abstract

**Background:** This review determined the most commonly used, and reliable, measures for assessing clinical outcomes for Multiple Sclerosis (MS).

**Objectives:** It was anticipated that this would facilitate the development of a common set of metrics, and aid reaching a consensus regarding the outcome measures that are typically used in the field of MS clinical research.

**Major Findings:** A thorough literature review of clinical outcome measures for MS produced 166 measures that have been used in this context. This list was then refined by discussion with a panel of consultant neurologists, which reduced the list to 23 commonly employed tools. This shortlist was then further refined through surveying 41 centres for MS treatment, which reduced the shortlist to 16 measures. The properties of these scales, in terms of their symptom/function domains, their specificity for MS, their administration characteristics, and their reliability and validity for MS, are all discussed.

**Conclusions:** Conclusions regarding the development of potential sets of assessment measures for MS, that encompasses broad symptom/function domains, and which are sensitive to the practical requirements of administration within clinical contexts, are explored.

**Key Words:** Assessment Sets; Multiple Sclerosis; Narrative Review; Outcome Measures; Psychometric Properties.
Introduction

Multiple Sclerosis (MS) is a chronic condition of the central nervous system\textsuperscript{1,2}. Although a central database for MS in the U.K. is being piloted\textsuperscript{3-5}, there are currently no definitive figures regarding those it affects. Estimates suggest that anywhere between 70,000 and 100,000 people in the U.K. have MS\textsuperscript{6-8}. MS is most often diagnosed between the ages of 20 and 40 years, although it can be diagnosed earlier or later in life\textsuperscript{3,6}, and it appears to be one of the most common disabling neurological diseases among young adults\textsuperscript{9}. MS is a complex condition that presents a wide spectrum of severities, symptoms, and impacts on functioning; the neurological damage impacts on physical\textsuperscript{10-12}, cognitive\textsuperscript{13-15}, and psychological and emotional\textsuperscript{16-18} functioning, as well as on quality of life\textsuperscript{19-21}. In order to enhance medical practice in diagnosis, care, and treatment, it is important to determine which measures are best employed in the assessment of MS symptoms/functions and clinical outcomes\textsuperscript{22,23}. This topic has recently been the subject of some study and debate in the context of clinical outcomes\textsuperscript{22,23}, and any databases or and registries aiming to further knowledge about MS clearly require strong measures at their cores\textsuperscript{23}. In fact, recently, some such databases have been criticised on the basis of the measures selected\textsuperscript{24}. The questions, raised by numerous task forces concerned with MS over the last two decades about the best set of measures for clinical outcomes for MS that can detect changes in impairment and evaluate a person’s MS symptoms and functioning, remain largely unanswered\textsuperscript{22,25}.

There are numerous measures available for the study of MS\textsuperscript{26}, and a number of reviews of these assessment instruments have previously been conducted\textsuperscript{20,22,27-34}. These reviews highlight several commonly used instruments, as well as outlining their properties. However, these reviews are not
always comprehensive in their coverage of the literature, frequently focusing on only a few selected measures. They do not always address the full range of MS symptomatology or all aspects of a person’s functioning, focussing instead on only one specific aspect of MS\textsuperscript{28,32,34,35}. Previous work does not always attempt to define standard and comprehensive packages of measures to encompass all aspects of MS (i.e. physical, cognitive, psychological and emotional, and quality of life functions), although this is increasingly recognised as an important step\textsuperscript{36,37}. Moreover, with a few exceptions\textsuperscript{38}, often previous review articles do not account for typical clinical-care usage, but rather are inclined to be research-oriented in outlook\textsuperscript{12,32}.

The review that is reported here aims to provide an up-to-date, consensus-based overview of the most extensively utilised clinical outcome measures for MS that could detect changes in levels of impairment, compatible with the practicalities and demands of clinical care, detailing their relative strengths, and their focus on particular symptom/function domains. In doing so, this review addresses current practice and views concerning the most effective instruments currently employed for assessing MS symptoms/functionality and treatments. This review has multiple objectives: (1) to survey and critique the measures in common use; (2) to examine the strengths and weaknesses of these measures in the context of MS data collection (e.g., their reliability and validity for this population); and (3) in the light of these findings to offer a helpful resource for clinicians and researchers working in the field of MS to enable them to choose the set of measures that are best suited for their particular needs. The study’s outputs will aim to be a helpful resource for clinicians and researchers working in the field of MS.
**Review Method**

**Search Strategy**

Search engines and computerised bibliographic databases were deployed to identify papers that have used measures of clinical outcomes for MS (PubMed, PsychINFO, Google Scholar, Web of Knowledge). In all cases, the search phrases, “clinical-outcome-measures” and “multiple-sclerosis”, were employed together, using a backward chronological saturation search strategy. Such a search strategy moves from the most recent year (2012) backwards, in chronological order, through the published literature, year by year. For every article identified, its methodology was examined to determine which measures had been used. This process continued until no further measures were identified from the search. The ‘saturation principle’ was used to determine when to stop (i.e. when no new measures were coming to light).

This method was designed to produce a list of outcome measures that reflects current and recent usage. This is important, given the fast-moving nature of medical and health care developments as a whole, and in this field, in particular, with many instruments rapidly superseded and replaced by alternative measures, or are altered and updated with newer versions.

**Validation**

To ensure that there were no major omissions from the initial list of outcome measures, two additional search and validation strategies were adopted. Firstly, the full list of instruments, derived from the current search, was compared to those listed in a number of recently published reviews of the literature on MS outcome measures\(^{20,22,27-30,32-34}\). Secondly, the full set of measures derived was compared to the list of measures provided by the NICE Guidelines\(^{36}\). This
rigorous search and validation process identified an initial total of 166 instruments that have been employed extensively as MS outcome measures.

