Health Technology Assessment 2004; Vol. 8: No. 9

Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome

JC Hobart, A Riazi, DL Lamping,
R Fitzpatrick and AJ Thompson

March 2004

Health Technology Assessment
NHS R&D HTA Programme

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome

JC Hobart,[1,2*] A Riazi,[1] DL Lamping,[3] R Fitzpatrick[4] and AJ Thompson[1]

[1] Neurological Outcome Measures Unit, Institute of Neurology, London, UK
[2] Peninsula Medical School, Derriford Hospital, Plymouth, UK
[3] Health Services Research Unit, London School of Hygiene and Tropical Medicine, UK
[4] Division of Public Health and Primary Health Care, Institute of Health Sciences, University of Oxford, UK

* Corresponding author

Declared competing interests of authors: none

Published March 2004

This report should be referenced as follows:

Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome. *Health Technol Assess* 2004;**8**(9).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE.

NHS R&D HTA Programme

The NHS R&D Health Technology Assessment (HTA) Programme was set up in 1993 to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS.

Initially, six HTA panels (pharmaceuticals, acute sector, primary and community care, diagnostics and imaging, population screening, methodology) helped to set the research priorities for the HTA Programme. However, during the past few years there have been a number of changes in and around NHS R&D, such as the establishment of the National Institute for Clinical Excellence (NICE) and the creation of three new research programmes: Service Delivery and Organisation (SDO); New and Emerging Applications of Technology (NEAT); and the Methodology Programme.

This has meant that the HTA panels can now focus more explicitly on health technologies ('health technologies' are broadly defined to include all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care) rather than settings of care. Therefore the panel structure was replaced in 2000 by three new panels: Pharmaceuticals; Therapeutic Procedures (including devices and operations); and Diagnostic Technologies and Screening.

The HTA Programme will continue to commission both primary and secondary research. The HTA Commissioning Board, supported by the National Coordinating Centre for Health Technology Assessment (NCCHTA), will consider and advise the Programme Director on the best research projects to pursue in order to address the research priorities identified by the three HTA panels.

The research reported in this monograph was funded as project number 95/01/03.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for any recommendations made by the authors.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The editors and publisher have tried to ensure the accuracy of this report but do not accept liability for damages or losses arising from material published in this report. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.

G

Abstract

Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome

JC Hobart,[1,2*] A Riazi,[1] DL Lamping,[3] R Fitzpatrick[4] and AJ Thompson[1]

[1] Neurological Outcome Measures Unit, Institute of Neurology, London, UK
[2] Peninsula Medical School, Derriford Hospital, Plymouth, UK
[3] Health Services Research Unit, London School of Hygiene and Tropical Medicine, UK
[4] Division of Public Health and Primary Health Care, Institute of Health Sciences, University of Oxford, UK
* Corresponding author

Objectives: To develop a patient-based, disease-specific measure of the health impact of multiple sclerosis (MS) for use in clinical trials and clinical practice.
Data sources: People with MS. Members of the MS Society of Great Britain and Northern Ireland.
Methods: Standard psychometric methods were used to develop the Multiple Sclerosis Impact Scale (MSIS-29) in three stages. Stage 1 (item generation): questionnaire items were generated from 30 patient interviews on the impact of MS on their lives, expert opinion and literature review. Stage 2 (item reduction and scale generation): the questionnaire developed in stage 1 was administered by postal survey to 1530 randomly selected members of the MS Society. Standard item reduction techniques were used to develop a rating scale from the pool of questionnaire items. Stage 3 (psychometric evaluation): the questionnaire was evaluated for data quality, scaling assumptions, acceptability, reliability and validity in a separate postal survey of 1250 MS Society members. Responsiveness was evaluated in 55 people admitted to hospital for rehabilitation and intravenous steroid treatment of MS relapses.
Results: Stage 1 resulted in a 129-item questionnaire. Stage 2 resulted in a 29-item rating scale measuring the physical and psychological impact of MS. The MSIS-29 satisfied all recommended psychometric criteria for rigorous measurement. Data quality was excellent: missing data were low, item test–retest reliability was high and scale scores could be generated for over 98% of respondents. Item descriptive statistics, item convergent and discriminant validity, and factor analysis supported summing items to produce two summary scores. MSIS-29 physical and psychological scale scores showed good variability, low floor and ceiling effects, good internal consistency and test–retest reliability. Correlations with other measures and confirmation of hypotheses about group differences provided evidence for the validity of the MSIS-29 as a measure of the physical and psychological impact of multiple sclerosis. Effect sizes provided preliminary evidence for responsiveness.
Conclusions: The 29-item MSIS-29 is a rigorous new measure of the physical and psychological impact of MS. All psychometric criteria were satisfied and there is preliminary evidence of responsiveness. The MSIS-29 is particularly appropriate for use in clinical trials to evaluate therapeutic effectiveness from the patient's perspective. Further critical evaluations of the MSIS-29 completed by people with neurologist-confirmed MS in different settings are suggested. Head-to-head comparisons of the psychometric properties of the MSIS-29 and other outcome measures for MS will help to determine the relative advantages of different instruments so that the choice of measures for studies can be evidence based.

Contents

List of abbreviations

AERA	American Educational Research Association
ANOVA	analysis of variance
APA	American Psychological Association
BI	Barthel Index
EDSS	Expanded Disability Status Scale
EQ-5D	EuroQol
ES	effect size
FAMS	Functional Assessment of MS
GHQ-12	General Health Questionnaire
GNDS/UKNDS	Guy's (now UK) Neurological Disability Scale
HRQOL-MS	Health-related Quality of Life Questionnaire for MS
ICC	intraclass correlation coefficient
ION	Institute of Neurology
IR	Irene Richardson
MD	missing data
MEF	maximum endorsement frequency
MRI	magnetic resonance imaging
MS	multiple sclerosis
MSIS-29	Multiple Sclerosis Impact Scale
MSQLI	MS Quality of Life Inventory
NA	not applicable
NCME	National Council on Measurement in Education
NHNN	National Hospital for Neurology and Neurosurgery
PCA	principal components analysis
SE	standard error
SF-36	Medical Outcomes Study 36-Item Short Form Health Survey
SIP	Sickness Impact Profile

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices in which case the abbreviation is defined in the figure legend or at the end of the table.

Executive summary

Background

Multiple sclerosis (MS) is an incurable progressive neurological disorder that has a profound impact on people's lives. Although a wide range of problems has been documented, the impact of MS from the individual's perspective has not been systematically and directly measured. There is no outcome measure that incorporates patients' own perspectives about the impact of MS that is sufficiently rigorous to be used in treatment trials, epidemiological studies and audit. This report describes the development and validation of a new instrument, the Multiple Sclerosis Impact Scale (MSIS-29), a rigorous measure of the physical and psychological impact of MS from the patient's perspective.

Objectives

To develop a patient-based, disease-specific measure of the health impact of MS that is clinically useful, and scientifically sound, and suitable for use as an outcome measure in clinical trials and in routine clinical practice.

Methods

Standard psychometric methods were used to develop the MSIS-29 in three stages.

- Stage 1 (item generation): questionnaire items were generated from 30 patient interviews on the impact of MS on their lives, expert opinion and literature review.
- Stage 2 (item reduction and scale generation): the questionnaire developed in stage 1 was administered by postal survey to 1530 randomly selected members of the MS Society. Standard item reduction techniques were used to develop a rating scale.
- Stage 3 (psychometric evaluation): the rating scale was evaluated for data quality, scaling assumptions, acceptability, reliability and validity in a separate postal survey of 1250 MS Society members. Responsiveness was evaluated in 55 people admitted to hospital for rehabilitation and intravenous steroid treatment of MS relapses.

Results

- Stage 1: a pool of 129 items was generated.
- Stage 2: the item pool was reduced to a 29-item measure of the physical (20 items) and psychological (nine items) impact of MS: the MSIS-29.
- Stage 3: the MSIS-29 satisfied all recommended psychometric criteria for rigorous measurement. Data quality was excellent: missing data were low (maximum 3.9%), item test–retest reliability was high (r = 0.65–0.90) and scale scores could be generated for >98% of respondents. Item descriptive statistics, item convergent and discriminant validity, and factor analysis supported summing items to produce two summary scores. MSIS-29 physical and psychological scale scores showed good variability, low floor and ceiling effects, good internal consistency (Cronbach's $\alpha \geq 0.91$) and test–retest reliability (intraclass correlation ≥ 0.87). Correlations with other measures, and confirmation of hypotheses about group differences, provided evidence for the validity of the MSIS-29 as a measure of the physical and psychological impact of multiple sclerosis. Effect sizes (physical scale = 0.82, psychological scale = 0.66) provided preliminary evidence for responsiveness.

Conclusions and recommendations

The 29-item MSIS-29 is a rigorous new measure of the physical and psychological impact of MS. All psychometric criteria were satisfied and there is preliminary evidence of responsiveness. The MSIS-29 is particularly appropriate for use in clinical trials to evaluate therapeutic effectiveness from the patient's perspective.

A limitation of the study is that the MS Society membership database was used to define the sampling frame; the percentage of people in the database with a neurologist-confirmed diagnosis of clinically definite MS, the disease type of those with MS and the representativeness of people who join charitable groups are unknown.

Critical evaluations of the MSIS-29 completed by people with neurologist-confirmed MS in different settings will identify its strengths and weaknesses, and further define its role in clinical practice and research. Head-to-head comparisons of the psychometric properties of the MSIS-29 and other outcome measures for MS will help to determine the relative advantages of different instruments so that the choice of measures for studies can be evidence based.

Chapter 1

Overview of report

This report describes the development and validation of the Multiple Sclerosis Impact Scale (MSIS-29), a rigorous new measure of the physical and psychological impact of multiple sclerosis (MS) from the patient's perspective.

The MSIS-29 was developed and tested in three stages. In stage 1 (item generation), a 129-item questionnaire was generated from 30 patient interviews, expert opinion and a review of the literature. In stage 2 (item reduction and scale generation), the 129-item questionnaire was administered by postal survey to 1530 randomly selected members of the MS Society to identify items for elimination on the basis of psychometric performance. This process generated the MSIS-29. In stage 3 (psychometric evaluation), a comprehensive evaluation of the psychometric properties of the MSIS-29 (data quality, scaling assumptions, acceptability, reliability, validity and responsiveness) was undertaken in a postal survey of 1250 MS Society members and 55 people admitted to hospital for rehabilitation or intravenous steroid treatment.

Chapter 2 describes the evaluation of therapeutic interventions for MS and, in some detail, the psychometric concepts and methods used to develop and validate the MSIS-29. Readers familiar with psychometric methods may prefer to omit this section.

Chapter 3 presents the methods and results of stages 1 (item generation) and 2 (item reduction and scale generation).

Chapter 4 presents the methods and results of stage 3 (psychometric evaluation).

Chapter 5 presents a discussion of study results, study limitations and the implications of the MSIS-29 for healthcare, and provides recommendations for future research.

The appendices include a copy of the MSIS-29 and instructions for administration and scoring of this measure.

Chapter 2
Background

Overview

This chapter describes the evaluation of therapeutic interventions for MS and the psychometric methods and concepts used. Readers familiar with psychometric methods may prefer to skip to Chapter 3 after the section on Health outcomes measurement.

Evaluation of therapeutic interventions for MS

MS is an incurable progressive neurological disorder that has a profound impact on individuals and their families. Although the incidence in the UK is relatively low (2500 new cases/year), the prevalence is much higher (85,000). This is because MS tends to begin in young age groups, is incurable and in the majority of people is progressive over many decades. Although MS has little effect on longevity, it has a major impact on physical function, employment and quality of life. It is a complex disorder with diverse effects, an unpredictable course, and variable manifestations that pose unique problems to patients and their families. Moreover, the cost of MS in the UK is estimated to be £1200 million per year[1] and is expected to increase.[2] Costs due to MS have been shown to increase as disability progresses.[3,4] Psychosocial costs are less easily quantified, but no less real.

As MS is a major public health concern in Britain, beneficial interventions are to be welcomed. However, the outcomes of therapeutic interventions must be rigorously evaluated if policy decisions and clinical practice are to be evidence based. The need for more rigorous evaluation of treatments for MS has recently become critically important for several reasons. First, an increasing number of therapeutic pharmaceutical agents aimed at altering the course of MS is being introduced and their effectiveness needs to be determined.[5] Second, because the relative benefits of different interventions are likely to be marginal, analyses of comparative effectiveness are necessary.[6] Third, as treatments are expensive and may be required on a long-term basis, decisions about interventions based on short-term evaluation may have long-term economic implications. Fourth, as resources for the treatment of MS are required for other aspects of service provision, including rehabilitation and community support, resource allocation must be equitable. Finally, it is important that the current limited resources are allocated appropriately.

Evidence-based policy and clinical decision-making require rigorous measurement of outcomes. This information is of value when the outcomes that are evaluated are appropriate to patients and the instruments that are used are clinically useful and scientifically sound. Outcome measures in MS have traditionally focused on physiological parameters of disease and simple, easy to measure entities such as mortality morbidity and duration of survival. Although these assessments are important, they only partly address patients' concerns,[7] offer little information about diverse clinical consequences, fail to address the personal impact of disease,[8] and are of limited relevance in conditions that do not affect longevity. As new treatments for MS are aimed at altering its natural history or modifying its impact, traditional outcomes are inadequate in a comprehensive evaluation of therapeutic effectiveness.

Over the past two decades, outcome measurement in MS has relied heavily on the Expanded Disability Status Scale (EDSS).[9] This is an observer (neurologist)-rated scale which grades 'disability' due to MS in 20 steps on a continuum from 0 (normal neurological examination) to 10 (death due to MS). The EDSS was developed on the basis of the extensive clinical experience of a neurologist specialising in MS. It addresses impairment (symptoms and signs) at the lower levels (0–3.5), mobility in the middle range (4.0–7.5), and upper limb (8.0–8.5) and bulbar function (9.0–9.5) in the higher levels. Although the EDSS evaluates disability, it was developed before psychometric methods became familiar to clinicians, was not based on recognised techniques of scale construction,[10] and did not directly involve people with MS. More importantly, the EDSS is rated by neurologists rather than by patients themselves and has limited measurement properties.[11,12]

The lack of validated MS-specific measures has led to the use of generic measures, such as the Medical Outcomes Study 36-Item Short Form Health Survey (SF-36),[13] Sickness Impact Profile (SIP)[14] and EuroQoL.[15] Although generic measures have the advantage of enabling comparisons across diseases, it is increasingly recognised that they do not cover some areas of outcome that are highly relevant in specific diseases,[16] and may have limited responsiveness.[17] Psychometric limitations of the SF-36 in MS include significant floor and ceiling effects,[18] limited responsiveness,[18] underestimation of mental health problems,[19] and a failure to satisfy assumptions for generating summary scores.[20] Disease-specific instruments, consisting of items and domains of health that are specific to a particular disease, are more relevant and important to patients and clinicians and consequently are more likely to be responsive to subtle changes in outcome.[7,17,21]

Several MS-specific measures have been developed since the mid-1990s. These include the Functional Assessment of MS (FAMS),[22] the MSQOL-54,[23] the MS Functional Composite,[24] the Leeds MSQoL scale,[25] the Guy's (now UK) Neurological Disability Scale (GNDS/UKNDS),[26] the MS Quality of Life Inventory (MSQLI),[27] and the health-related quality of life questionnaire for MS (HRQOL-MS).[28] While all are encouraging, one limitation of these measures is that none was developed using the standard psychometric approach of reducing a large item pool generated *de novo* from people with MS. The FAMS and MSQOL-54 were developed by adding MS-specific items to existing measures, an approach that has been demonstrated to have some limitations.[29] The HRQOL-MS was developed through factor analysis of items from two generic measures and one MS-specific measure, and the MSQLI combines a large number of existing disease-specific and generic instruments. Items for the GNDS were developed through expert clinical opinion rather than on the basis of interviews with people with MS. Consequently, an outcome measure that is MS specific and combines patient perspective with rigorous psychometric methods will complement existing instruments. The aim of this study was to develop such a measure.

Health outcomes measurement: history, concepts and theory

The scientific discipline of health measurement[30] grew in response to the need to supplement clinical judgement with reliable and valid patient-based measures of health outcomes. Recently, there has been increasing recognition of the importance of assessing more patient-relevant consequences of disease, a practice that is now considered essential in a comprehensive evaluation of healthcare.[16,31] As measurement of such outcomes will influence decisions that affect patient welfare, policy development and the expenditure of public funds, it is essential that rigorous measurement instruments are used in healthcare evaluation.[32]

A simple but useful classification considers health outcomes in neurology to be either physician or patient-based. The most frequently studied physician-based outcomes in MS are magnetic resonance imaging (MRI) and relapse rate. Although these physician-based outcomes have the patient's interests at heart, they only address the pathological basis of MS and evaluate health in terms of quantity. They do not provide a complete picture of disease impact as they offer limited information about the diverse clinical consequences of MS and fail to incorporate subjective assessments of health.[16]

Patient-based outcomes are the consequences of disease and treatment that are considered important to patients. Patients are the best source of information about therapeutic benefit defined in terms of functioning and well-being.[33] As patients, their carers and their physician differ in their interpretation of the impact of illness,[34–39] it is important to elicit information from patients about which outcomes are important. This is supported by irrefutable evidence that patients can provide reliable and valid judgements of health status and the benefits of treatment.[40,41] Indeed, patient report has been described as the ultimate measure of health status.[42] In addition, the self-report method affords considerable methodological advantages over other methods of instrument administration.[43] For example, large numbers of geographically disparate patients can be accessed by postal survey, thus reducing selection bias while minimising patient discomfort and research staff involvement.

It is common to consider measures apart from traditional indicators of biological functioning as a single category of quality of life measures.[33] However, as quality of life encompasses factors not generally considered to be part of health per se (e.g. income and environment), the terms health-related quality of life,[7] health status[33] and functional status are commonly used interchangeably.[44,45] The title of a measure should be as descriptive as possible of the construct measured.

Psychometric theory

Although health measurement as a distinct discipline emerged in the 1980s,[46–48] it is derived from well-established theories and methods of measurement in the field of social sciences the origins of which can be traced to the mid-1800s. The basic scientific principles of measurement were established by mathematical psychologists interested in the human being as a measuring instrument. By studying how people make subjective judgements about measurable physical stimuli (e.g. length, weight, loudness), they developed the science of psychophysics: the precise and quantitative study of how human judgements are made.[49] The investigation of overt responses to physical stimuli requires precise methods, referred to as psychophysical methods, for presenting the stimuli and for measuring responses.[50]

The work of psychophysicists seems far removed from health measurement. In fact, it established the fundamental principles of subjective measurement which are as equally relevant to judgements about health as to judgements about physical stimuli. The psychophysicists demonstrated three important findings about human judgement: that subjective judgement is a valid approach to measurement, that humans make judgements about abstract comparisons in an internally consistent manner, and that accurate judgements can be made on ratio rather than simple ordinal scales. It is notable that psychophysical methods are still used in neurology; thermal threshold testing is based on the principle of the just noticeable differences in temperature detection, and audiometry on a person's response to different sound frequencies.

While the psychophysicists were measuring subjective judgements about physical stimuli that could be independently and objectively measured and verified, experimental psychologists were attempting to measure human attributes for which there were no independent physical scales of measurement (e.g. intelligence, personality, attitudes).[50] Darwin's empirical demonstration of evolution in the *On the Origin of Species* in 1859 was the impetus behind the study of individual differences in psychology.[51] It was reasoned that if animals inherit ancestral characteristics, and if individual differences influence their ability to adapt and survive, so individual differences in humans would have functional significance and could be inherited. Galton, who followed Darwin and believed that the human race could be bettered through controlled mating (eugenics), realised that human characteristics must be measured in a standardised manner before their inheritance could be studied. He coined the term 'mental test' for any measure of a human attribute, and set about the large-scale testing of sensory discrimination and motor function in the belief that people with the most acute senses would be the most gifted and most knowledgeable.[51] However, when Galton's colleague Pearson developed and applied the correlation coefficient, it became clear that results from these simple sensory and motor tests bore almost no relationship to measures of intellectual achievement, such as school grades.[52] This finding prompted the development of the mental test movement, with a widespread interest in the development and application of mental testing, and the measurement of individual differences.

A major advance in mental testing[53] was made when Thurstone demonstrated that psychophysical scaling methods could be used to measure accurately psychological attributes.[54,55] This finding prompted the development of psychological (or psychometric) scaling methods, which are defined as procedures for constructing scales for the measurement of psychological attributes.[49] Spurred on by the practical need to measure diverse outcomes, the mental test movement flourished between 1930 and 1950 with the spread of standardised testing for assessing educational achievement, measuring attitudes and personality, and selecting and screening personnel. In addition, scientific interest in methods of testing led to the development of psychometrics as a prominent discipline within psychology and established the cornerstones of the scientific evaluation of measuring instruments based on reliability and validity testing.[49,56]

The growth and development of psychometrics required standards for the development and evaluation of measurement instruments. The first of these was introduced in 1954 by a committee of the American Psychological Association (APA).[57] The following year similar guidelines were prepared by a committee representing the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME).[58] Subsequently, standards have been published by the Committee to Develop Standards for Educational and Psychological Testing, which represents the APA, AERA and NCME,[59–61] along with a commitment to the continual review of measurement standards in psychology and education.[61]

Thus, when healthcare evaluation needed methods for measuring patient-orientated outcomes, the

technology already existed. Since the 1970s, the focus of healthcare evaluation has moved to the measurement of function (the ability of patients to perform the daily activities of their lives), how patients feel and their own evaluation of their health in general.[40] The primary source of this information is standardised surveys,[46] for which psychometric techniques of scale construction are highly appropriate.[40]

Two studies in the USA confirmed the value of psychometric methods in assessing health outcomes. The Health Insurance Study,[62] a randomised experiment conducted by The Rand Corporation between 1974 and 1981, demonstrated that psychometric methods can be used to generate reliable and valid measures for assessing changes in health status for both adults and children in the general population. Following on from this, the Medical Outcomes Study[40,63] demonstrated that psychometric methods of scale construction and data collection were successful for measuring health status in samples of sick and elderly people. This study also demonstrated that psychometrically equivalent short-form measures could be constructed from the original longer forms,[64] thereby reducing respondent and administrative burden and improving measurement efficiency. These two pivotal studies confirmed that psychometric methods, borrowed from the social sciences, generated scientifically sound and clinically useful health measures.

Psychometric theory posits that when a concept cannot be measured directly (e.g. health status), it can be measured by asking a series of questions or items, each of which addresses a different aspect of the same concept.[65] Analysis of a large number of items generated by clearly defined standard techniques allows one to reduce the number of items and to construct scales.[10] Instruments developed according to psychometric principles must then be formally evaluated to ensure that they measure the outcome of interest in a manner that is reliable (consistent, stable over time and reproducible), valid (measure what they purport to or are intended to measure) and responsive (able to detect clinically important change over time).[10,52,61,66–72]

Instrument development

Below is an overview of the psychometric methods with appropriate references. Further information on these methods is reported in the references cited.

Development of an outcomes measurement instrument in accordance with psychometric principles involves two stages: generation of a large item pool, followed by reduction of the initial item pool to form the final instrument. Items can be generated from a variety of sources, including patients, consensus opinion of experts in the field, literature review and critical review of existing measures. They are then pretested on a small sample to assess how easily they can be understood and completed, whether there are ambiguities of wording, whether there are any irrelevant, misleading or offensive items, and whether the content of each item is appropriate. Items are revised based on pretesting to produce a version to be evaluated in the preliminary field test.

Pretesting is critical for identifying problems with a questionnaire, such as problems with question content, which can cause confusion with the overall meaning of an item, as well as misinterpretation of individual terms or concepts. Pretesting is a broad term that incorporates many different methods or combination of methods[73] in the prefield and field testing phase. Examples of prefield techniques include respondent focus groups and cognitive laboratory interviews. The latter consist of one-to-one interviews using a structured questionnaire in which respondents describe their thoughts while answering the questions (this is also called the 'think aloud' interviews). Field techniques include behaviour coding, respondent debriefings, interviewer debriefings, split-panel tests, analysis of item non-response rates and analysis of response distributions.[73] For more information, see Ref. 73.

The purpose of the first field test is to reduce the number of items and to develop scales. The instrument is administered to a large sample of patients and results are analysed using standard psychometric techniques for item analysis.[10,66] First, items with poor response rates and very high or low endorsement frequencies (proportion of people who endorse each response alternative) are eliminated. The remaining items are analysed using a variety of techniques including exploratory factor analysis to determine the underlying dimensions (factors) of the instruments. Items are analysed for redundancy, homogeneity and discrimination ability. Redundancy can be described as the extent to which a pair of items measures the same construct. Homogeneity refers to the fact that all of the items are tapping different aspects of the same attribute, and not different parts of different traits.[10,48,74] Discrimination ability is the extent to which an item (or scale) can discriminate between those individuals who differ in the construct being measured.[48] Based on these results, items are

retained or eliminated and grouped into subscales to produce a final version of the instrument.

Instrument evaluation

Instrument evaluation is the assessment of six scientific properties: data quality, scaling assumptions, acceptability, reliability, validity and responsiveness.

Data quality

Indicators of data quality such as item non-response and missing scale scores determine the extent to which an instrument can be used successfully in a clinical setting. They reflect respondents' understanding and acceptance of a measure and help to identify items that may be irrelevant, confusing or upsetting to patients.[75] Data quality can be determined by calculating per cent missing data for items, item test–retest reproducibility and per cent computable scale scores. Item test–retest reproducibility is the degree to which an item of the questionnaire yields stable scores over time among respondents who are assumed not to have changed on the domains being assessed. When there are missing items, a scale score can be calculated provided that 50% or more of the items are completed. A psychometrically sound method of imputing data is to replace missing items with a person-specific mean score, the average score across completed items for that respondent.[13,76]

Scaling assumptions

Having developed an instrument in the manner outlined above, and having used factor analysis to group items into subscales, one can now make the following assumptions about the final version: first, that items are correctly grouped into scales and that items in the same scale measure the same construct; and secondly, that the items of each scale can be summed without weights to produce scale scores. These assumptions can be evaluated by examining five criteria.

Equivalence of item variances

If items of the same scale are summed to produce a score it is assumed that the responses to items do not require standardisation or weighting.[77] This assumption relies on items being roughly parallel and therefore having symmetrical item-response distributions and exhibiting equivalent means and standard deviations.

Equivalence of corrected item–total correlations

If items of the same scale are summated to provide a score, it is assumed that each item in the same scale contains the same proportion of information about the construct being measured. This assumption is met if item–total correlations (correlation between the score of an individual item and the scale total score) are approximately equal. The item–total correlation is corrected for overlap by subtracting the item score so that estimates of the item–total relationship are not spuriously inflated.[78] Recently, Ware and colleagues[79] stated that this criterion can be considered satisfied when values exceed 0.30, even if they vary. No empirical justification of this criterion is given.

Item convergent validity

If all items in a scale are measuring the same underlying intangible construct or 'latent variable'[80] each item should be substantially linearly related to the total score computed from other items in that group. This criterion of item convergent validity is supported if an item correlates substantially with its own scale. Different authorities interpret different values for corrected item–total correlations as substantial. These include 0.20,[48] 0.30[10] and 0.40.[81]

Item-discrimination criteria

If the items of an instrument are correctly grouped into scales, items within a particular grouping should correlate more highly with the concept they are hypothesised to represent than with the other concepts measured by the instrument. Hypothesised groupings of items are supported when correlations between an item and its own scale (item–own-scale correlation) are significantly higher than with other scales of the measure (item–other-scale correlation). The extent of this item-discrimination criterion can be gauged by calculating scaling success rates.[82] A scaling success occurs when the item–own-scale correlation is two–standard-errors (SE) or more greater than the item other scale correlation (SE of a correlation coefficient = $1/\sqrt{N}$).[83] An overall scaling success rate for a scale is the percentage of item scaling successes relative to the total number of item–own-scale correlations. Correlations within 2 SE of the corresponding convergent correlations indicate limited item discrimination. Therefore, a definite scaling success is defined as item–own-scale correlations greater than item–other-scale correlations by 2 SE or more. Possible scaling success is defined as item–own-scale correlations greater than item–other-scale correlations by less than 2 SE. Possible scaling failure is defined as item own correlations less than item–other-scale correlations by less than 2 SE. Definite scaling failure is defined as item–own-scale correlations less than item–other-scale correlations by 2 SE or more.

Factor analysis
Exploratory factor analysis is used to reduce the number of items and develop scales in the preliminary field test stage of the study. In this technique a number of decisions is taken that can have a substantial impact on the results and their interpretation. The resulting item structure of the instrument depends on choices regarding the factor model [principal components analysis (PCA) or common principal axis factor analysis], the number of factors that are appropriate, the rotation method selected and the other items that are included in the analysis.[84] In addition, the interrelationship of variables is left unspecified and it is impossible to test directly alternative theoretical structures underlying the data. Consequently, confirmatory factor analysis will be performed at a later stage after the instrument has been developed to assess the underlying structure of the final instrument.

Acceptability
An instrument is considered acceptable when score distributions adequately represent the true distribution of health status in the sample.[81]

Item score distributions are considered acceptable when four criteria are met: approximately equal endorsement across response categories;[49] maximum endorsement frequencies, calculated as the percentage of responses for the most frequently endorsed response category; less than 80%, for dichotomous response options [for multipoint (polychotomous) response options, this criterion is less, and there are no published guidelines][66] and minimal item floor and ceiling effects, calculated as the percentage of responses for the lowest and highest scores, respectively. Although there are no widely accepted criteria for maximum item floor and ceiling effects, two published recommendations are 75%[85] and 90%.[86]

Scale score distributions are considered acceptable when four criteria are satisfied: scores should span the full scale range;[46] mean scores should be situated near the scale midpoint;[87] scale floor and ceiling effects, calculated as the percentage of responses for the minimum and maximum scores, respectively, are minimal; and score distributions are not excessively skewed.[75] There are no widely accepted criteria for floor and ceiling effects and skewness for scales. Current recommendations are that scale floor and ceiling effects should not exceed 15%[88] or 20%[89] and that skewness statistics should be within the –1 to +1 range.[85]

Reliability
The reliability of an instrument is defined as the extent to which it is free from random error.[49] As reliability increases (or decreases), scores are more (or less) consistent and, therefore, measured variance reflects true variance in the construct (or random error). In keeping with this definition, reliability coefficients estimate the proportion of total score variance that is due to true score variance.[49] In practice, the evaluation of reliability is in terms of two different aspects of a measure: internal consistency and reproducibility.[82]

Internal consistency is the extent to which items are interrelated.[84] Three indicators of internal consistency can be derived: corrected item–total correlations, Cronbach's α coefficients and homogeneity coefficients.

Corrected item–total correlations have been discussed above in the section on 'Scaling assumptions' (p. 7). The higher the correlation, the higher the variance shared by the item and the total score, and the higher the reliability of the item.