**Refinements**

This extensive list was subjected to two stages of refinement, in order to produce a summary of contemporary clinical outcome measures most often used for MS. Firstly, the full list was provided to an independent panel of consultant neurologists with expertise in MS, convened by the U.K. MS Society, for scrutiny and comment. This process involved discussion of these measures, moderated by a member of the MS Society who was independent to the authors of this study. The panel members were asked to comment on the full list, and to indicate which measures they typically employed for the assessment of MS outcomes. This refinement process was directed at developing a much reduced version of this list. This first refinement process resulted in 23 measures being identified.

The 23 measures identified as most used in MS clinical contexts, were sent, as part of a questionnaire, to 82 U.K. centres involved in the treatment of MS. These centres were asked to indicate which of these measures they employed, and whether they used any other measures in addition to those listed. The teams who were consulted at each of the treatment sites were interdisciplinary, and typically comprised a Consultant Neurologist, a Head MS Nurse, a Physiotherapist or a Speech Therapist or an Occupational Therapist. The teams all contributed to the questionnaire responses, and the Head MS Nurse usually returned the questionnaire after consultation with their other team members. Responses were obtained from 41 centres (50%).
These responses indicated that none of the centres used measures that were not already included in the shortened list from the expert panel consultation. All of the 23 measures, except for two (Brief Repeatable Battery of Neuropsychological Tests, and the Percentage of Patients Remaining Relapse Free) were employed by at least one of these centres. There were 21 measures identified as being in contemporary use for MS clinical outcome assessments. The mean use across the centres for any specific scale was 20% of the centres (± 17%), with the use ranging from 2% of the centres employing the least-used measure, to 76% of the centres employing the most-used measure (see Figure 1).

Figure 1

Measurement Scales

The identified measures were subject to further analysis by searching the literature in order to identify their features, and their psychometric properties. The standard types of psychometric properties were identified (see Figure 2) to enable the measures to be compared with one another. In all cases, at least two independent assessments of these properties were identified from the literature, where possible, from sources other than the author. Of the remaining 21 measures, four were removed. Clinical Relapse Rate/Severity/Time to First Relapse was removed, because it could not be considered as a measure with psychometric properties; Quality-Adjusted Life Years, depends on other measures for its usefulness; and a further two, Evoked Response Potential, and Magnetic Resonance Imaging, were generic technologies that encompass a wide range of measures, which are not readily susceptible to psychometric analysis.
Additionally, details regarding the Fatigue Impact Scale, and its modified version, were merged into one for the purposes of this review.

Figure 2

Figure 1 shows the measures categorised by the predominant domains that they address. In addition, each instrument is classified as MS-specific or generic, and as clinician-administered or patient-completed. Figure 3 shows the psychometric properties, as measured on MS samples, and main qualities of each measure.

Figure 3

Physical Disability

*The Modified Ashworth Scale*\(^{39,40}\) measures muscle spasticity for each joint on a 6-point scale. The resulting score has variable inter-rater agreement, some being good (86\(^{39}\)%), but this reliability is better for upper joints (80.6\%) than for lower joints (63.9\%)\(^{41,42}\). As motor function can vary dramatically between extremities within individuals with MS, it makes little sense to evaluate the internal consistency of this scale, although some attempted this, and found moderate values (0.78\(^{37}\)). The test has varying test-retest reliabilities, 93.4\% for upper joints, and 71.1\% for lower joints, with some studies reporting very low test-retest reliability\(^{37}\). However, the variability of these reliabilities across joints has caused some concern\(^{43,44}\). The resulting spasticity score has rather poor concurrent validity with similar disability scales, such as the
Outcome measures for MS

Spasm Frequency Scale\(^4\), and electromyography\(^5\), and a poor correlation with the EDSS\(^6\). Upper joints have better concurrent validity than lower joint scores\(^7\). The variable reliability and poor validity have lead some to question its usage for MS\(^8\), and the modified modified version has been produced, which appears to show better psychometric properties in its initial trials\(^9\).

**Expanded Disability Status Scale (EDSS)**\(^{10}\) is a measure of impairment and impacts of MS, which scores disability on a 20-point scale (0 = normal to 10 = death), on the basis of a neurological examination. In addition to the overall score, there are also 8 associated functional systems assessment scales. There is a patient-rated equivalent, the *Patient Determined Disease Steps*, which is a single-item scale, relating to 8 levels of disability. This is a generic measure of disability, but it is mainly used for MS, and it correlates between 0.70\(^{11}\) and 0.93\(^{12}\) with the clinician-administered EDSS. The EDSS has moderately good reliability characteristics, having inter-rater correlations measured as between 0.32 and 0.76\(^{13}\), and 0.5 and 0.7\(^{14}\). It has an internal reliability of 0.88 to 0.96\(^{15}\), and test-retest reliability reported as between 0.42 and 0.66\(^{16}\), and 0.61 and 0.94\(^{17}\). It also has good concurrent validity, in that it correlates well with the Multiple Sclerosis Functional Composite score, with measures of disability (0.84)\(^{18}\), and with physical, but not mental, measures of quality of life\(^{19-21}\). It correlates well with patient-rated measures of physical disability\(^{22}\). It has poor sensitivity (24%), but good specificity (79%)\(^{23}\). However, a major drawback is its lack of responsiveness to clinical change (effect size = 0.1)\(^{24,25}\), being better for less severely disabled individuals\(^{26}\). In particular, changes in the EDSS score from one time to another often fall within the error that would be typically expected on the basis of inter-rater agreement\(^{27}\). To help overcome these problems it has been suggested that two EDSS determinations are made prior to intervention, and either the mean value (to increase power) or
the lower value (to increase sensitivity) be employed\textsuperscript{61}. It is generally agreed that it would need substantial modification to be used as a main tool of clinical change\textsuperscript{62}, but its widespread use and long history mean that some have suggested that it continue to be used as part of a battery of assessment tools\textsuperscript{22}.

**Functional Independence Measure (FIM)**\textsuperscript{63} assesses the ability of a patient to conduct routine daily activities, also often used for older people and those who have suffered a stroke. The instrument contains 18 items, which produce a composite measure of independence when assessed: 13 of these items are concerned with motor function (based on the Barthal Index); and 7 items are related to cognitive function. A shorter version, the alpha-FIM, is available, which has only 6 items\textsuperscript{64}. All items are rated on a 1 (total assistance required) to 7 (full independence) scale. It can be combined with the Functional Assessment Measure (FAM), which adds 12 items concerned with cognition, similarly completed by a health care professional\textsuperscript{65}. The FIM has an inter-rater reliability of between 0.83 and 0.99\textsuperscript{66-68}, although this is greater for the motor items (0.95) than for the cognition items (0.78)\textsuperscript{67}, suggesting that the FIM should not be used without the FAM for assessment of cognitive independence\textsuperscript{69}. The FIM also has good internal reliability (0.94\textsuperscript{66}, 0.98\textsuperscript{12}). It has good test-retest reliability, measured as between 0.95\textsuperscript{67} and 0.99\textsuperscript{12}. It has good concurrent validity with both the EDSS\textsuperscript{66}, and the Barthal Index\textsuperscript{70}. As the FIM was heavily based on the Barthal Index for motor items, this is not surprising. However, the validity is less good for the neurological aspects that it measures. There are no data for its sensitivity and specificity. A key criticism is that it is not responsive to change\textsuperscript{12}, with effect sizes of between 0.3\textsuperscript{69} and 0.46\textsuperscript{12}, and lower for the cognitive items. This lack of responsiveness to change may be the result of floor and ceiling effects\textsuperscript{71}.  

Guy's Neurological Disability Scale (GNDS)\textsuperscript{72} is a measure of general clinical disability, and asks the patient to assess their current clinical state over the last month (a patient self-administered version is also available\textsuperscript{73}). It gives an overall measure of disability and sub-scores covering 12 domains. It has an internal reliability of around 0.80\textsuperscript{72,74}. This instrument has a very good test-retest reliability of over 0.96\textsuperscript{72,75} for the overall score, and of 0.68 to 0.99 for the sub-scales. However, it has mixed results in terms of concurrent validity; correlating 0.64 with the EDSS, and 0.76 with the Barthal Index\textsuperscript{75}. As would be predicted, it has lower correlations with scales that do not measure physical disability\textsuperscript{76}. There are no data on its sensitivity or specificity. The GNDS appears to be reasonably responsive (effect size, 0.58\textsuperscript{12}, although lower effect sizes have also been noted\textsuperscript{77}. Its duration of administration, and its uncertain responsiveness, means it tends to be used largely in clinical practice, rather than for trial purposes\textsuperscript{27}.

Multiple Sclerosis Functional Composite (MSFC)\textsuperscript{78} assesses the level of impairment across three domains at the time of testing (i.e. leg function is assessed by a 25-feet walking task, arm function by the nine-hole peg test, and cognitive function by use of the paced auditory serial addition test). The assessment provides analysis of the above three aspects of functioning, as well as an overall composite measure (although this summary measure has been critics as being ‘abstract’ and difficult to interpret clinically\textsuperscript{22}. The measure has an inter-rater reliability of 0.95\textsuperscript{79} to 0.96\textsuperscript{80} for the composite score, and 0.93 to 0.99 for the components\textsuperscript{80}. It has an internal reliability of 0.97 for the overall composite score\textsuperscript{81}. The test-retest measures are good, ranging between 0.87\textsuperscript{56} and 0.90\textsuperscript{82} for the overall score, and 0.93 to 0.98 for the subtests\textsuperscript{80}. However, there
Outcome measures for MS are some practice effects that are particularly noticeable on the first few applications. It has good concurrent validity scores with a range of other measures of disability: 0.80 with the EDSS; and it correlates well with MRI measures of brain atrophy. It shows weaker association with measures of emotional functioning. It has moderate sensitivity (51%) and it is moderately good at predicting both disability and MRI results. It has a strong (86%) specificity. Its responsiveness is measured as being between 0.62 and 0.71 in terms of the area under the curve in the ROC, and has a reported effect size of 0.50, although responsiveness to clinical trials can be compromised by the absence of a test for vision, and this value is better for the leg function than for the upper limb and cognitive measures. Moreover, the responsiveness of the cognitive component is not strong, and this should not be relied upon as a sole measure of cognitive change for MS. As a result many have suggested replacing the existing cognitive component (the paced auditory serial addition task), with the symbol digit modalities test, to enhance the test.

**Multiple Sclerosis Walking Scale-12 (MSWS-12)** assesses patient-reported disruption to their walking, and its 12 items measure patient-perception of their walking quality during the past 2 weeks. This measure has a strong internal reliability of 0.97, and good test-retest reliability ranging from 0.86 to 0.96, and over 0.78 for the individual items. It has good concurrent validity, as it correlates 0.80 overall with the EDSS (although this relationship is mainly driven by strong correlations at lower levels of disability), and 0.77 with the MSIS-29 physical scale, and 0.82 with physical scales from the SF-36. However, its correlation with accelerometry scores are moderate, and range from 0.38 to 0.70, as are its correlations with other objective measures of walking. There are no data on its sensitivity and specificity. A strong feature of
the tool is that it has good levels of responsiveness\textsuperscript{35}, with effect sizes being greater than 1.00\textsuperscript{77,95}, but the variance between differently disabled groups may make it less than optimal for comparing between samples\textsuperscript{96}.