Cronbach's α provides an estimate of reliability based on all possible correlations between two sets of items within a scale.[82] Although widely interpreted as such, strictly speaking α is not a measure of unidimensionality. Rather, α is a measure of level of mean intercorrelation weighted by variances. It will be higher when there is homogeneity of variances among items than when there is not. Furthermore, the formula for α also takes into account the number of items on the theory that the more items, the more reliable a scale will be.[90–92] That is, when the number of items in a scale is higher, α will be higher even when the estimated average correlations are equal. Alpha coefficients exceeding 0.80 are considered acceptable for scales used to make group comparisons, whereas the more stringent criterion of 0.90–0.95 is required for scales used to make individual comparisons.[10]

As α coefficients are related to scale length[90–92] Ware and colleagues[46] recommend that 'homogeneity' coefficients are also reported as indices of internal consistency. Homogeneity coefficients are simply the average item intercorrelations for scales; it is recommended that values exceed 0.30.[87] They are of particular value when comparing the internal consistency of instruments with differing numbers of items within their subscales.

Reproducibility
Reproducibility evaluates whether an instrument yields the same results on repeated assessments,

assuming that respondents have not changed on the domain being measured.[93] Examples of reproducibility are parallel-forms, rater and test–retest reproducibility. Parallel-forms reproducibility is used when psychometrically identical versions of the same questionnaire are developed (to overcome the effects of memory or learning). Rater reproducibility is of importance in non-self-report measures, and is concerned with agreement between two or more ratings made by the same observer (intra-rater) or different observers (inter-rater) for the same patients. Thus, test–retest reproducibility is the most relevant form of reproducibility for patient-based outcome measures because parallel forms of measures do not usually exist and most measures are self-completed. It is examined by readministering the instrument to the same respondents after a specified period. If the results from the two time points have high agreement, the instrument demonstrates high test–retest reproducibility. Although there is no rule about the length of the test–retest interval, it needs to be sufficiently long to ensure that respondents are unlikely to recall their previous answers, but not so long that changes in health have occurred.[10] Although the recommended range of the test–retest interval is between 2 and 14 days,[66] this must be influenced by the nature of the study.

Correlation coefficients are frequently used to measure test–retest reproducibility. This method has been criticised on the basis that the results may be highly correlated but systematically different.[94] Therefore, an intraclass correlation coefficient (ICC), a measure of agreement, is recommended. This uses analysis of variance (ANOVA) to determine how much of the total variability in scores is due to true differences between individuals and how much to variability in measurement.[95] Recommended minimum standards for reproducibility are 0.80 for group comparisons and 0.90–0.95 for group comparisons.[10]

Validity

Validity can be broadly defined as the extent to which an instrument measures the concept it purports or is intended to measure.[61,96–98] Validity of measurement cannot be proven; rather, accumulating evidence is gathered, much as in a court case.[72] There are three types of validity: content, criterion-related and construct.[96]

Content validity

This refers to how well an instrument covers the construct being measured. Appropriate methods of item generation and selection help to ensure content validity. For example, as only persons with MS can truly define the aspects of health status affected by the disease, they act as the ultimate expert opinion. By involving a broad spectrum of MS patients and field testing large samples, omission of important domains and thus poor content validity are less likely. Nevertheless, it provides only weak evidence for the validity of a scale.

Criterion-related validity

This examines the degree to which a measure correlates with gold standard (criterion) measures obtained at a similar point in time (concurrent validity) or at a later time (predictive validity). Both types of criterion-related validity are expressed as correlations between the scale (predictor) and the criterion. However, as it is rare to find gold standard measures in the field of health status, more indirect approaches are recommended to evaluate validity.[99]

Construct validity

This process is used to establish the validity of a measurement instrument when no criterion or universe of content is accepted as entirely adequate to define the attribute being measured.[96] Construct validity involves testing hypotheses about how the instrument is expected to perform and examining the extent to which empirical data support these hypotheses.[96] Although there are several methods for determining construct validity, two categories have been distinguished: internal and external construct validity,[100] or psychometric and clinical tests of validity.[101] In the absence of gold standard measures of health status, both types of validity should be evaluated, as they are independent, complementary and on their own insufficient.[101]

Internal construct validity involves statistical analyses of scale scores to determine whether hypotheses concerning the theoretical structure of the instrument are supported. These analyses include PCA, within-scale correlations and relative validity.[101,102] Evidence for construct validity is provided if factor analysis confirms that the instrument consists of distinct scales that have items consistent with those hypothesised, and if item discrimination criteria are supported (see 'Scaling assumptions', p. 7). Further evidence for construct validity is provided if correlations between the scales of an instrument conform to hypotheses about the magnitude and pattern of correlations. Relative validity assessment determines the degree to which the component scales of an instrument measure the underlying concept as defined by the most valid scale. For any

groups that are of interest (e.g. those who are more or less disabled), the measurement precision of an instrument is quantified as the degree to which it separates these two groups (the difference between the mean scores) relative to the variance within the groups. *F*-statistics, derived from a one-way ANOVA, take both of these attributes into account as they indicate the ratio of between-groups (systematic) variance to within-group (error) variance.[101] The higher the *F*-statistic, the greater the measurement precision. By comparing a number of instruments in the same sample, relative measurement precision is estimated as the ratio of pairwise *F*-statistics (*F* for one measure divided by *F* for another) and indicates, as a percentage, how much more (or less) precise one measure is compared with another at detecting group differences.[64] In practice, the instrument with the largest *F*-statistic can be chosen as the arbitrary standard and assigned a relative measurement precision of 1. By comparing different scales, relative validity can be estimated by the ratio of pairwise *F*-statistics (*F* for one measure divided by *F* for another).

In contrast, external (empirical) construct validity or clinical tests of validity examine the relationships between the score on a given scale and external variables measured simultaneously or at a different point in time. This is an attempt to demonstrate that the instrument (1) measures what it is supposed to measure (convergent construct validity), (2) does not measure what it is not designed to measure (discriminant construct validity), (3) distinguishes between groups in predictable ways (group differences construct validity), and (4) produces results consistent with theoretical expectation (hypothesis testing).[96,97]

Responsiveness

Responsiveness is the ability of an instrument to measure clinically important change over time. While reliability and validity are the major determinants of the scientific robustness of a measure, the ability of an instrument to detect clinically significant change is also essential when evaluating the relative benefits of different interventions. This is particularly important when treatments are associated with small but significant benefits (a feature of current-day interventions in MS), which may be undetected by measures that are unresponsive. In such cases a clinically appropriate, reliable and valid, but unresponsive instrument is of limited value.

Although several methods have been used to assess the responsiveness of an instrument, there is little consensus about which method is best and how results should be reported.[103] The most common method of determining responsiveness is to examine the change scores following an intervention of known efficacy. Results are reported as an effect size, a standardised change score. There are different ways of calculating effect sizes, depending on whether the denominator is the standard deviation of baseline scores,[104] the standard deviation of change scores [standardised response mean[105] or the standard error of change scores (*t*-statistics)[106]]. These different methods of calculating effect sizes generate estimates of different magnitude and there is no consistent relationship between them.[107] Responsiveness measures using effect sizes are termed prospective methods.[108]

Another method of estimating the ability of instruments to detect change is by comparing change scores on a health status instrument with an external criterion of change, such as a transition question, also referred to as the global scale of change.[109] In this method, either patients or clinicians assess the amount of change retrospectively using a transition question (e.g. 0 = no change, 1 = minimal improvement, 2 = moderate improvement, 3 = marked improvement). Responsiveness can then be determined in a number of ways, for example, correlating change scores with the transition question (high correlations indicate greater responsiveness).[110] Alternatively, the minimum clinically important difference can be calculated[111] by dividing the mean change score for minimally improved/deteriorated patients by the mean change score for unchanged patients. Finally, the coefficient proposed by Guyatt and colleagues[112] can be calculated (mean change score in patients judged to have changed divided by the standard deviation of change scores in patients judged to have not changed). Norman and colleagues[108] defined these as retrospective methods of examining responsiveness as they involve the determination of subgroups of patients on the basis of their degree of change, and then the retrospective computation of responsiveness. Recently, Norman and colleagues compared prospective and retrospective methods of reporting responsiveness[108] and demonstrated that there is no consistent relationship between the results generated by the two methods.

As each method of reporting responsiveness has significant limitations, it is important that the relative responsiveness of competing measures is examined. This analysis is rarely undertaken.

Chapter 3
Development of the MSIS-29

Overview

This chapter outlines the development of the MSIS-29, a 29-item questionnaire designed to assess the impact of MS on people's lives. The MSIS-29 was developed and tested in three stages. In stage 1 (item generation), a 129-item questionnaire was generated from 30 patient interviews, expert opinion and a review of the literature. In stage 2 (item reduction and scale generation), the questionnaire was administered by postal survey to 1530 randomly selected members of the MS Society. Standard item reduction techniques were used to develop a 29-item scale (MSIS-29) measuring the physical (20 items) and psychological (nine items) impact of MS. In stage 3 (psychometric evaluation), six psychometric properties of the MSIS-29 (data quality, scaling assumptions, acceptability, reliability, validity and responsiveness) were evaluated in two studies. Data quality, scaling assumptions, acceptability, reliability and validity were evaluated in an independent sample of 1250 members of the MS Society. The responsiveness of the MSIS-29 was evaluated in 55 people with MS admitted for inpatient rehabilitation or intravenous steroids for treatment of relapse. This chapter presents the methods and results of stages 1 and 2. The methods and results of stage 3 are presented in Chapter 4.

Methods

The MSIS-29 was developed at the Neurological Outcome Measures Unit of the Institute of Neurology (ION)/National Hospital for Neurology and Neurosurgery (NHNN), Queen Square, London.

Item generation

Generation of an item pool

A pool of 129 items concerning the health impact of MS was generated from three sources: semistructured interviews of people with MS, multidisciplinary expert opinion and a comprehensive literature review.

Thirty people with MS attending the MS clinical service of the NHNN consented to participate in semistructured interviews. They were selected to represent as much diversity of illness as possible in terms of disability, duration of illness, age of onset and educational level. None of the patients who were asked to participate refused to be interviewed. The sample included men and women in all diagnostic categories (i.e. primary progressive, secondary progressive and relapsing–remitting MS), who represented the entire range of disability and illness duration and an age range similar to that of the British population of people with MS. *Table 1* presents the characteristics of the sample of patients who participated in the semistructured interviews.

Interviews lasted for an average of 1 hour, and were tape recorded, transcribed and then content analysed. The interviews were carried out by a single investigator, Irene Richardson (IR), at either the patients' homes or in a consultation room at NHNN. Statements relating to the health impact of MS on people's lives were extracted from all interviews by IR and in parallel by one of the co-investigators (Ray Fitzpatrick) for approximately one-third of the interviews. The extraction process involved highlighting any phrase or a sentence made by the patient that referred to the health impact of MS on their lives. Where two sets of statements had been extracted, comparisons were made. Agreement was high.

The extraction process was reviewed twice: first, to adjust the level of inclusiveness required, in particular to identify the areas covered by the interviews which did not relate to quality of life issues (e.g. people's reaction to their diagnosis); and second, for the last eight interviews, only completely new items, which did not belong to any of the categories identified as irrelevant, were extracted. In total, 3750 health impact statements were extracted from the interviews (mean 125, range 64–212).

Extracted statements were then classified into 11 broad categories to facilitate presentation and readability (symptoms, activities of daily living, emotional impact of MS, doctor-related statements, drug side-effects, financial strain, required planning, public response, relapses, impact on and responses of significant others, and wheelchair-related statements). Thus, these were

TABLE 1 *Characteristics of samples*

Variable[a]	Samples		
	Semistructured interviews	First field test	Second field test
N[b]	3	766	713
Gender			
Female	56	74	71
Age			
Mean (SD)	41 (12)	51 (12)	52 (12)
Range	23–70	2–87	18–82
Ethnicity			
White	100	98	98
Years since MS onset			
Mean (SD)	12 (11)	19 (12)	19 (11)
Range	1–36	1–56	1–59
Mobility indoors			
Walks unaided	40	–[c]	32
Walks with an aid	23	–	40
Uses a wheelchair	37	–	28
Mobility			
Can walk	NA	79	–
Cannot walk	NA	21	–
Marital status			
Married	77	66	70
Living with others		83	81
Employment status			
Retired due to MS	63	54	56
Employed		18	19
Type of MS			
Primary progressive	13.3	Unknown	Unknown
Secondary progressive	43.4	Unknown	Unknown
Relapsing–remitting	43.3	Unknown	Unknown

[a] All values are percentages unless specified otherwise.
[b] For whom both physical and psychological scale scores could be computed.
[c] Question not asked.
NA, not applicable.

the emergent themes of the interviews; these categories came about by simply content analysing the statements into a more manageable format. In each category, statements were further organised into subcategories (e.g. the broad category of symptoms included subcategories such as spasms, numbness or fatigue).

The elimination of redundant items was conducted in two stages (via series of discussion among the study panel; see Acknowledgements section). First, redundant statements within each individual on a particular subcategory were eliminated. That is, statements made by the same individual with a high degree of overlap were discarded until only one relevant statement was retained. Next, redundant statements between patients were eliminated. For example, the statement 'I have spasms' was retained but 'I have muscular spasm at night' and 'all my muscles were going into spasms' were discarded (these three statements were all made by different individuals). For both stages of elimination, it was decided to retain statements that were broader in content than specific in content, and which captured that particular subcategory succinctly. Any disagreements were discussed among team members and agreement was reached in all cases.

Through this process, a first list of items was extracted containing 117 statements covering the whole range of issues raised by people with MS during the interviews. Items were again chosen to avoid idiosyncratic and highly specific responses. For example 'can only walk short distances' was chosen in preference to 'can only walk 200 yards'. After discussions, it was agreed by all team members that statements regarding 'coping with MS', 'positive impact of MS' and 'diagnosis' statements were to be excluded, as the intention was for the questionnaire to focus on the impact of MS on their daily life.

An additional 38 items were generated from the review of the literature and from interviews with health professionals at the NHNN who were involved in the care of people with MS (i.e. neurologists, neuropsychologists, nurses, occupational therapists, physiotherapists, social workers, speech and language therapists). There was no information to identify which of the 38 items were from healthcare professionals. However, all the items from the final measure (MSIS-29) can be referenced back to interviews from patients.

A preliminary 129-item questionnaire consisting of two sections was developed (see Appendix 1). Section 1 included items to evaluate how people perceive the impact of MS on various aspects of their lives. Section 2 included items to evaluate the extent of physical limitations of people with MS. These two sections were based on the most appropriate way to group the items without changing patients' words. The time-frame specified for all items was the previous 2 weeks before completion of the questionnaire. Although the choice of time-frames is arbitrary, 2 weeks was chosen for three very specific reasons. First, during pretesting, a number of patients commented that 2 weeks was the most appropriate time-frame. Second, MS clinicians commented that 2 weeks was the most clinically appropriate. Third, a 2-week time-frame is most suitable for use in clinical trials.

Response options

Examination of the content of the initial pool of 129 items indicated that two distinct question stems and response scales were required. The majority of items (n = 97) were best represented by the stem 'How much have you been bothered by…' with a five-point response option (1 = not at all, 2 = a little, 3 = moderately, 4 = quite a bit, 5 = extremely). The remaining items (n = 32) about activity limitations were best represented by the stem 'How much has your MS limited your ability to…' with a six-point response option (1 = not at all, 2 = a little, 3 = moderately, 4 = quite a bit, 5 = extremely, 6 = can't do at all).