**Cognitive Impairment**

*Mini-Mental State Examination (MMSE)*\textsuperscript{97} screens for dementia and cognitive impairment at the time of testing. It comprises 11 items that measure a number of domains (a standardised version exists, but this has not been tested for MS\textsuperscript{98}). It has a good internal reliability of 0.89 to 0.95\textsuperscript{97}, but variable test-retest reliability, ranging between 0.65 and 0.89\textsuperscript{97}. It does display some moderate concurrent validity – it correlates 0.78 with the verbal IQ score from the WAIS\textsuperscript{97}. However, it has been noted that it is not well suited for the patterns of disability seen in MS\textsuperscript{99}, and it has poor sensitivity for this population of between 28\%\textsuperscript{100} and 36\%\textsuperscript{101}. It does have good specificity (89\%)\textsuperscript{101}, but poor, and unreliable, responsiveness, given its weak test-retest reliability\textsuperscript{102}. Given the latter problems, some have recommended that it not be used for people with MS\textsuperscript{103}, and it is more commonly employed as part of a battery of tests to describe the sample characteristics, rather than to assess change due to an intervention\textsuperscript{104}.

**Psychological/Emotional**

*Beck's Depression Inventory (BDI-II)*\textsuperscript{105,106} measures depression over the last week. There are 21 items, each rated on a 4-point scale, relating to symptoms of depression, which gives an overall score. It has an internal reliability score for people with MS of between 0.86\textsuperscript{107} and 0.94\textsuperscript{108}. It has a test-retest reliability of 0.93 across people with a range of disorders\textsuperscript{105}. It has good concurrent validity for MS\textsuperscript{109}, and correlates well (0.65) with psychiatric ratings of
depression, and with the SF-20 sub-scales\textsuperscript{108}, and 0.71 with the Hamilton Depression Scale\textsuperscript{105}. It has between 81\%\textsuperscript{110} and 84\%\textsuperscript{111} sensitivity, and 70\% specificity, in all phases of the disease\textsuperscript{111}. Its responsiveness is also reasonable\textsuperscript{109,112}. However, there is some discussion about whether some of the items from the BDI-II should be removed for use with MS, as they overlap with some of the symptoms and impacts of MS\textsuperscript{107,113}.

**General Health Questionnaire (GHQ)**\textsuperscript{114} is used for assessing general psychiatric problems over the prior ‘several weeks’, and it comes in many versions, including: the GHQ-12 (which gives an overall score); the GHQ-28 (which gives scores for somatic problems, anxiety, social dysfunction, and severe depression, as well as an overall score); the GHQ-30 (which gives an overall score); and the GHQ-60 (which gives an overall score). The most used for MS are the GHQ-12 and GHQ-28, being the shortest versions. The GHQ-12 has an internal reliability of 0.91\textsuperscript{115}, and the GHQ-28 has an internal reliability of 0.90\textsuperscript{116}. The GHQ-28 has a test-retest reliability of 0.69 in a general population\textsuperscript{116}. There are reports of a retest/practice effect impacting on the score\textsuperscript{117}. The GHQ-28 has a moderate concurrent validity (0.44 to 0.66) with a range of health measures\textsuperscript{76}, and correlates 0.83 with the present state examination\textsuperscript{118}. There are mixed reports about the sensitivity of the GHQ-12, these being 36\% to 67\%\textsuperscript{76} or 72\% to 92\%\textsuperscript{119}, the GHQ-28 being reported to have better sensitivity\textsuperscript{120}, and is recommended as one of the stem tools for identification of depression in chronic illness\textsuperscript{121}. The GHQ-12 has good specificity, being measured as over 74\%\textsuperscript{76,119,120}. Reports on the responsiveness of the GHQ-12 are also mixed, varying from effects sizes of 0.15\textsuperscript{122} to 0.51\textsuperscript{123}. 
Hospital Anxiety and Depression Scales (HADS)\textsuperscript{124} assess levels of anxiety and depression, excluding somatic symptoms to avoid overlap with any physical symptoms of the patient. Originally designed for use by hospital general medical outpatients, it has since been widely employed in primary care settings, and also in on-line surveys\textsuperscript{125}, and contains 14 items (7 for anxiety and 7 for depression), that relate to the last week. Patients are recommended not to take very long to respond, as their immediate replies should better reflect their actual state than more considered reactions\textsuperscript{124}. This instrument has excellent reliability, having an internal reliability of 0.83 for anxiety\textsuperscript{21,126}, and of 0.81\textsuperscript{126} to 0.82\textsuperscript{21} for depression. Its test-retest reliability is 0.89 for anxiety and 0.86 for depression\textsuperscript{127}. It has reasonable concurrent validity correlates between 0.49 and 0.83 with similar instruments\textsuperscript{128}, and 0.62 for anxiety, and 0.55 for depression, with the appropriate Beck scales\textsuperscript{76}. It has mixed reports concerning sensitivity, it has 88.5\% sensitivity for anxiety, and a 90\% sensitivity for depression, when tested against objectively rated symptoms\textsuperscript{129}, but between 29\% and 46\% for anxiety, and 25\% and 46\% for depression, when measured against other self report scales\textsuperscript{76}. Its specificity is excellent, being measured as between 80\% and 90\% for anxiety, and 84\% to 90\% for depression, regardless of the anchor measure used\textsuperscript{76,129}. However, the anxiety scale tends to be more sensitive to General Anxiety Disorder than to other forms of anxiety\textsuperscript{129}.

Quality of Life

EuroQol-5 Dimensional Questionnaire (EQ-5D)\textsuperscript{130} is a measure of the impacts of disease on various aspects of health. There are 5 specific, Likert-type, questions addressing the patient’s state on the day of assessment (plus optional questions relating to socio-economic variables). It has reasonable test-retest reliability reported as 0.63-0.80\textsuperscript{131}, and 0.81\textsuperscript{132}. It has reasonable
concurrent validity, correlating 0.70 with the SF-60\textsuperscript{132}, and 0.55-0.81 with a range of similar generic quality of life instruments\textsuperscript{12}. However, its relationship with EDSS scores in MS are uncertain\textsuperscript{133}, and it does not correlate with objective measures of health and functioning\textsuperscript{134}. It is not regarded to have sensitivity for MS\textsuperscript{76}, and the absence of a fatigue score is an issue in this regard\textsuperscript{135}, and there are no direct specificity data. Its responsiveness is less than that of other instruments\textsuperscript{12}, with some ceiling effects and with 40% of severely disabled individuals likely to omit the items relating to physical status\textsuperscript{132}. This measure is considered good for population description, but it may be too insensitive for measuring individual changes. Although it is commonly used for cost estimates in health-economics\textsuperscript{136}, some have suggested that disease-specific tools, such as the MSIS-29, are more sensitive in this context\textsuperscript{137}.

**Fatigue Impact Scale and Modified (FIS/MFIS)\textsuperscript{138,139}** measures the impact of fatigue on daily life over the last 4 weeks. This instrument comprises 40 items, and gives an overall fatigue impact score and three sub-scales. (The 21-item MFIS\textsuperscript{139} is a reduced version of the FIS, taking 5 to 10 minutes to complete, giving an overall score and three subscales; an even briefer 5-item version takes 2 to 3 minutes to complete). The FIS has good indicators of overall internal reliability of 0.98\textsuperscript{138}, and 0.85\textsuperscript{140}, with the individual sub-scales all having internal reliabilities of over 0.87. The overall MFIS score has an internal reliability measured from 0.81\textsuperscript{139} to 0.92\textsuperscript{141}, with sub-scales ranging between 0.88 and 0.92\textsuperscript{141}. The test-retest reliabilities of the overall FIS score is measured as between 0.81\textsuperscript{142} and 0.93\textsuperscript{122}, with the sub-scales ranging from 0.68 to 0.85\textsuperscript{138,142}. For the MFIS, the overall test-retest reliability varies between 0.82-0.85\textsuperscript{141,143,144}. There is moderate concurrent validity of the FIS with the SF-36\textsuperscript{142}, and between the MFIS and the Fatigue Severity Scale (0.66)\textsuperscript{144}, 0.68 \textsuperscript{145}. In terms of its sensitivity, 78% of people with MS are identified...
correctly by the FIS\textsuperscript{138}, with the MFIS have a greater then 90% sensitivity\textsuperscript{146}. There are no data on specificity for the FIS, but the MFIS has greater than 90% specificity\textsuperscript{146}. The FIS has a greater than 0.7 effect size in terms of its responsiveness\textsuperscript{147}, with the MFIS having effect sizes of greater than 0.5\textsuperscript{148,149}, but it is not greatly responsive over short periods of time\textsuperscript{144}. Recently, the usefulness of the overall fatigue impact score of these scales has been called into question, although the sub-scales, used on an individual basis, are considered to be safe\textsuperscript{150}, and, as a result, the 22-item Unitary FIS (U-FIS) has been developed to give an overall fatigue score\textsuperscript{151}. Nevertheless, this is the tool recommended to assess fatigue-related quality of life\textsuperscript{152}.

\textit{Leeds Multiple Sclerosis Quality of Life Scale (LMSQoL)}\textsuperscript{153} is a measure of the impact of MS on quality of life at the time of completion. This scale has 8 items, each scored on a 4-point Likert-type scale, and it gives an overall index of quality of life. The scale has a good internal reliability, being reported as between 0.71 and 0.86\textsuperscript{21,76,153}, and good test-retest reliability of 0.85\textsuperscript{153}. While there are some reports of moderate concurrent validity: 0.68 with the SF-36, and 0.83 with measures of well-being\textsuperscript{153}, and it correlates with diary reports of the impact of MS\textsuperscript{154}, recent questions have been raised about its validity when compared to the MSQoL\textsuperscript{76}. There are no data for its sensitivity and specificity. It has moderate responsiveness, having reported effect sizes of 0.34\textsuperscript{76,155}, and 0.45\textsuperscript{156}, with little sign of floor or ceiling effects\textsuperscript{157}.

\textit{Multiple Sclerosis Impact Scale (MSIS-29)}\textsuperscript{158} measures the impact of MS on daily living over the preceding 2 weeks, rather than on the resulting quality of life. Its 29 items produce an overall score and sub-scales relating to both physical and psychological domains. The overall score is considered to be of debatable use\textsuperscript{159}. The internal reliability of the two sub-scales is very good
ranging from 0.88 to 0.96 for the Physical, and 0.85 to 0.91 for the Psychological\textsuperscript{158,160-162}, as are their test-retest reliabilities, which range from 0.86 to 0.94 for the Physical, and from 0.81 to 0.87 for the Psychological\textsuperscript{158,161-163}. The score correlates moderately well with similar instruments, giving this scale reasonable concurrent validity: the Physical scale correlates 0.66 with the EDSS and 0.69 with the GNDS\textsuperscript{162}, and greater than 0.5 with other measures of disability\textsuperscript{160,164}, the Psychological scale correlates greater than 0.6 with other scales measuring mental functioning\textsuperscript{160}. The sensitivity is good\textsuperscript{137}: 78\% for the Physical scale\textsuperscript{161,165}, and 73\% for the Psychological scale\textsuperscript{161}. Reports on its specificity are mixed, some results being good (greater than 80\% for both scales\textsuperscript{161}, but some being only moderate (51\% for the Physical scale\textsuperscript{165}. The responsiveness of the Physical scale, which is good (0.82 effect size), is better than that of the Psychological scale, which is moderate (0.66 effect size)\textsuperscript{122}. The area under the curve in its ROC is good when anchored against the EDSS, 0.72\textsuperscript{165}, but lower values have been reported\textsuperscript{166}.