Pretesting

The preliminary questionnaire (including the instructions, item-stems, items and response options) was reviewed for content, wording and clinical appropriateness by patients, clinicians and researchers who were involved in its development. The preliminary version of the questionnaire was then pretested, first, in an independent and heterogeneous sample of 20 people with MS who were attending the NHNN. These people were selected to be representative of the general MS population (see *Table 2* for a breakdown of patient characteristics). They were asked to fill in the questionnaire in the presence of the project coordinator and to comment on it by identifying items and instructions that were unclear, ambiguous, irrelevant, misleading or offensive, and to make suggestions for alterations to the questionnaire. A formal cognitive laboratory interview using the 'think-aloud' approach was used. Second, the 30 people who had been interviewed (in the item generation stage) were all sent the preliminary questionnaire. These people were asked to fill in the questionnaire and to comment on it in the same manner as described above (the response rate was 70%; 21 patients returned the questionnaire). In addition, ten of these people were contacted by telephone to discuss the questionnaire. The final version of the questionnaire for the first field test consisted of 129 items (section 1, 97 items; section 2, 32 items), The questionnaire also contained open-ended questions on any comments or suggestions about the questionnaire and socio-demographic questions.

Item reduction and scale formation

The 129-item questionnaire was administered by postal survey to 1530 people, randomly selected and geographically stratified, from the membership database of the MS Society of Great Britain and Northern Ireland. This sampling frame has the advantage of being representative of people with MS in the MS Society membership

TABLE 2 *Pretesting patient characteristics (*n = *20)*

	Patients
Gender	
Male	8
Female	12
Age (years)	
20–29	0
30–39	8
40–49	5
50–59	5
60–69	1
70–79	1
Type of MS	
Relapsing-remitting	5
Primary progressive	2
Secondary progressive	13
Mobility indoors	
Walks unaided	6
Walks with aid	7
Wheelchair	7

database. However, one disadvantage is that the representativeness of people who join charitable groups is unknown; those who join such groups may be the most affected by their condition and/or least able to cope with illnesses. A further disadvantage is that not all members have MS. Therefore, based on results of a pilot study,[113] a target sample size of 1530 was chosen to ensure 500 completed questionnaires with no missing data. A subsample of 400 people was randomly selected from the larger sample to study item test–retest reproducibility to ensure 125 completed questionnaires on two occasions with no missing data. Patients in the test–retest sample received two questionnaires in the same envelope: one to complete immediately (time 1), and a second in a sealed envelope with instructions to open and complete 10 days later (time 2). A postcard reminder to complete the time 2 questionnaire was sent on day 7. Non-responders received reminders (letter and questionnaire) at 2 and 4 weeks.[114] In the test–retest subsample, non-responders to the time 2 questionnaire did not receive a reminder.

Item reduction was an iterative process that followed a predetermined plan with three stages. The aim of stage 1 was to eliminate items on the basis of excessive missing data and redundancy. The aim of stage 2 was to define the most valid way of representing the remaining items as measurement scales. The aim of stage 3 was to refine these measurement scales.

Stage 1: eliminating items

First, item level missing data were examined. The published criterion of ≥ 10% was used to indicate excessive missing data and eliminate items.[115]

Second, the item–item correlation matrix of the items was examined. When a pair of items correlates very highly one of these items is considered redundant.[116] For each pair of items correlating at three sequential levels: ≥ 0.80, ≥ 0.75 and ≥ 0.70, the psychometric properties of those items were examined in terms of per cent missing data, endorsement frequencies, floor and ceiling effects, and item test–retest reproducibility. The item with the worst properties was eliminated. Where items were psychometrically equivalent, a decision was made by consensus opinion, considering all aspects of the items (e.g. clinical relevance, clarity, item length).

Third, the psychometric properties of the remaining items were examined. The predetermined criteria for item elimination were: maximum endorsement frequency ≤ 35% (extrapolated from 66), aggregate endorsement frequency of any two-item adjacent response categories ≤ 10%[115] and item test–retest reproducibility ≤ 0.50.[117]

Stage 2: defining scales

The remaining items were entered into a factor analysis. First, all items were entered into a PCA without rotation to determine whether there were any rogue items that should be eliminated.[118] Next, principal axis factoring with varimax rotation was undertaken.[84] Multiple criteria were used to determine how many factors to rotate: Eigenvalues greater than 1,[119] the scree test,[120] the 5% rule,[121] and trial rotations.[46] All potential factor solutions were examined for cross-loading (items loading on two or more factors by > 0.40 and/or items loading on two or more factors within 0.1 of each other,[118] clinical interpretability of item content and replicability of results in random split-half samples.

Stage 3: refining scales

Item groups modelled through factor analysis were then examined to determine whether they satisfied recommended psychometric criteria for summed rating scales. This included acceptability, scaling assumptions, reliability and validity. In addition, item convergent and discriminant validity were tested in random half samples.

Results

In total, 1202/1530 (78.6%) questionnaires were returned. Of these, 436 were returned blank (change of address or deceased $n = 113$, did not have MS $n = 207$, declined to participate $n = 97$, no reason given $n = 19$). The response rate was 63.3% [response rate = (1202 – 436)/(1530 – 113 – 207)]. Therefore, item analyses were performed on data for 766 people with MS (*Table 1*).

Stage 1: eliminating items

Please refer to *Table 3* for the number of items removed at each stage and Appendix 1 for the items removed. None of the items failed criteria for missing data. A total of 36 items was eliminated owing to item redundancy. A total of 51 items failed one or more of the predetermined criteria for item elimination. There were 42 items remaining at this stage.

Stage 2: defining scales

In total, 42 items were entered into a factor analysis. Neither PCA, nor principal axis factoring

TABLE 3 *Analysis plan for item reduction and scale formation*

No. of items	Analyses and reasons for item elimination
129	First, item-level MD were examined. It was predetermined that items with ≥ 10% MD would be eliminated. No items failed this criterion. Next, item–item correlations were examined at three sequential levels: ≥ 0.80, ≥ 0.75, ≥ 0.70. When a pair of items correlated above this level, they were examined, and the one with the worst psychometric properties (listed below) was eliminated. A total of 36 items was eliminated from these analyses (27, 5, and 4, respectively)
93	The remaining 93 items were examined against three criteria: • Floor/ceiling effect, or MEF > 35%; • aggregated endorsement frequency of two adjacent response options < 10%; • item test–retest reliability < 0.50 A total of 51 items failed ≥ 1 of these criteria and were eliminated (See Appendix 1, stage 1 for eliminated items: 129 to 42)
42	The remaining 42 items were entered into a factor analysis. Exploratory analyses (principal components and principal axis) requesting factors with Eigenvalues ≥ 1.0 unity did not produce a clear solution. Therefore, all solutions with 2–7 factors were examined. The two-factor solution was the most clinically and psychometrically appropriate, but three items were then deleted as they loaded similarly on both factors. This gave 39 items: two scales of 26 and 13 items. Their content concerned the physical and psychological impact of MS
39	The item convergent and discriminant validity of the 39 items was examined in random half-samples (n = 379, n = 372). Three items in the physical scale and two items in the psychological scale registered probable scaling failures in both random half-samples and were removed to generate a 34-item scale: 23-item physical scale; 11-item psychological scale
34	The item convergent and discriminant validity of the 34 items was examined in the random half-samples. Three items in the physical scale and one item in the psychological scale registered reproducible scaling failures in random half-samples. These items were removed to generate a 30-item scale: 20-item physical scale; 10-item psychological scale
30	Examination of item convergent and discriminant validity in the random half-samples indicated that one item in the psychological scale registered a probable scaling failure in both samples and was removed to generate a 29-item scale: 20-item physical scale; 9-item psychological scale (See Appendix 1, stage 2 for eliminated items: 42 to 29)
29	All items registered definite scaling successes in both random half-samples

MD, missing data; MEF, maximum endorsement frequency.

generated a clear factor structure when factors with Eigenvalues > 1.0 were requested. Similar findings were demonstrated when the 102 items (see above) were factor analysed. Consequently, the study group examined all factor solutions with between two and seven factors generated by principal axis factoring with varimax rotation of the 42 items, cross-validated in random half samples. *Table 4* summarises the results of these analyses and the criteria applied. The two-factor solution best met the empirical criteria of reducing number of cross-loading items, good split-half replicability, and factors that were broadly conceptually and clinically interpretable at this stage.[118] As *Table 4* indicates, three of the 42 items ('bothered by numbness or loss of sensation', 'bothered by problems with vision when reading' and 'bothered by constipation') did not load on either factor ≥ 0.40 and were eliminated, leaving 39 items. Of these 39 items, a total of three cross-loaded, that is loaded ≥ 0.40 onto both factors. However, the difference between the magnitude of the loadings was > 0.10. Two items were from 26-item factor 1 ('bothered by difficulties planning things on a day-to-day basis' and 'bothered by feeling that you have missed out on things because of your MS') and one item was from 13-item factor 2 ('bothered by feeling frustrated'). The study group discussed the content of the two factors. The consensus opinion was that their best interpretation was the impact of MS on physical functioning (factor 1) and psychological well-being (factor 2).

TABLE 4 *Summary of solutions with two to seven factors when 42 items were entered into a factor analysis*

Solution	Amount of variance explained (%)	Total no. of cross-loading items (where both loadings > 0.4), (items not eliminated)	No. of cross-loading items with difference between the loadings < 0.1 (eliminated)	No. of items not loading anywhere > 0.4 (eliminated)	No. of items remaining	Split-half replicability	Factorial clarity
2 factor	47	3	0	3	39	Good	Interpretable
3 factor	50	11	8	3	31	Some are acceptable, but fair amount of factor swapping (esp. in model 2)	Interpretable
4 factor	53	9	3	5	34	Acceptable for factor 1, but poor for later factors (esp. model 1)	Interpretable, although some difficulty distinguishing F1 and F2
5 factor	55	11	4	3	35	NA	Difficult
6 factor	57	10	6	3	33	NA	Difficult
7 factor	58	10	5[a]	3	34	NA	Difficult

[a] One of these items loaded on three factors. The difference between the largest and smallest loading was >0.1, but other differences were < 0.1. Therefore, the item was excluded.

Stage 3: refining scales

The study group considered that five items were conceptually difficult to retain within their respective factors. Three items were in the physical scale: 'bothered by feeling that you have missed out on things because of your MS', 'bothered by the effect of MS on your spouse/partner or family' and 'bothered by having to change long-term work plans'. The other two items were in the psychological scale: 'bothered by pins and needles, tingling or burning sensations' and 'bothered by pain'. These five items were removed, leaving a 34-item instrument: 23-item physical impact scale and 11-item psychological impact scale.

The psychometric properties of the 34-item instrument were examined. All criteria examined were satisfied. These included acceptability, scaling assumptions, reliability and validity. However, on testing item convergent and discriminant validity in random half-samples, one item from each scale registered probable scaling successes in both samples, indicating a limited ability to discriminate between the two scales. These items, 'bothered by hot or cold temperatures making your MS worse' (physical impact scale) and 'bothered by feeling frustrated' (psychological impact scale), were eliminated, leaving 32 items.

Two other items were eliminated from the physical scale despite good psychometric properties. 'MS has limited my ability to walk indoors' was removed by the study group for two reasons. First, its content was considered to be well covered by another item ('difficulties moving about indoors'). Second, the aim was to develop a scale that was applicable to all people with MS, not just those who were able to walk. 'Bothered by difficulties planning things on a day-to-day basis' was removed because the study group thought it did not fit well with the conceptual definition of physical impact. It was assumed that this item performed well because there is a relationship between physical function and difficulties planning things on a day-to-day basis.

Removal of these items resulted in a 30-item instrument: 20-item physical impact scale and ten-item psychological impact scale. The psychometric properties of this instrument were examined. Testing of item convergent and discriminant validity in random half-samples indicated that one item in the psychological scale, 'bothered by getting tired when you do things', scored probable scaling success rates in both samples. This finding was consistent with clinical logic as this item reflected the construct of psychological impact less than the other items. The item was removed.

The psychometric properties of the 29-item instrument were examined. All criteria were satisfied. Item convergent and discriminant validity testing in random half-samples registered 100% definite scaling success rates in both samples. This iterative process of item reduction generated a 29-item instrument, the MSIS-29,

TABLE 5 *Score distributions and reliability of the MSIS-29 in first field test*

Variable	MSIS-29 scale	
	Physical (*n* = 751)	**Psychological (*n* = 751)**
No. of items	20	9
Scale range (median)	20–100 (60)	9–45 (27)
Sample range	20–100	9–45
Mean (SD)	61.9 (20.3)	23.3 (8.4)
% Floor effect (*n*)	1.2 (9)	1.6 (12)
% Ceiling effect (*n*)	0.8 (6)	0.3 (2)
Skewness/SE skew	–0.220/0.089	0.266/0.089
Type of reliability		
Internal consistency (*n* = 751)		
Item intercorrelation: range (mean)	0.30–0.68 (0.49)	0.26–0.69 (0.45)
Corrected item-total correlation (range)	0.53–0.78	0.47–0.77
Cronbach's α coefficient	0.95	0.88
Reproducibility (*n* =156)		
Test–retest (ICC)	0.92	0.82

with a 20-item physical impact scale and nine-item psychological impact scale. The psychometric properties of this instrument were examined in independent samples – the second field test.

The MSIS-29 included 26 items with five-point response options and three items with six-point response options. The latter three items were rescaled (category 5 combined with 6) so that all items had the same number of response options. This rescaling did not change the psychometric properties of the scales. All of the final 29 items can be referenced back to statements derived from the qualitative interviews with MS patients.

The physical and psychological summary scores are generated by summing individual items and then transforming these to a 0–100 scale using the formula:[40]

$$\frac{100 \times (\text{observed score} - \text{minimum score})}{(\text{maximum score} - \text{minimum score})}$$

High scores indicate poorer health. For respondents with missing data, missing values were imputed using a respondent-specific mean score in cases where at least 50% of the items in a scale had been completed.[13]

Preliminary psychometric analyses of data collected in the first field test indicated that the MSIS-29 satisfied standard criteria for acceptability and reliability (*Table 5*).

Chapter 4

Psychometric evaluation of the MSIS-29

Overview

This chapter outlines the psychometric evaluation of the MSIS-29. Six psychometric properties of the MSIS-29 were evaluated: data quality, scaling assumptions, acceptability, reliability, validity and responsiveness. Data quality, scaling assumptions, acceptability, reliability and validity were evaluated in a randomly selected sample of 1250 members of the MS Society. Responsiveness was evaluated in 55 people with MS admitted for inpatient rehabilitation or for intravenous steroids for treatment of relapse. The results demonstrated that the MSIS-29 satisfied all psychometric properties.

Data quality, scaling assumptions, acceptability, reliability and validity

Methods

Item reduction analyses produced the MSIS-29, a 29-item measure that includes two scales: physical impact (20 items) and psychological impact (nine items). All items could be referenced back to statements made by patients during interviews. The psychometric properties of the MSIS-29 were comprehensively evaluated in an independent sample. A second postal survey of randomly selected and geographically stratified members of the MS Society (n = 1250) was undertaken to evaluate data quality, scaling assumptions, acceptability, reliability and validity. The sample was divided randomly into three subsamples to evaluate convergent validity and test–retest reliability (n = 500, 500, 250). Respondents in the two larger subsamples completed the MSIS-29, demographic questions and three other health measures. Respondents in the first validity subsample completed the SF-36, EuroQol (EQ-5D)[15] and postal Barthel Index (BI).[122] Respondents in the second validity subsample completed the FAMS, EQ-5D and the 12-item version of the General Health Questionnaire (GHQ-12).[123] Respondents in the third test–retest subsample completed the MSIS-29 on two occasions separated by a 10-day interval. The same postal survey methods as used in the first field test were also used in the second field test, including an initial mailing followed by postal reminders at 1, 2 and 4 weeks.

t-Tests were used to compare the three subsamples described above, and between time 1 and time 2 scores for the test–retest reproducibility subsample to determine the similarity and reproducibility of the samples.