\textit{Multiple Sclerosis Quality of Life-54 (MSQoL-54)}\textsuperscript{157} assesses the quality of life of individuals with MS over the last 4 weeks (the 54 items include the 36 items from the generic SF-36 questionnaire). This tool gives an overall score for quality of life, comprising two health subdomains (Physical and Mental). The overall score has an internal reliability measured as between 0.84\textsuperscript{167} and 0.96\textsuperscript{168}, with the sub-scales having internal reliabilities ranging from 0.75 to 0.96\textsuperscript{157}. The test-retest reliability coefficients range between moderate 0.61 and good 0.96\textsuperscript{157,169}. The overall measure correlates moderately with MS symptoms measured by the ICD\textsuperscript{167}, it also correlates moderately with the EDSS and MSFC (0.49 to 0.67), and with the Leeds MS QoL scale\textsuperscript{170}. Both the Physical and Mental Health sub-scales correlate well with the Fatigue Severity Scale\textsuperscript{171}. There are no data on its sensitivity and specificity, although the addition of 18 MS-
specific questions to the generic SF-36 should improve these properties relative to the generic form\textsuperscript{157}. Reports on its responsiveness are mixed\textsuperscript{172} with effect sizes of over 0.7 for the physical scales, and between 0.57 and 0.7 for the mental scales\textsuperscript{166}, while other reports on responsiveness have been poor\textsuperscript{123}. There has been some criticism of potential floor and ceiling issues that may limit its usefulness as an outcome measure\textsuperscript{28}.

\textbf{Short Form SF-36 Health Scale}\textsuperscript{173} measures the impact of a disorder on the functioning of an individual over the last 4 weeks. Derived from the GHQ-28\textsuperscript{114}, this form contains 36 items, producing an overall score relating to impacts on health, as well as two sub-scales (physical health and mental health). The overall scale has an internal reliability of 0.67 to 0.94\textsuperscript{173}, with the sub-scales ranging between 0.77 and 0.96\textsuperscript{54,174}. The test-retest score results are mixed, and are between 0.64 and 0.96 for the subscales\textsuperscript{174}. However, there have been recent suggestions that the psychometric properties of the two main scales (physical and mental) are not as good as the overall score, or those of the sub-scales on which they are based\textsuperscript{158}. It has moderate concurrent validity (0.6) with symptom severity\textsuperscript{174}. The physical scale has moderate correlations (0.6) with the FIM and the EDSS, and the mental sub-scale correlates 0.5 with the GHQ\textsuperscript{123,133}. It has moderate sensitivity for body functions and activities in MS\textsuperscript{175}, but there are no data on its specificity. The effect sizes show poor responsiveness, ranging from 0.01 to 0.30\textsuperscript{123,176}.

\textbf{Discussion}

The findings suggest a relatively wide range of instruments are commonly used in MS assessment, and those employed measure many varied MS symptoms/functions, and reflect the need to capture the diverse nature of this disorder’s symptomatology and impact on a person’s
functioning. These scales evaluate the symptoms/functioning in each of four broad categories: physical (e.g., physiological, fatigue, movement); cognitive (e.g., memory, attention); psychological and emotional (e.g., depression, anxiety); and quality of life (i.e. impacts of the symptoms/functioning on daily life and living). Naturally, there is overlap across the aspects that an individual scale measures, and between these broad categories. It is not always straightforward to ascertain which domain a questionnaire indexes; the content of some measures covers more than one domain. For instance, there can be difficulty discriminating the direct symptoms of MS and their effects on functioning from the impacts of those symptoms/functions on daily life. Some aspects of MS (e.g., gait, pain, incontinence, depression) may be physical or psychological problems and also quality of life problems, and it is not necessarily easy to divide them into separate domains. The impact of physical issues on an individual (e.g., producing depression) is not readily distinguishable from their impacts on what an individual can or cannot do (i.e. their quality of life). In developing test-sets, attention should be paid to these issues. However, it is useful to compare the assessment tools along the four dimensions, to allow comparison between the measures, and to develop appropriate MS assessment sets (see Figures 1 and 3).

In terms of physical symptoms, the measures that address these mainly focus on mobility and motor function, and deal with central nervous system damage. The majority of such tools are clinician-administered (Modified Ashworth Scale, EDSS, FIM, MSFC), with only two such scales being patient-completed (GNDS and MSWS-12). This may reflect the facts that these symptoms/functions are observable, and measurable, by the clinician, and that a patient’s judgement of their physical symptoms and functioning may be confounded with other factors
(e.g., mental state). Of the clinician-administered scales, three deal largely with mobility and motor function (Ashworth, EDSS, and MSFC). All of these give reliable measures of the present levels of disability. The patient-completed scales concerning motor symptoms/functioning (some aspects of the GNDS, and the MSWS-12) give reliable measures of current levels of disability, although not as reliable as the clinician-administered scales. However, the clinician-administered scales (especially the EDSS) suffer from a common problem of not being particularly responsive to change in disability status, possibly resulting from the nature of the physical problems themselves, in that, once mobility has been impaired, it may not then readily return to its former unimpaired state (especially in some types of MS). The patient-completed scales appear more responsive to change in disability status, but it is not clear whether this reflects their measure of the disability itself, or a person’s perception of their disability, which may be different, and the latter may be responsible for responsiveness to change. In this context, it is informative to consider current models of disability. For example, the World Health Organisation’s (WHO) International Classification of Functioning, Disability, and Health (ICF) acknowledges that ‘disability’ is a constructed experience – that is, the person’s own perception of their symptoms/functioning, in addition to any objective assessment of those issues, is a prime aspect of disability. In addition, the literature suggests that better personal perceptions of health are stronger predictors of confidence regarding a person’s involvement and participation in their community (which is also a key element defining disability in the WHO's ICF framework).

There are only two clinician-administered measures that deal with cognitive impairments (e.g., memory, attention, problem solving, information processing), which are the MMSE, and one component of the MSFC. However, these measures are rather poor, both in terms of their
psychometric properties, and in their lack of sensitivity to MS cognitive impairments. In addition, neither of these captures the potential range of impairments, focusing only on very specific and limited tests. These criticisms are also potentially true of the patient-completed measure of cognition contained in the GNDS. The measurement of this domain, by the commonly employed scales, is potentially the weakest, and some consideration should be given to employment of better validated tests of cognitive function, or, at least, some of their sub-scales, such as the WAIS.

A third area of symptomatology or functioning that some of these instruments address is the psychological and emotional impacts of MS. Some of these symptoms/functions may result from CNS damage, but many will also result from the impact of MS on the ability of a person to conduct activities in which they wish to engage. Due to the subjective nature of these symptoms and their impacts on functioning, all of the scales that assess these aspects are patient-completed. In terms of depression, both the BDI and the HADS have very good psychometric properties for MS. The HADS has a slight advantage in that it is shorter, and it measures both anxiety and depression. The EQ-5D, GHQ-28, GNDS, and the SF-36, all contain measures of psychological function, that correlate with the BDI and HADS, but they are not necessarily related to clinical definitions of depression, and also reflect the impacts of these problems, rather than the problems themselves. In terms of fatigue, the FIS and MFIS are fatigue-specific instruments, of which the MFIS is thought to have an advantage psychometrically, especially when used for the separate impacts of fatigue on physical and mental functioning. The only measure that deals specifically with pain symptoms is the EQ-5D, although other measures assess the impact of pain on limiting...
activities. All of these measures have reasonable psychometric properties, but not all of them are MS-specific, and this may merit consideration.

In regards to quality of life, the psychometrics of almost all of these measures are reasonably robust, although not all of these measures are MS-specific. Typically, it is the case that, the longer the measure, the better the psychometric properties. Most of these longer measures give an index of both physical and mental quality of life functioning. Of the nine quality of life measures (EQ-5D, FIS, GHQ, LMSQoL, MFIS, MSIS-29, MSQoL-54, SF-36), four appear to offer very good coverage of quality of life issues, as well as having excellent psychometric properties (GHQ, MSIS-29, MSQoL-54, SF-36). Of these, the MSIS-29 is MS-specific, the MSQoL-54 has some MS-specific questions, and also encompasses the generic SF-36. However, in choosing an appropriate questionnaire, issues regarding the MS-specificity of a scale’s items should be considered. It has been recommended that the MFIS is used, when working with people with MS, as it was developed on an MS population, however, the MSIS-29 is MS-specific, and this measure offers better psychometric properties than the MFIS. The generic GHQ and the SF-36 do similar things to one another, as the SF-36 is derived from the GHQ, and so would not be employed together. The MSQoL-54 is MS-specific, and encompasses the SF-36, and it is similar, in itself, to the MSIS-29, which is shorter. So a choice may be between using the GHQ and the MSIS-29, or using the MSQoL-54. An ideal approach is to choose a set of tests in which there is some overlap across the various measures; for example, when assessing quality of life, it would be desirable to use measures that examine the impact of physical symptoms/functioning in conjunction with other measures that examine the objective physical symptoms/functioning. In addition, a generic HRQoL measure (such as the EQ-5D) should be
included in any set of outcome measures to enable comparison with population norms, with other conditions, and to facilitate health economic analyses using QALYs, which generic HRQoL measures can help to produce.

In summary, the key to better assessment of MS is to develop tailored balanced sets of measures that are both complementary and comprehensive. Given the multiple symptom/function domains of MS, which vary in quantity and quality across patients, it would be very difficult to prescribe a single recommended battery of measures to detect changes in impairment; these would vary according to the needs of the clinician, researcher, or patient. Such assessment sets should include measures for each of the four areas discussed above, with consideration to time efficiency, minimising replication or scale redundancy, and not fatiguing the patient. Provided here is a resource from which suitable sets of commonly-used measures can be selected, determined by the specific requirements of each situation. This review could aid such developments, and be a useful resource in itself.
Acknowledgments

Recognition and gratitude go to Professor Rhys Williams who helped to initiate this project. Thanks go to Ed Holloway and to the Neurologists who volunteered their input and time, as well as to Professor Ceri Phillips for his helpful suggestions.
Disclosure Statement

This work was supported by a grant from the Multiple Sclerosis Society, UK. There are no conflicts of interest.
References