The data quality, scaling assumptions, acceptability, reliability and validity of the MSIS-29 were evaluated using standard methods.[10,48,82]

Data quality

Data quality[75] was determined to be high if per cent missing data for items was low, item test–retest reproducibility (ICC)[124] was high (≥ 0.50) and per cent computable scale scores was high.

Scaling assumptions

Scaling assumptions were examined by determining whether items in each scale had roughly similar response-option frequency distributions, roughly equivalent mean and variances, and substantial ($r > 0.30$) and roughly equivalent item–total correlations. Items were considered correctly grouped into scales when item–own-scale correlations exceeded item–other-scale correlations by at least 2 SE ($1/\sqrt{n}$)[79] and when the results of factor analysis support hypothesised item groups.

Acceptability

Acceptability was determined by examining score distributions. Ideally, items should have all response categories similarly endorsed,[49] low maximum endorsement frequencies, and low floor and ceiling effects. There are few published criteria for these descriptive statistics. However, it is recommended that item floor and ceiling effects are below 75%[85] and 90%.[86]

Scales were considered acceptable when observed scores were well distributed,[40] mean scores were near the scale midpoint,[87] floor and ceiling effects were less than 20%,[88] and skewness statistics were between –1 and +1.[85]

Reliability

Two types of reliability, internal consistency (Cronbach's α coefficients)[88] and test–retest

reproducibility (intraclass correlation coefficient)[124] were examined. Estimates should exceed 0.80.[10]

Validity

The aim of the validity studies was to examine evidence that the MSIS-29 was a measure of the physical and psychological impact of MS. Three types of validity were examined. Internal validity[100] was determined by examining the intercorrelations between the two MSIS-29 scales. Physical and psychological subscales were expected to be moderately correlated (0.30–0.70) because they are measuring related but different constructs. Convergent and discriminant validity[96] were evaluated by examining the extent to which correlations between MSIS-29 scales and other measures (SF-36, BI, EQ-5D, FAMS, GHQ-12) and variables (age, gender, duration of MS) were consistent with predictions. For example, the MSIS-29 physical scale should correlate highly ($r > 0.70$) with other measures of physical function (e.g. SF-36 physical function, BI, FAMS mobility scale, EQ-5D mobility dimension). See *Table 6* for a more explicit account of the predicted relationships between MSIS-29 and other measures.

Group differences validity was evaluated by examining MSIS-29 scores for groups expected to differ in a predictable way. For example, people who were retired owing to their MS were expected to have higher scores (i.e. poorer health) than people who were employed. People with the most difficulties in mobility and self-care, as categorised by the dimensions of the EQ-5D, were expected to have the highest scores on the physical scale, followed by those with some difficulties, and lastly, by those with no problems. Differences in the psychological scores between these subgroups were not expected to be statistically significant. People who were extremely anxious or depressed as categorised by the anxiety/depression dimension of the EQ-5D were expected to have the highest scores on the psychological scale, followed by those who were moderately anxious or depressed, and lastly, by those who were not anxious or depressed. Differences in the physical scores between these subgroups were not expected to be statistically significant. Men and women were expected to have similar scores, and people with or without a degree were expected to have similar scores. The predictions for group differences validity should be reflected in pairwise F statistics generated by ANOVA (see 'Validity', p. 22).

Table 7 summarises the criterion used to determine adequacy for each psychometric property evaluated.

Results

In total, 1023/1250 (81.8%) questionnaires were returned. Of these, 310 were returned blank (change of address or deceased $n = 63$, did not have MS $n = 155$, did not wish to participate $n = 64$, no reason given $n = 28$). The response rate was 69.1% [response rate = (1023 – 310) (1250 – 63 – 155)] and was similar to the first field test. Analyses were performed on data for 713 people with MS. In the test–retest subsample, 90.6% ($n = 136$) of people who returned the time

TABLE 6 *Expected correlations between MSIS-29 and other measures*

		MSIS scales[a]	
Measure	**Scoring direction**	**MSIS-29 Physical**	**MSIS-29 Psychological**
SF-36 physical	+	−−−	−−
BI	+	−−−	−−
FAMS mobility	+	−−−	−−
EQ-5D mobility	–	+++	++
SF-36 mental health	+	−−	−−−
FAMS emotional well-being	+	−−	−−−
EQ-5D anxiety/depression scale	–	++	+++
GHQ-12	–	++	+++
Age	NA	+ or –	+ or –
Sex	NA	+ or –	+ or –
Duration of MS	NA	+ or –	+ or –

[a] The direction and number of + and – signs reflect the direction and magnitude of correlations: +/–, weak positive/negative correlation ($r < 0.30$); ++/−−, moderate positive/negative correlation ($0.30 < r < 0.70$); +++/−−−, strong positive/negative correlation ($r > 0.70$).

TABLE 7 *Summary of psychometric properties evaluated and the criteria used for determining the adequacy of the MSIS-29*

Psychometric property	Criterion for adequacy
Data quality	Missing item data < 10% High ($r \geq 0.50$) item test–retest reliability High % computable scale scores
Scaling assumptions	Similar response option frequency distribution Skewness –1 to +1 Similar mean scores and variances Similar and substantial ($r > 0.30$) item–total correlations Item–total correlations exceed item–other correlations by at least 2 SE Factor analysis supports hypothesised item groups
Acceptability	
Item acceptability	Approximately equal endorsement across response options Low maximum endorsement frequencies Minimal item floor and ceiling effects
Scale acceptability	Scores span the full scale range Mean scores near midpoint Floor and ceiling effect < 20% Skewness –1 to +1
Reliability	Cronbach's $\alpha > 0.80$ Intraclass correlation coefficient > 0.80
Validity	
Internal validity	Moderate ($r = 0.30$–0.70) intercorrelations between MSIS-29 physical and psychological scales
External validity	
Convergent and discriminant validity	High correlations ($r > 0.70$) between MSIS-29 physical scale and scales measuring physical function High correlations ($r > 0.70$) between MSIS-29 psychological scale and scales measuring psychological function Low correlations between MSIS-29 physical scale and scales measuring psychological function Low correlations between MSIS-29 psychological scale and scales measuring physical function
Group differences validity	Higher physical and psychological scores in retired than employed people Greater differences in physical than psychological scores in people with increasing difficulties in mobility and self-care as defined by EQ5-D Greater differences in psychological scores than their physical scores in people with increasing anxiety or depression as defined by EQ5-D Similar scores between men and women Similar scores between people with and without degrees
Responsiveness	Effect sizes large (> 0.80) to moderate (= 0.50)

1 questionnaire returned the time 2 questionnaire. The characteristics of patients in the second field tests were similar to those of patients in the first field test (*Table 1*). There were no significant differences in demographic characteristics between patients in the three samples.

Data quality (Tables 8 and 9)

Missing data for items were low (range 1.1–3.6%). Eighty-four per cent of respondents endorsed all 29 items (100% complete data), 8.4% of respondents missed out one item and 3.2% of respondents missed out two items. Ninety-seven per cent of respondents had ≥ 90% complete data. Therefore, MSIS-29 scale scores could be computed for 703 respondents (98.6%). In total, 98% of MSIS-29 physical and 98.7% of psychological scale scores were computable. Item test–retest reproducibility was large (*Table 9*). These results indicate that data quality was high.

Scaling assumptions (Tables 8 and 10–12)

Item response option frequency distributions for the items in each scale of the MSIS-29 were

TABLE 8 *MSIS-29: summary of item descriptive statistics in second field test (n = 713)*

MSIS-29 scale	Data completeness		Item descriptive statistics (range)				
	% Missing items	% Computable scale scores	Mean	SD	Floor	Ceiling	Skewness
Physical	1.7–3.6	98.0	2.54–3.83	1.20–1.56	5.9–32.9	14.7–39.5	–0.863 to + 0.405
Psychological	1.1–1.8	98.7	2.64–3.28	1.27–1.40	12.8–29.0	10.7–22.2	–0.294 to + 0.400

TABLE 9 *MSIS-29: test–retest reproducibility in second field test (n =129)*

	Scale scores							Test–retest reproducibility (ICC)	
	Time 1		Time 2		Paired samples t-test			Item	Scale
Scale	Mean	SD	Mean	SD	Mean difference	t-Value	p-Value	Range (mean)	ICC
Physical	55.3	26.9	55.3	27.1	–0.021	–0.028	0.978	0.65–0.90 (0.81)	0.94
Psychological	45.6	25.7	45.3	25.9	0.335	0.293	0.770	0.72–0.82 (0.78)	0.87

roughly symmetrical and not unduly skewed. Items within each scale had similar means scores and standard deviations. All item–own-scale correlations were high (range 0.49–0.86) and exceeded item–other-scale correlations (range 0.33–0.56) by at least 2 SE of a correlation coefficient [$> 2 \times (1/\sqrt{n}) = 0.08$; range 0.12–0.40] (*Tables 11* and *12*). Both scales registered 100% definite scaling success rates, and principal axis factoring of the 29 items, cross-validated in two random split-half samples, generated two factors that were consistent with the hypothesised physical and psychological scales.

Acceptability

Item acceptability (*Table 10***)**

Frequency distributions for item response options were well distributed. Item floor and ceiling effects were low and ranged from 5.9 to 32.9% (item floor effects) and from 10.7 to 39.5% (item ceiling effects). MEF was ≤ 39.5% (response option '5' for item 19). These results indicate that the items satisfy the criteria for acceptability.

Scale acceptability (*Table 13***)**

Scale scores spanned the entire scale range and were in the acceptable range for skewness, mean scores were near the scale midpoint, and floor and ceiling effects were low (maximum 3.9%). These results suggest that the MSIS-29 scales satisfy the criteria for acceptability.

Reliability (Tables 9 and 14)

Internal consistency and test–retest reproducibility exceeded the recommended criterion of 0.80 for reliability in group comparison studies. There were no statistically significant differences in MSIS-29 scores between the three subsamples or between time 1 and time 2 scores for the test–retest reproducibility subsample. These results indicate that the MSIS-29 satisfies the criteria for reliability.

Validity (Tables 15 and 16)

MSIS-29 physical and psychological scores were moderately correlated (0.62), indicating that the two scales measure related but distinct constructs. *Table 15* provides evidence for the convergent and discriminant validity of MSIS-29 scales. The direction, magnitude and pattern of correlations are consistent with predictions. For example, the MSIS-29 physical scale correlates most highly with the FAMS mobility scale, the SF-36 physical functioning scale and the BI, and least with the EQ-5D anxiety/depression dimension, the SF-36 role emotional scale and the FAMS family/social well-being scale. Similarly, the MSIS-29 psychological scale correlates most highly with the SF-36 mental health scale, the FAMS thinking/fatigue scale and the GHQ-12, and least with EQ-5D mobility and self-care dimensions and the BI. In addition, both MSIS-29 scales show low correlations with age, gender and duration of MS, indicating that they are not biased by these variables. Some correlations are not consistent with predictions. For example, the MSIS-29 physical scale correlates more highly than expected with the FAMS emotional well-being scale.

The MSIS-29 confirms hypothesised group differences (*Table 16*). As predicted, mean scores

TABLE 10 *MSIS-29 total sample: item descriptive statistics (n = 713)*

			Item frequency distribution %[a]					Item descriptive statistics[b]		
Item	**Valid**	**MD *n* (%)**	**1**	**2**	**3**	**4**	**5**	**Mean**	**SD**	**Skewness**
01	699	14 (2.0)	5.9	11.2	13.7	32.8	36.5	3.83	1.20	–0.863
02	694	19 (2.7)	21.2	21.9	22.9	19.3	14.7	2.84	1.35	0.126
03	691	22 (3.1)	11.7	14.2	19.8	27.5	26.8	3.43	1.33	–0.447
04	693	20 (2.8)	8.5	15.7	17.0	28.1	30.6	3.57	1.30	–0.526
05	688	25 (3.5)	17.7	18.8	20.3	19.6	23.5	3.12	1.42	–0.100
06	691	22 (3.1)	9.8	25.5	20.4	25.9	18.4	3.18	1.27	–0.081
07	688	25 (3.5)	14.1	19.3	21.2	26.0	19.3	3.17	1.33	–0.183
08	693	20 (2.8)	11.1	16.0	18.2	27.1	27.6	3.44	1.34	–0.426
09	687	26 (3.6)	32.9	20.1	19.7	15.1	12.2	2.54	1.39	0.405
10	690	23 (3.2)	26.2	22.6	18.6	17.7	14.9	2.72	1.41	0.252
11	692	21 (2.9)	12.4	20.4	18.5	24.6	24.1	3.28	1.36	–0.227
12	699	14 (2.0)	17.0	16.0	14.6	23.0	29.3	3.32	1.47	–0.323
13	699	14 (2.0)	15.3	21.2	19.5	22.7	21.3	3.14	1.37	–0.106
14	700	13 (1.8)	19.0	16.4	13.9	22.3	28.4	3.25	1.49	–0.257
15	698	15 (2.1)	20.9	23.5	17.8	20.3	17.5	2.90	1.40	0.104
16	698	15 (2.1)	12.3	15.3	17.8	24.8	29.8	3.44	1.38	–0.429
17	701	12 (1.7)	24.7	16.7	15.8	14.1	28.7	3.05	1.56	–0.027
18	699	14 (2.0)	8.4	16.5	13.4	30.5	31.2	3.60	1.30	–0.582
19	696	17 (2.4)	15.8	13.1	10.5	21.1	39.5	3.55	1.50	–0.575
20	695	18 (2.5)	13.2	15.7	14.7	26.6	29.8	3.44	1.40	–0.452
21	701	12 (1.7)	19.4	27.0	23.5	19.4	10.7	2.75	1.27	0.220
22	700	13 (1.8)	26.7	24.4	19.3	17.3	12.3	2.64	1.36	0.326
23	702	11 (1.5)	12.8	18.4	18.5	28.1	22.2	3.28	1.34	–0.294
24	702	11 (1.5)	20.7	28.2	21.2	16.8	13.1	2.74	1.32	0.298
25	700	13 (1.8)	19.7	25.4	22.3	20.3	12.3	2.80	1.30	0.173
26	703	10 (1.4)	18.1	27.5	22.3	18.8	13.4	2.82	1.30	0.209
27	705	8 (1.1)	15.5	26.4	22.3	22.0	13.9	2.92	1.29	0.096
28	703	10 (1.4)	24.6	21.3	18.2	20.9	14.9	2.80	1.40	0.144
29	701	12 (1.7)	29.0	25.5	16.5	17.1	11.8	2.57	1.37	0.400

[a] Computed for all completed questionnaires.
[b] All cases where scale can be computed.

for people who were retired owing to MS were significantly higher than for those who were still employed. In contrast, mean scores for men and women, and those with or without a degree or professional qualification were not significantly different. Also as predicted, mean MSIS-29 scores for people with increasing problems in mobility, self-care and anxiety/depression, as defined by the EQ-5D, demonstrate a stepwise increase in magnitude and statistically significant *F*-statistics (ratio of between-group to within-group variance). Specifically, mean physical scores for people with the most difficulties in mobility and self-care as measured by EQ-5D were highest, followed by those with some difficulties, and lastly, by those with no problems. Mean psychological scores for people who were extremely anxious or depressed as measured by the EQ-5D were highest, followed by those who were moderately anxious or depressed, and lastly, by those who were not anxious or depressed. The higher the *F*-statistic, the greater the measurement precision,[101] or the extent to which an instrument can detect small differences in the construct being measured.[40]

By comparing different scales (MSIS-29 physical and psychological scales) in the same sample, relative validity can be estimated by the ratio of pairwise *F*-statistics (*F* for one measure divided by *F* for another). Pairwise *F*-statistics indicate that the MSIS physical scale is more valid for detecting group differences in mobility and self-care, while the psychological scale is more valid for detecting group differences in anxiety/depression.

Responsiveness

Overview

Responsiveness was examined in two hospital-based samples: people admitted for rehabilitation and people admitted for intravenous steroids for

TABLE 11 *Item to scale correlations in second field test* (n = *713)*

Item		Item–scale correlations		
		Own scale[a]	Other scale[b]	Difference[c]
	MSIS-29 Physical impact scale			
1	Do physically demanding tasks	0.73	0.41	0.32
2	Grip things tightly (e.g. turning on taps)	0.70	0.43	0.27
3	Carry things	0.79	0.41	0.38
4	Problems with balance	0.73	0.42	0.31
5	Difficulties moving about indoors	0.86	0.47	0.39
6	Being clumsy	0.77	0.53	0.24
7	Stiffness	0.68	0.47	0.21
8	Heavy arms and/or legs	0.70	0.44	0.26
9	Tremor of the arms or legs	0.67	0.49	0.18
10	Spasms in your limbs	0.68	0.44	0.24
11	Your body not doing what you want it to	0.81	0.46	0.35
12	Having to depend on others to do things for you	0.85	0.45	0.40
13	Limitations in your social and leisure activities at home	0.79	0.55	0.24
14	Being stuck at home more than you would like to be	0.79	0.56	0.23
15	Difficulties in using your hands in everyday tasks	0.78	0.51	0.27
16	Having to cut down the amount of time you spent on work or other daily activities	0.77	0.54	0.23
17	Problems using transport (e.g. car, bus, train, taxi)	0.78	0.45	0.33
18	Taking longer to do things	0.84	0.53	0.31
19	Difficulties doing things spontaneously (e.g. going out on the spur of the moment)	0.80	0.50	0.30
20	Needing to go to the toilet urgently	0.60	0.42	0.18
	MSIS-29 Psychological impact scale			
21	Feeling unwell	0.69	0.54	0.15
22	Problems sleeping	0.49	0.33	0.16
23	Feeling mentally fatigued	0.69	0.45	0.24
24	Worries related to your MS	0.67	0.55	0.12
25	Feeling anxious or tense	0.77	0.49	0.28
26	Feeling irritable, impatient or short tempered	0.72	0.40	0.32
27	Problems concentrating	0.68	0.48	0.20
28	Lack of confidence	0.72	0.49	0.23
29	Feeling depressed	0.75	0.45	0.30

[a] For the physical impact scale this is the corrected item–total correlation with the physical impact total score; for the psychological scale this is the corrected item–total correlation with the psychological impact total score.
[b] For the physical impact scale this is the item-total correlation with the psychological impact total score; for the psychological scale this is the item–total correlation with the physical impact total score.
[c] (Item–own-scale correlation) – (item–other-scale correlation), an indicator of extent to which each item discriminates between the two scales.

TABLE 12 *MSIS-29: scaling assumptions in second field test*

	Correlation			Scaling success *n* (%)		Scaling failure *n* (%)	
MSIS-29 scale	**Item–own**	**Item–other**	**Own–other[a]**	**Definite**	**Probable**	**Probable**	**Definite**
Physical	0.60–0.86	0.41–0.57	0.16–0.39	20/20 (100)	0/20 (0)	0/20 (0)	0/20 (0)
Psychological	0.49–0.77	0.34–0.56	0.12–0.31	9/9 (100)	0/9 (0)	0/9 (0)	0/9 (0)

[a] (Item–own-scale correlation) – (item–other-scale correlation).

TABLE 13 *MSIS-29: scale acceptability in second field test*

	Score range[a]			% Floor and ceiling effects		
MSIS-29 scale	**Scale (midpoint)**	**Observed**	**Mean (SD)**	**Floor effects**	**Ceiling effects**	**Skewness**
Physical	0–100 (50)	0–100	56.0 (26.6)	0.9	3.9	–0.285
Psychological	0–100 (50)	0–100	45.5 (25.2)	1.7	1.9	0.172

[a] 0 = best health, 100 = worst health.

TABLE 14 *MSIS-29: internal consistency in second field test*

MSIS-29 scale	Intercorrelations between items Range (mean)	Item–total correlations[a] Range	Cronbach's α
Physical	0.40–0.79 (0.58)	0.60–0.86	0.96
Psychological	0.30–0.70 (0.52)	0.49–0.77	0.91

[a] Corrected for overlap.

TABLE 15 *Convergent and discriminant construct validity of the MSIS-29 in the second field test*

		MSIS-29 scale[a] *r*[b] (*n*)	
Measure	**Scale/dimension**	**Physical**	**Psychological**
SF-36[c]	Physical functioning	–0.79	–0.41
	Role–physical	–0.43	–0.40
	Bodily pain	–0.45	–0.50
	General health perceptions	–0.48	–0.53
	Vitality	–0.49	–0.55
	Social functioning	–0.64	–0.56
	Role–emotional	–0.29	–0.52
	Mental health	–0.41	–0.76
FAMS[d]	Mobility	–0.88	–0.50
	Symptoms	–0.55	–0.64
	Emotional well-being	–0.68	–0.68
	General contentment	–0.64	–0.58
	Thinking and fatigue	–0.56	–0.73
	Family/social well-being	–0.37	–0.50
EQ5D[e]	Mobility	0.61	0.23
	Self-care	0.69	0.37
	Usual activities	0.69	0.42
	Pain/discomfort	0.44	0.43
	Anxiety/depression	0.36	0.68
GHQ-12[f]	Total	0.46	0.68
Postal BI[g]	Total	–0.71	–0.35
Age (n = 678)		0.22	0.03
Sex (n = 686)		0.05	–0.05
Years since diagnosis (n = 629)		0.19	0.03

[a] High scores indicate worse health.
[b] Pearson product–moment correlation coefficients.
[c] High scores indicate better health (n = 263–280).
[d] High scores indicate better health (n = 233–259).
[e] High scores indicate worse health (n = 520–550).
[f] High scores indicate better health (n = 248 and 249, respectively).
[g] High scores indicate better health (n = 260 and 243, respectively).

treatment of their relapse. Results were examined in the total sample. The MSIS-29 was successful at detecting changes in physical and psychological impact in this sample, and was better at doing so than other measures of similar constructs.

Methods

A preliminary responsiveness study was undertaken in consecutive admissions to the NHNN between 1 February and 1 August 2000 for rehabilitation and intravenous steroid treatment.

TABLE 16 *MSIS-29 group differences and relative validity*

Variable	MSIS-29 score	
	Physical	**Psychological**
Employment status		
Employed (n = 107)	30.6 ± 23.1	31.1 ± 22.5
Retired due to MS (n = 390)	64.3 ± 23.0	49.9 ± 24.9
Mean difference (p)	–33.7 (< 0.001)	–18.8 (< 0.001)
EQ-5D mobility dimension		
No problems in walking about (n = 61)	17.5 ± 17.2	27.7 ± 23.1
Some problems in walking about (n = 389)	56.4 ± 21.7	46.6 ± 23.5
Confined to bed (n = 70)	82.6 ± 16.8	51.4 ± 28.1
F (p)[a]	164.3 (< 0.001)	19.3 (< 0.001)
Relative validity[b]	1.0	0.12
EQ-5D self-care dimension		
No problems with self-care (n = 227)	35.5 ± 22.1	34.5 ± 23.0
Some problems with self-care (n = 235)	66.6 ± 16.9	51.5 ± 22.0
Unable to wash or dress myself (n = 76)	85.2 ± 15.3	58.9 ± 27.3
F (p)	256.7 (< 0.001)	46.2 (< 0.001)
Relative validity	1.0	0.18
EQ-5D anxiety/depression dimension		
Not anxious or depressed (n = 229)	45.8 ± 27.4	27.1 ± 17.8
Moderately anxious or depressed (n = 277)	62.2 ± 22.7	55.6 ± 19.3
Extremely anxious or depressed (n = 38)	75.3 ± 22.7	81.6 ± 16.3
F (p)	39.5 (< 0.001)	231.3 (< 0.001)
Relative validity	0.17	1.0
Gender		
Female (n = 489)	55.0 ± 26.9	46.1 ± 25.8
Male (n = 197)	58.1 ± 26.1	43.5 ± 23.6
Mean difference (p)[c]	–3.1 (0.165)	2.6 (0.197)
Degree or professional qualification		
Yes (n = 183)	53.2 ± 26.7	41.6 ± 25.8
No (n = 491)	56.7 ± 26.5	46.8 ± 24.8
Mean difference (p)	–3.5 (0.131)	–5.3 (0.133)

[a] One-way ANOVA with Duncan's post hoc comparisons.
[b] Calculated as the ratio of paired F-values using the largest as the denominator.
[c] Independent samples t-tests, equality of variances not assumed.
Data are shown as mean ± SD.

People were excluded if they appeared to have severe cognitive impairment and were later confirmed to have such an impairment by formal neuropsychological assessment.

People admitted for rehabilitation completed the MSIS-29 on admission and discharge, whereas those admitted for intravenous steroid treatment completed the MSIS-29 on admission and 6 weeks later at their outpatient review. To compare the responsiveness of the MSIS-29 with other measures of the same health construct, all participants were asked to complete a battery of other questionnaires. These included the SF-36, FAMS, GNDS and GHQ-12.

The two samples were pooled and responsiveness was determined by calculating effect sizes (ES),[104] mean change score (admission minus discharge) divided by the standard deviation of admission scores. These are interpreted as small (ES < 0.20), medium (ES = 0.50) or large (ES > 0.80).[125] The statistical significance of the change scores was determined using paired samples t-tests.[106]

Results

Four people recruited to the responsiveness sample were excluded because of cognitive impairment. In total, 55 people completed the questionnaires at both time 1 and time 2. *Table 17* describes the characteristics of the pooled

TABLE 17 *Characteristics of responsiveness sample (n = 55)*

Characteristic	Value
Gender: % female	66
Age (years): mean ± SD (range)	45.0 ± 13.0 (23–83)
Years since MS onset: mean (SD); range	16 ± 12 (1–60)
Marital status (%)	
Married	64
Percent living with others	82
Employment status (%)	
Retired due to MS	31
Employed 44	
Type of MS (%)	
Primary progressive	5.5
Secondary progressive	47.3
Relapsing–remitting	47.3
Mobility indoors (%)	
Walk unaided	24
Walk with an aid	49
Wheelchair dependent	27

responsiveness sample. Although this sample is small, its characteristics are similar to the larger field test shown in *Table 1*.

Scores for MSIS-29 physical and psychological scales were lower at time 2 than at time 1 (*Table 18*), indicating improvement following treatment by inpatient rehabilitation or intravenous steroid treatment. Change scores for both scales were similar in magnitude and statistically significant. Effect sizes were large to moderate.

Compared with questionnaires measuring similar constructs, the MSIS-29 physical scale is the most responsive. Although the GHQ-12 is the most responsive, the MSIS-29 psychological scale performs relatively well compared with other measures of psychological impact.

TABLE 18 *Preliminary responsiveness of MSIS-29 and other measures*

Instrument	Effect size[a]	Relative responsiveness of MSIS
Measures of physical function		
MSIS-29 physical impact scale	0.82	1.0
SF-36 physical function dimension	0.46	0.56
FAMS mobility scale	0.60	0.73
GNDS	0.53	0.65
EDSS	0.42	0.51
Measures of psychological function		
MSIS-29 psychological impact scale	0.66	1.0
SF-36 mental health dimension	0.36	0.54
FAMS emotional health dimension	0.44	0.66
GHQ	0.76	1.15
MSIS-29	**Physical scale**	**Psychological scale**
Time 1: mean ± SD	64.4 ± 23.0	48.4 ± 26.7
Time 2: mean ± SD	45.6 ± 23.4	30.7 ± 22.3
Change score: mean ± SD (*p*-value)	18.8 ± 19.6 ($p < 0.001$)	17.7 ± 24.6 ($p < 0.001$)

[a] Mean change score divided by SD of admission scores.

Chapter 5
Discussion

Overview

The aim of this study was to develop an MS-specific outcome measure that combines the patient perspective with a rigorous scientific approach. This aim was addressed by generating items from in-depth patient interviews, using the self-report method of administration, selecting items on the basis of psychometric performance in a large field test and applying rigorous psychometric methods. Extensive field testing in the two independent samples confirmed that the MSIS-29 satisfies criteria as a summed rating scale and is an acceptable, reliable and valid measure of the physical and psychological impact of MS. Furthermore, there is preliminary evidence that the MSIS-29 is responsive to change. This chapter presents a discussion of study results, study limitations and the implications of the MSIS-29 for healthcare, and provides recommendations for future research.

Discussion of results

Results from this study confirm that the MSIS-29 satisfies criteria as a summed rating scale and is an acceptable, reliable and valid measure of the physical and psychological impact of MS. Stringent criteria for item selection were adopted in an attempt to develop an instrument with strong psychometric properties. To create a responsive scale, items were selected that discriminated well between individuals, and items with maximum endorsement frequencies over 40% were eliminated. Similarly, to reduce overlap between the MSIS-29 physical and psychological scales, items that did not show good evidence of item convergent and discriminant validity were eliminated. Such a rigorous approach to the development and validation of health outcome measures is important because the results of studies are dependent on the quality of the measures used for data collection. Furthermore, the limitations of measures cannot be overcome easily by improvements in study design and powerful statistical methods.[126]

The terms quality of life, health-related quality of life, health status, and disability are often used interchangeably or without specific reference to what they measure. Measures that are intended to assess such concepts are collectively referred to as 'patient-based outcome measures'.[127] The present researchers specifically chose not to describe the MSIS-29 as a measure of health-related quality of life, health status or disablement, as these terms are ambiguous and can mislead investigators when they are selecting measures for clinical trials.

There were some unexpected results. Only two distinct dimensions of health, physical and psychological impact, appear to underlie the diverse 129-item pool. Although many other dimensions of health, such as symptoms, were included in the initial version of the questionnaire, psychometric analyses did not support these multiple dimensions. These results support previous findings[28,128] that have suggested that a two-dimensional model, consisting of physical and psychological health, explains the construct of subjective health status. Another unexpected result is that the MSIS physical and psychological scales are correlated to a similar degree with the FAMS emotional well-being scale. However, correlations between the MSIS physical scale and other measures of psychological distress (GHQ-12, SF-36 mental health dimension and EQ-5D anxiety/depression dimension) are low to moderate. These findings require further investigation of the relationship between MSIS physical and psychological scales and other MS-specific scales.

Preliminary evidence of the responsiveness of the MSIS-29 was obtained in a hospital-based sample. The MSIS-29 physical scale showed particularly good responsiveness, as indicated by a large effect size, in comparison with other measures of physical function that showed much smaller effect sizes. The MSIS-29 psychological scale also showed good responsiveness compared with other measures of psychological function. Only the GHQ-12 showed better responsiveness than the MSIS-29 psychological scale. In contrast to a clinician-rated scale such as the EDSS, the MSIS-29 is sensitive to changes that patients themselves have defined as important. As these results are preliminary, owing to the small sample size, further evaluations of responsiveness must be undertaken in different samples and settings to confirm the responsiveness of the MSIS-29 compared with other measures.