Outcome measures for MS - 35


Figure 1: Psychometric characteristics

<table>
<thead>
<tr>
<th>Property</th>
<th>Measured</th>
<th>Assesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Reliability</td>
<td>Chronbach’s Alpha</td>
<td>The degree to which all items are related to one another</td>
</tr>
<tr>
<td>Test-Retest Reliability</td>
<td>Correlation between test administered at two separate times</td>
<td>The freedom from unsystematic error over time</td>
</tr>
<tr>
<td>Concurrent Validity</td>
<td>Correlation between measure and other measures in the area</td>
<td>Whether the scale compares well with other similar tools</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>Test-identified positives divided by all actual positives</td>
<td>The degree that the measure identifies individuals with the disorder (true positives)</td>
</tr>
<tr>
<td>Specificity</td>
<td>Test-identified negatives divided by all actual negatives</td>
<td>The degree to which the measure does not identify people without the disorder (true negatives)</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>Effect sizes, or area under the receiver operating characteristics curve</td>
<td>The ability to detect changes in the measure over time</td>
</tr>
</tbody>
</table>
Table 2: Characteristics of measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Generic or MS Specific</th>
<th>Clinician Administered or Patient Completed</th>
<th>Approximate Length of Time to Administer or Complete</th>
<th>Domains Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified Ashworth Scale</td>
<td>Generic</td>
<td>Clinician</td>
<td>Short (depending on number of joints assessed)</td>
<td>Physical Disability: Muscle spasticity, joint by joint</td>
</tr>
<tr>
<td>Beck Depression Inventory</td>
<td>Generic</td>
<td>Patient</td>
<td>10 min</td>
<td>Psychological/Emotional: Symptoms of depression (e.g., hopelessness, irritability, guilt, and somatic problems)</td>
</tr>
<tr>
<td>Euro Qol - 5 Dimension Questionnaire</td>
<td>Generic</td>
<td>Patient</td>
<td>5 min</td>
<td>Quality of Life: Mobility, self-care, usual activities, pain, anxiety/depression, and general health.</td>
</tr>
<tr>
<td>Expanded Disability Status Scale</td>
<td>Generic (mainly used for MS)</td>
<td>Clinician administered and rated</td>
<td>Variable (depending on neurological examination)</td>
<td>Physical Disability: Impairment and disability (with 8 additional scales: pyramidal, cerebellar, brainstem, sensory, bowel and bladder, visual, cerebral, and other functioning)</td>
</tr>
<tr>
<td>Fatigue Impact Scale and Modified</td>
<td>Generic (developed on an MS population)</td>
<td>Patient</td>
<td>20 min</td>
<td>Quality of Life: Overall impact of fatigue, and sub-scales relating to the: physical, cognitive, and psychosocial, impacts of fatigue</td>
</tr>
<tr>
<td>Functional Independence Measure</td>
<td>Generic</td>
<td>Clinician or trained layperson</td>
<td>30 min</td>
<td>Quality of Life: Independence and ability of a patient to conduct routine daily activities concerned with motor function (e.g., self-care, sphincter control) and cognitive function (primarily social and communication function)</td>
</tr>
<tr>
<td>Outcome measures for MS</td>
<td>General Health Questionnaire</td>
<td>Generic</td>
<td>Patient</td>
<td>Up to 8 min (depending on version used)</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------</td>
<td>---------</td>
<td>---------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>Guy’s Neurological Disability Scale</td>
<td>MS Specific</td>
<td>Clinician or Patient</td>
<td>30 min</td>
<td>Physical Disability: Overall disability (12 sub-scales: cognition, mood, vision, speech, swallowing, upper and lower limb function, bladder and bowel function, sexual function, fatigue, and other)</td>
</tr>
<tr>
<td>Hospital Anxiety and Depression Scales</td>
<td>Generic</td>
<td>Patient</td>
<td>5 to 10 min</td>
<td>Psychological/Emotional: Anxiety and Depression</td>
</tr>
<tr>
<td>Leeds Multiple Sclerosis QoL Scale</td>
<td>MS Specific</td>
<td>Patient</td>
<td>5 min</td>
<td>Quality of Life: Overall index of quality of life, mostly focuses on fatigue and social issues</td>
</tr>
<tr>
<td>Mini Mental State Examination</td>
<td>Generic</td>
<td>Clinician</td>
<td>20 min</td>
<td>Cognitive Impairment: Cognitive impairment across a number of domains (e.g., attention, memory, orientation, arithmetic)</td>
</tr>
<tr>
<td>Multiple Sclerosis Functional Composite</td>
<td>MS Specific</td>
<td>Clinician</td>
<td>Variable (depending on disability level)</td>
<td>Physical/Cognitive Impairment: Impairment across 3 domains (leg function, arm function, and cognitive function)</td>
</tr>
<tr>
<td>Multiple Sclerosis Impact Scale</td>
<td>MS Specific</td>
<td>Patient</td>
<td>10 min</td>
<td>Quality of Life: Impact of MS on daily living on both physical and psychological aspects</td>
</tr>
<tr>
<td>Multiple Sclerosis QoL-54</td>
<td>MS Specific (generic component)</td>
<td>Patient</td>
<td>11 to 18 min according to Vickrey et al. 1995 (at least 30 min according to Bandari et al. 1994)</td>
<td>Quality of Life: Overall quality of life with two domains Physical Health (physical function, health perceptions, energy/fatigue, role limitations - physical, pain, sexual function, social function, and health distress); and Mental Health (anxiety, depression, social dysfunction, and overall quality of life).</td>
</tr>
</tbody>
</table>
Outcome measures for MS - 50

<table>
<thead>
<tr>
<th>Condition</th>
<th>Scale</th>
<th>Test Type</th>
<th>Time</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Sclerosis Walking Scale</td>
<td>MS Specific (has been used for other disorders)</td>
<td>Patient</td>
<td>5 min</td>
<td>Physical Impairment: Disruption to walking and walking quality</td>
</tr>
<tr>
<td>Health Scale Short Form-36</td>
<td>Generic</td>
<td>Patient</td>
<td>10 to 15 min</td>
<td>Quality of Life: Overall impact of a disorder on functioning (with sub-scales for physical health and mental health, which, themselves, are based on a number of sub-scales that measure: physical functioning, role limitation, pain, general health, vitality, social functioning, and mental health.)</td>
</tr>
</tbody>
</table>
Table 3: Quality of the psychometric properties of measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Internal Reliability</th>
<th>Test-Retest Reliability</th>
<th>Concurrent Validity</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified Ashworth Scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beck Depression Inventory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Euro Qol - 5 Dimension Questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expanded Disability Status Scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modified/ Fatigue Impact Scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional Independence Measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Health Questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guy’s Neurological Disability Scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital Anxiety and Depression Scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leeds MS QoL Scale</td>
<td>Mini Mental State Examination</td>
<td>MS Functional Composite</td>
<td>Multiple Sclerosis Impact Scale</td>
<td>Multiple Sclerosis QoL-54</td>
<td>Multiple Sclerosis Walking Scale</td>
<td>Health Scale Short Form-36</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------------</td>
<td>------------------------</td>
<td>---------------------------------</td>
<td>--------------------------</td>
<td>-------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key:**

<table>
<thead>
<tr>
<th>Not Applicable</th>
<th>Poor (&lt; 0.4)</th>
<th>Moderate (0.4-0.7) /Mixed or Varied Values/Some Caution Required</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Outcome measures for MS - 52
<table>
<thead>
<tr>
<th>Outcome measures for MS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (&gt; 0.7)</td>
<td></td>
</tr>
<tr>
<td>No Data Available</td>
<td></td>
</tr>
</tbody>
</table>