Study limitations

One potential limitation in this study is the use of the MS Society membership database to define the sampling frame. It is known that many members of the MS Society are partners, friends or relatives of people with MS. Therefore, the researchers specifically asked people who did not have MS to tick a box on the front of the questionnaire and to return it blank. The results suggest that a minimum of 56% of members are people with MS. However, the percentage of people in the database with a neurologist-confirmed diagnosis of clinically definite MS, the disease type of those with MS and the representativeness of people who join charitable groups are unknown. The estimates indicate, however, that a random sample was from approximately 35% (28,000) of the total UK population of people with MS.

Implications for health care

The MSIS-29 was purposely developed to be short and simple enough for routine use in a wide range of healthcare applications. It offers the opportunity to measure rigorously the impact of MS and evaluate treatment effectiveness from the patient's perspective. Recently, several disease-modifying treatments, of which interferon-β is the most widely publicised, have become available for the treatment of MS. The MSIS-29 provides a scientifically rigorous method of evaluating the effectiveness of new interventions in relation to current treatments. There is clear consensus about the need for outcome measures to evaluate models of care.[129–132] However, current evaluations of interferons for MS highlight some of the difficulties in measuring outcomes in this chronic progressive, incurable and unpredictable disease. First, the outcomes that are evaluated may be of limited relevance to patients' day-to-day lives.[133,134] Second, it is essential to show the relative efficacy of new and existing therapies to assess the proportional benefit of different interventions.[135] Third, evaluations of treatment effectiveness are often based on results of studies of a small number of patients that use poor quality measurement instruments.[136]

The MSIS-29 is the first patient-based measure for MS that truly incorporates patients' views, as it was developed by reducing an item pool generated *de novo* from people with MS. To create a responsive scale, which is one of the most important attributes of an instrument that is to be used in clinical trials, items were selected that discriminated well between individuals, and items with MEF over 40% were eliminated. The preliminary evidence suggests that MSIS-29 shows good responsiveness.

The MSIS-29 can be used in cross-sectional studies to describe the impact of MS from the patient's perspective, in longitudinal studies to monitor the natural history of MS and, most importantly, in clinical trials to evaluate therapeutic effectiveness from the patient's perspective. Furthermore, the availability of reliable, valid and responsive patient-based outcome measures is central to an improved understanding of the impact of MS and its relationships with other indicators of disease activity, such as neuroimaging and neurophysiology.

The MSIS-29 has been developed for use in both clinical trials and evidence-based clinical practice, to monitor the progress of people with MS. Therefore, it is highly relevant to the NHS. In light of the disease-modifying drugs that will be available, MSIS-29 provides a further valuation of the effectiveness of these therapies, as well as of other treatments such as neurorehabilitation. It can be used in routine data collection and clinical governance, which is also relevant to the NHS.

Recommendations for future research

There are several recommendations for future research. These are now discussed in order of priority. First, further evaluations of the MSIS-29 are needed as the psychometric properties of health outcome measures are sample dependent and cannot be established in a single study.[137] Evaluations of the performance of the MSIS-29 with different patient samples and in different settings will help to clarify its strengths and weaknesses and further define its role in clinical practice and research. Second, head-to-head comparisons of the full spectrum of psychometric properties of the MSIS-29 and existing MS-specific outcome measures such as the MS Functional Composite,[138] GNDS[26] and Leeds MSQoL scale[25] should be undertaken. This will determine the advantages and disadvantages of different instruments and how they complement each other. Most importantly, such head-to-head comparisons will provide an evidence-based framework to guide investigators in the selection of outcome measures for research and audit. Third, as traditional psychometric methods were used to develop and evaluate the MSIS-29, it is also important that newer psychometric methods such as Rasch item analyses[139] and Item Response Theory models[67] are used to evaluate the MSIS-29. Fourth, the specificity of the MSIS-29 to MS and applicability to other neurological conditions should be tested. The MSIS-29 was developed from in-depth

interviews with people with MS, so it is most suitable for use with people with MS. However, as with all such tools, it may be applicable for people with other disabling neurological conditions, and this may be another area for future research. Fifth, although steps were taken to ensure that the pretesting sample was sufficiently large and representative of the general MS population, a larger follow-up interview survey would have been useful. Thus, further interviews should be conducted to obtain feedback, especially regarding questions that are deemed irrelevant by subgroups (e.g. ethnic minority groups and older people). Finally, the assessment of the MSIS-29 by other investigators in different cohorts of patients is also a recommendation for future research.

Interpreting scores

Despite the widespread use of health outcome measures, there is no systematic strategy for translating scores generated by such instruments into clinical decisions. It is standard practice with any standardised measure, including those used in clinical laboratory testing, academic achievement, personnel selection and psychological testing, to interpret scores in relation to normative values for the population. These are generally expressed as standard deviation units, percentiles or percentages, thus enabling comparisons across samples, constructs and measures. However, as such norm-based interpretations are unfamiliar to clinicians and patients, and may have limited clinical meaning, content-based referencing is the preferred method for the interpretation of scores.[30,140] Using this method of interpreting scores, changes on health measures are anchored to clinically relevant change.[141] Although clinical interpretation of scores is relatively straightforward for single-item measures such as the EDSS, where each score has a specific clinical meaning, it is less clear for multi-item measures in which an overall score may represent varying combinations of item scores. Determining the clinical significance of changes even on single-item measures is not simple. For example, the recently published randomised placebo-controlled study of interferon-β in secondary progressive MS[142] demonstrated significantly lower disability in the treatment group compared with placebo, as measured by a difference of 0.13 EDSS points. Although this difference was shown to be statistically significant, its clinical significance is unknown. The relationship between statistical significance and clinical meaning is poorly studied.[143]

Some authors have suggested a simplistic preliminary method of interpreting scores for Likert scales.[144] Based on this, a simple method of interpreting MSIS-29 scores would be to categorise scores of 0–19 as ' no problems', 20–39 as 'few problems', 40–59 as 'moderate problems', 60–79 as 'quite a few problems' and 80–100 as 'extreme problems'. Although simplistic, age- and disease-related norms are needed for comparisons, the accumulation of MSIS-29 data in future studies will enable population- and content-based norms to be established.[30] In the mean time, MSIS-29 scores can be compared with mean scores from the random sample of people from the MS Society. Scores from groups of individuals with MS can be compared with the mean scores of people from the MS Society to identify how much that group's score differs from the MS Society sample.

Several other methods have been proposed for the clinical interpretation of scores on health measures. These include relating scores or score changes to the cost of healthcare utilisation,[142] major life events,[146] preference weightings,[147] equivalence with the impact of other diseases[148] or visual representations (mapping) of the relationships between perceptions and behaviours.[149] All of these methods have their limitations[141] (e.g. major life events are uncommon and their impact is variable), prompting Deyo and colleagues[150] to recommend the use of a limited number of measures, and Lydick and Yawn[151] to add that the continued collection of data concerning clinical anchors will enable clinicians, over time, to become increasingly familiar with the clinical significance of particular levels of change. The latter again highlights the need for continuous collection of MSIS-29 data to determine the clinical significance of change scores.

Conclusions

Rigorous measurement of health outcomes underpins research and clinical practice. In chronic progressive disorders such as MS, it is essential that outcome measures incorporate the patient's perspective. The 29-item MSIS-29 is a rigorous new measure of the physical and psychological impact of MS, which is evaluated from the perspective of patients themselves. Comprehensive evaluation of the psychometric properties of the MSIS-29 demonstrated that data quality was high and scaling assumptions were satisfied. Acceptability, reliability and validity were supported, and preliminary evidence of responsiveness was demonstrated. The MSIS-29 is appropriate for use in clinical trials to evaluate therapeutic effectiveness from the patient's perspective.

Acknowledgements

We would like to thank the people with multiple sclerosis who participated in this study, the Multiple Sclerosis Society of Great Britain and Northern Ireland for their support, Mr Peter Cardy and Dr Iain Smith (Advisory Committee), Ms Irene Richardson for undertaking the patient interviews and research assistance during the first field test, Dr Sarah Smith for assistance with the psychometric analyses during the item reduction process, Dr Sara Schroter and Dr Stefan Cano for their contribution to the item reduction process, and Ms Laura Camfield for research assistance during the second field test.

This study was funded by the NHS Health Technology Assessment Programme, but the views and opinions expressed do not necessarily reflect those of the Department of Health.

We would like to thank the following for kindly giving their permission to reproduce data from the published original articles: Arnold Journals for extracts from Riazi A, Hobart JC, Lamping DL, Fitzpatrick R, Thompson AJ. Evidence-based measurement in multiple sclerosis: the psychometric properties of the physical and psychological dimensions of three quality of life rating scales. *Mult Scler* 2003;**9**:411–19. BMJ Publishing Group for extracts from Riazi A, Hobart JC, Lamping DL, Fitzpatrick R, Thompson AJ. Multiple Sclerosis Impact Scale (MSIS-29): reliability and validity in hospital based samples. *J Neurol Neurosurg Psychiatry* 2002;**73**:701–4. Oxford University Press for extracts from Hobart JC, Lamping DL, Fitzpatrick R, Riazi A, Thompson AJ. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. *Brain* 2001;**124**:962–73.

References

1. Holmes J, Madgwick T, Bates D. The cost of multiple sclerosis. *Br J Med Econ* 1995;**8**:181–93.
2. Hatch J. The economic impact of multiple sclerosis. *MS Management* 1996;**3**(1):40.
3. Harvey C. Economic costs of multiple sclerosis: how much and who pays? Health Services Research Report No. ER-6005. New York: National Multiple Sclerosis Society; January 1995.
4. Prouse P, Ross-Smith K, Brill M, Singh M, Brennan P, Frank A. Community support for young physically handicapped people. *Health Trends* 1991;**23**:105–9.
5. Thompson AJ, Noseworthy JH. New treatments for multiple sclerosis: a clinical perspective. *Curr Opin Neurol* 1996;**9**:187–98.
6. Hobart JC, Thompson AJ. Clinical trials of multiple sclerosis. In Reder AT, editor. *Interferon therapy of multiple sclerosis*. New York: Marcel Dekker; 1996. pp. 398–407.
7. Guyatt GH, Freeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;**118**:622–9.
8. Jenkinson C, Peto V, Fitzpatrick R, Greenhall R, Hyman N. Self-reported functioning and well-being in patients with Parkinson's disease: comparison of the Short-Form Health Survey (SF-36) and the Parkinson's Disease Questionnaire (PDQ-39). *Age Ageing* 1995;**24**:505–9.
9. Kurtzke JF. Rating neurological impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983;**33**:1444–52.
10. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.
11. Sharrack B, Hughes RAC, Soudain S, Dunn G. The psychometric properties of clinical rating scales used in multiple sclerosis. *Brain* 1999;**122**:141–59.
12. Hobart JC, Freeman JA, Thompson AJ. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. *Brain* 2000; **123**:1027–40.
13. Ware JE Jr, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey manual and interpretation guide*. Boston, MA: Nimrod Press; 1993.
14. Bergner M, Bobbitt RA, Pollard WE, Martin DP, Gilson BS. The Sickness Impact Profile: validation of a health status measure. *Med Care* 1976;**14**: 57–67.
15. EuroQoL Group. EuroQoL: a new facility for the measurement of health-related quality of life. *Health Policy* 1990;**16**:199–208.
16. Peto V, Jenkinson C, Fitzpatrick R, Greenhall R. The development and validation of a short measure of functioning and well-being for individuals with Parkinson's disease. *Qual Life Res* 1995;**4**:241–8.
17. Patrick D, Deyo R. Generic and disease-specific measures in assessing health status and quality of life. *Med Care* 1989;**27**(3 Suppl):S217–32.
18. Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Clinical appropriateness: a key factor in outcome measure selection. The 36-item Short Form Health Survey in multiple sclerosis. *J Neurol, Neurosurg Psychiatry* 2000;**68**:150–6.
19. Norvedt MW, Riise T, Myer K-M, Nyland HI. Performance of the SF-36, SF-12 and RAND-36 summary scales in a multiple sclerosis population. *Med Care* 2000;**38**:1022–8.
20. Hobart JC, Freeman JA, Lamping DL, Fitzpatrick R, Thompson AJ. The SF-36 in multiple sclerosis: why assumptions must be tested. *J Neurol Neurosurg Psychiatry* 2001;**71**:363–70.
21. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Importance of sensitivity to change as a criterion for selecting health status measures. *Qual Health Care* 1992;**1**:89–93.
22. Cella DF, Dineen K, Arnason B, Reder A, Webster KA, Karabatsos G, *et al*. Validation of the Functional Assessment of Multiple Sclerosis quality of life instrument. *Neurology* 1996;**47**:129–39.
23. Vickrey BG, Hays RD, Harooni R, Myers LW, Ellison GW. A health-related quality of life measure for multiple sclerosis. *Qual Life Res* 1995;**4**: 187–206.
24. Rudick R, Antel J, Confavreux C, Cutter G, Ellison G, Fischer J, *et al*. Recommendations from the National Multiple Sclerosis Society Clinical Outcomes Assessment Task Force. *Ann Neurol* 1997;**42**:379–82.
25. Ford HL, Tennant A, Johnson MH. Developing a disease-specific quality of life measure for people with multiple sclerosis. *Clin Rehabil* 2001;**15**: 247–58.
26. Sharrack B, Hughes RAC. The Guy's Neurological Disability Scale (GNDS): a new disability measure for multiple sclerosis. *Mult Scler* 1999;**5**:223–33.

27. Fischer JS, Rocca NL, Miller DM, Ritvo PG, Andrews H, Paty D. Recent developments in the assessment of quality of life in multiple sclerosis. *Mult Scler* 1999;**5**:251–9.

28. Pfennings LEMA, Cohen L, Van der Ploeg HM, Bramsen I, Polman CH, Lankhorst GJ, *et al*. A health-related quality of life questionnaire for multiple sclerosis patients. *Acta Neurol Scand* 1999; **100**:148–55.

29. Freeman JA, Hobart JC, Thompson AJ. Does adding MS specific items to a generic measure (SF-36) improve measurement. *Neurology* 2001; **57**:68–74.

30. McDowell I, Jenkinson C. Development standards for health measures. *Journal of Health Services Research and Policy* 1996;**1**:238–46.

31. Quality of life and clinical trials [editorial]. *Lancet* 1995;**346**:1–2.

32. McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*. 2nd ed. Oxford: Oxford University Press; 1996.

33. Ware JE Jr. The status of health assessment 1994. *Annu Rev Public Health* 1995;**16**:327–54.

34. Gothan A, Brown R, Marsden C. Depression in Parkinson's disease: a quantitative and qualitative analysis. *J Neurol Neurosurg Psychiatry* 1986; **49**:381–9.

35. Brown R, MacCarthy B, Jahanshahi M, Marsden C. Accuracy of self-reported disability in patients with parkinsonism. *Arch Neurol* 1989;**46**:955–9.

36. Hays R, Vickery B, Hermann B, Perrine K, Cramer J, Meador K, *et al*. Agreement between proxy reports and self-reports of quality of life in epilepsy patients. *Qual Life Res* 1995;**4**:159–65.

37. Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *J Clin Epidemiol* 1992;**45**:743–60.

38. Brosseau L. The inter-rater reliability and construct validity of the Functional Independence Measure for multiple sclerosis subjects. *Clin Rehabil* 1994;**8**:107–15.

39. Vickrey BG, Hays RD, Engel J, Spritzer K, Rogers WH, Rausch R, *et al*. Outcome assessment for epilepsy surgery: the impact of measuring health-related quality of life. *Ann Neurol* 1995;**37**:158–66.

40. Stewart AL, Ware JE Jr, editors. *Measuring functioning and well-being: the Medical Outcomes Study approach*. Durham, NC: Duke University Press; 1992.

41. Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care. I: Applications and uses in assessment. *BMJ* 1992;**305**:1074–7.

42. Ware JE Jr. Measuring patients' views: the optimum outcome measure. *BMJ* 1993; **306**:1429–30.

43. Hobart JC, Freeman JA, Lamping DL. Physician and patient oriented outcomes in chronic and progressive neurological disease: which to measure? *Curr Opin Neurol* 1996;**9**:441–4.

44. Ware JE Jr. Standards for validating health measures: definition and content. *Journal of Chronic Diseases* 1987;**40**:473–80.

45. Bergner M. Quality of life, health status, and clinical research. *Med Care* 1989;**27**(3 Suppl):S148–56.

46. Ware JE Jr, Brook RH, Davies-Avery A, Williams KN, Stewart AL, Rogers WH, *et al*. Conceptualization and measurement of health for adults in the health insurance study. Vol. I. Model of health and methodology. Report No. R-1987/1-HEW. Santa Monica, CA: Rand Corporation; May 1980.

47. McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*. Oxford: Oxford University Press; 1987.

48. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press; 1989.

49. Guilford JP. *Psychometric methods*. 2nd ed. New York: McGraw-Hill; 1954.

50. Nunnally JC Jr. Tests and measurements: assessment and prediction. New York: McGraw-Hill; 1959.

51. Rogers T. *The psychological testing enterprise: an introduction*. Pacific Grove, CA: Brooks/Cole; 1995.

52. Nunnally JC. *Introduction to psychological measurement*. New York: McGraw-Hill; 1970.

53. Torgerson WS. *Theory and methods of scaling*. New York: Wiley; 1958.

54. Thurstone LL. A method for scaling psychological and educational tests. *J Educ Psychol* 1925; **16**:433–51.

55. Thurstone LL. Attitudes can be measured. *Am J Sociol* 1928;**33**:529–54.

56. Nunnally JC. *Psychometric theory*. New York: McGraw-Hill; 1967.

57. American Psychological Association, Committee on Test Standards. Technical recommendations for psychological tests and diagnostic techniques. *Psychol Bull* 1954;**51**(2 Suppl):201–38.

58. American Educational Research Association, National Council on Measurement in Education. *Technical recommendations for achievement tests*. Washington, DC: National Education Association; 1955.

59. American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association; 1966.

60. American Psychological Association, American Educational Research Association and National Council on Measurement in Education. *Standards for education and psychological tests*. Washington, DC: American Psychological Association; 1974.

61. American Educational Research Association, American Psychological Association, Education and National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington, DC: American Psychological Association; 1985.

62. Brook RH, Ware JE Jr, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, *et al*. Overview of adult health status measures fielded in Rand's Health Insurance Study. *Med Care* 1979;**17**(7 Suppl):1–131.

63. Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, *et al*. Functional status and well-being of patients with chronic conditions. Results from the medical outcomes study. *JAMA* 1989;**262**:907–13.

64. McHorney CA, Ware JE Jr, Rogers W, Raczek AE, Lu JFR. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. *Med Care* 1992; **30**(5):MS253–65.

65. Testa MA, Simonson DC. Assessment of quality-of-life outcomes. *N Engl J Med* 1996;**334**:835–40.

66. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 2nd ed. Oxford: Oxford University Press; 1995.

67. Lord FM, Novick MR. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley; 1968.

68. Brown FG. *Principles of educational and psychological testing*. Hinsdale, IL: Dryden Press; 1970.

69. Allen MJ, Yen WM. *Introduction to measurement theory*. Monterey, CA: Brooks/Cole; 1979.

70. Carmines EG, Zeller RA. *Reliability and validity assessment*. Newbury Park, CA: Sage; 1979.

71. Anastasi A. *Psychological testing*. 6th ed. Upper Saddle River, NJ: Prentice-Hall; 1988.

72. Kaplan RM, Saccuzzo DP. *Psychological testing: principles, applications, and issues*. 3rd ed. Pacific Grove, CA: Brooks/Cole; 1993.

73. Waksberg J, Ferber R, Sheatsley P, Turner A. *What is a survey? How to conduct pretesting*. Alexandria, VA: American Statistical Association; 1995.

74. Green, SB, Lissitz RW, Muliak SA. Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement* 1977;**37**:827–38.

75. McHorney CA, Ware JE Jr, Lu JFR, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. *Med Care* 1994;**32**:40–66.

76. Ware JE Jr, Davies-Avery A, Brook RH. Conceptualization and measurement of health for adults in the health insurance study. Vol. VI. Analysis of relationships among health status measures. Report No. R-1987/6-HEW. Santa Monica, CA: Rand Corporation; November 1980.

77. Likert RA. A technique for the development of attitudes. *Arch Psychol* 1932;**140**:5–55.

78. Howard KI, Forehand GC. A method for correcting item-total correlations for the effect of relevant item inclusion. *Educational and Psychological Measurement* 1962;**22**:731–5.

79. Ware JE Jr, Harris WJ, Gandek B, Rogers BW, Reese PR. MAP-R for Windows: multitrait/multi-item analysis program – revised user's guide. Boston, MA: Health Assessment Laboratory; 1997.

80. DeVellis RF. *Scale development: theory and applications*. London: Sage; 1991.

81. Ware JE Jr, Davies-Avery A, Donald CA. Conceptualization and measurement of health for adults in the health insurance study. Vol. V. General health perceptions. Report No.: R-1987/5-HEW. Santa Monica, CA: Rand Corporation; September 1978.

82. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, *et al*. Evaluating quality of life and health status instruments: development of scientific review criteria. *Clin Ther* 1996;**18**:979–92.

83. Hays, RD, Hayashi T. Beyond internal consistency reliability: rationale and user's guide for Multi-Trait Analysis Program on the microcomputer. *Behav Res Methods Instrum Comput* 1990;**22**:167–75.

84. Nunnally JC. *Psychometric theory*. 2nd ed. New York: McGraw-Hill; 1978.

85 Holmes WC, Bix B, Shea JA. SF-20 score and item distributions in a human immunodeficiency virus-seropositive sample. *Med Care* 1996;**34**:562–9.

86. Leplege A, Rude N, Ecosse E, Ceinos R, Dohin E, Pouchot J. Measuring quality of life from the point of view of HIV-positive subjects: the HIV-QL31. *Qual Life Res* 1997;**6**:585–94.

87. Eisen M, Ware JE Jr, Donald CA, Brook RH. Measuring components of children's health status. *Med Care* 1979;**17**:902–21.

88. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health

status surveys adequate? *Qual Life Res* 1995; **4**:293–307.

89. Holmes WC, Shea JA. Performance of a new, HIV/AIDS-targeted quality of life (HAT-QoL) instrument in asymptomatic seropositive individuals. *Qual Life Res* 1997;**6**:561–71.
90. Fiske DW. Some hypotheses concerning test adequacy. *Educational and Psychological Measurement* 1966;**26**:69–88.
91. Tyler TA, Fiske DW. Homogeneity indices and test length. *Educational and Psychological Measurement* 1968;**28**:767–77.
92. Cortina JM. What is coefficient alpha? An examination of theory and applications. *J Appl Psychol* 1993;**78**:98–104.
93. Spearman CE. The proof and measurement of association between two things. *Am J Psychol* 1904;**15**:72–101.
94. Bland JM, Altman DG. Statistical methods for assessing the agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.
95. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; **86**:420–8.
96. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.
97. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychol Bull* 1959;**56**:81–105.
98. Campbell DT. Recommendations for APA test standards regarding construct, trait, or discriminant validity. *Am Psychol* 1960;**15**:546–53.
99. Patrick DL, Erickson P. *Health status and health policy*. Oxford: Oxford University Press; 1993.
100. Bohrnstedt GW. Measurement. In: Rossi PH, Wright JD, Anderson AB, editors. *Handbook of survey research*. New York: Academic Press; 1983. pp. 69–121.
101. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993;**31**:247–63.
102. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 1985;**28**:542–7.
103. Norman GR. Issues in the use of change scores in randomized trials. *J Clin Epidemiol* 1989; **42**:1097–105.
104. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;**27**(3 Suppl):S178–89.
105. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;**28**:632–8.
106. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Control Clin Trials* 1991;**12**:142–58.
107. Hobart JC, Freeman JA, Greenwood RJ, McLellan DL, Thompson AJ. Responsiveness of outcome measures: beware the effect of different effect sizes. *Ann Neurol* 1998;**44**:519A.
108. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol* 1997; **50**:869–79.
109. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Transition questions to assess outcomes in rheumatoid arthritis. *Br J Rheumatol* 1993; **32**:807–11.
110. Mackenzie CR, Charleston ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. *Arch Intern Med* 1986;**146**:1325–9.
111. Juniper E, Guyatt G, Goldstein R. Determining a minimal important change in a disease-specific quality of life instrument. *J Clin Epidemiol* 1994; **47**:81–7.
112. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases* 1987;**40**:171–8.
113. Hobart JC, Lamping DL, Fitzpatrick R, Riazi A, Thompson AJ. The Multiple Sclerosis Impact Scale (MSIS-29): a rigorous patient-based measure of the impact of MS [abstract]. *Ann Neurol* 2000; **48**:448.
114. Dillman D. *Mail and telephone surveys: the total design method*. New York: Wiley; 1978.
115. The WHOQOL Group. The World Health Organisation Quality of Life Assessment (WHOQOL): development and general psychometric properties. *Soc Sci Med* 1998; **46**:1569–85.
116. Juniper EF, Guyatt GH, Streiner DL, King DR. Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epidemiol* 1997;**50**:233–8.
117. Duruoz MT, Poiraudeau S, Fermanian J, Menkes C-J, Amor B, Dougados M, *et al*. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. *J Rheumatol* 1996;**23**:1167–72.
118. Ferguson E, Cox T. Exploratory factor analysis: a user's guide. *International Journal of Selection and Assessment* 1993;**1**:84–94.

119. Guttman LA. Some necessary conditions for common-factor analysis. *Psychometrika* 1954; **19**:149–61.

120. Cattell RB. The scree test for the number of factors. *Multivariate Behavioural Research* 1966; **1**:245–76.

121. Guertin WH, Bailey JP Jr. *Introduction to modern factor analysis*. Ann Arbor, MI: Edwards Brothers; 1970.

122. Gompertz P, Pound P, Ebrahim S. A postal version of the Barthel Index. *Clin Rehabil* 1994;**8**:233–9.

123. Goldberg DP, Hillier VF. A scaled version of the General Health Questionnaire. *Psychol Med* 1979;**9**:139–45.

124. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966;**19**:3–11.

125. Cohen J. *Statistical power analysis for the behavioural sciences*. Hillside, NJ: Lawrence Erlbaum; 1969.

126. Fleiss JL. *The design and analysis of clinical experiments*. New York: Wiley; 1986.

127. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for clinical trials. *Health Technol Assess* 1998;**2**(14).

128. Ware JE Jr, Kosinski MA, Keller SD. SF-36 physical and mental health summary scales: a user's manual. Boston, MA: Health Institute, New England Medical Centre; 1994.

129. Hopkins A, Costain D, editors. *Measuring the outcomes of medical care*. London: Royal College of Physicians of London; 1990.

130. Frater A, Costain D. Any better? Outcome measures in medical audit. *BMJ* 1992;**304**:519–20.

131. Delmothe A, editor. *Outcomes into clinical practice*. London: BMJ Publishing; 1994.

132. Jenkinson C, editor. *Measuring health and medical outcomes*. London: University College London Press; 1994.

133. Harvey P. Why interferon beta-1b was licensed is a mystery. *BMJ* 1996;**313**:297–8.

134. Rous E, Coppel A, Haworth J, Noyce S. A purchaser experience of managing new expensive drugs: interferon beta. *BMJ* 1996;**313**:1195–6.

135. Ferner RE. Newly licensed drugs. *BMJ* 1996; **313**:1157–8.

136. Richards RG. Interferon beta in multiple sclerosis. *BMJ* 1996;**313**:1159.

137. Stewart AL, Hays RD, Ware JE Jr. The MOS Short-Form General Health Survey: reliability and validity in a patient population. *Med Care* 1988; **26**:724–35.

138. Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J, *et al*. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain* 1999;**122**:871–82.

139. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press; 1960.

140. Ware JE Jr. Content-based interpretation of health status scores. *Medical Outcomes Trust Bulletin* 1994; **2**(4):3.

141. Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res* 1993;**2**:221–6.

142. European Study Group on Interferon beta-1b in Secondary Progressive MS. Placebo-controlled multicentre randomised trial of interferon beta-1b in treatment of secondary progressive multiple sclerosis. *Lancet* 1998;**352**:1491–7.

143. Cohen J. *Statistical power analysis for the behavioural sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.

144. Jenkinson C, Fitzpatrick R, Swash M, Levvy G. ALSAQ User Manual: Amyotrophic Lateral Sclerosis Assessment Questionnaire. Oxford: Health Services Research Unit, University of Oxford; 2001.

145. Ware JE Jr, Manning WGJ, Duan N, Wells KB, Newhouse JP. Health status and the use of outpatient mental health services. *Am Psychol* 1984;**39**:1090–100.

146. Testa MA, Anderson RB, Nackley JF, Hollenberg NK, Group at QoLHS. Quality of life and hypertensive therapy in men: a comparison of captopril with enalapril. *N Engl J Med* 1993; **328**:907–13.

147. Hadorn DC, Uerbersax J. Large scale health outcomes evaluation: how should quality of life be measured? Part 1 – Calibration of a brief questionnaire and a search for preference subgroups. *J Clin Epidemiol* 1995;**48**:607–18.

148. Brook RH, Ware JE Jr, Rogers WH, Keeler EB, Davies AR, Donald CA. Does free care improve adult health? Results from a randomised controlled trial. *N Engl J Med* 1983;**309**:1426–34.

149. Cornwall A. Body mapping in health. *Rapid Rural Appraisal Series* 1991;**12**:69–76.

150. Deyo RA, Andersson G, Bombardier C, Cherkin DC, Keller RB, Lee CK. Outcome measures for studying patients with low back pain. *Spine* 1994; **18** (Suppl):2032S–6S.

151. Lydick E, Yawn BP. Clinical interpretation of health-related quality of life data. In: Staquet MJ, Hays RD, Fayers PM, editors. *Quality of life assessment in clinical trials: methods and practice*. Oxford: Oxford University Press; 1998. pp. 299–314.

Appendix 1

The 129-item long-form questionnaire and item reduction strategy

Stage 1

Bold items = 36 items eliminated on the basis of redundancy
Italic items = 51 items eliminated on the basis of psychometric performance

1. Numbness or loss of sensation?
2. Pins and needles, tingling, or burning sensations?
3. Tremor of your arms or legs?
4. Hot or cold temperatures making your MS worse?
5. Pain?
6. **Weakness anywhere in your body?**
7. Heavy arms and/or legs?
8. Spasms in your limbs?
9. Stiffness?
10. **Problems with your balance?**
11. Difficulties moving about outdoors?
12. Difficulties moving about indoors?
13. *Falling over?*
14. Being clumsy?
15. *Dizziness?*
16. *Problems speaking clearly?*
17. *Tinnitus (ringing in the ear)?*
18. *Problems finding or remembering words when speaking?*
19. Problems with your vision when reading?
20. *Problems with your vision when doing things other than reading?*
21. **Memory problems?**
22. **Problems concentrating?**
23. **Problems keeping your attention when doing things?**
24. **Difficulties thinking clearly?**
25. *Difficulties organising things?*
26. *Difficulties learning new ways of doing things?*
27. *The effect of MS on your interest in sex?*
28. *The effect of MS on your ability to have sex?*
29. **The effect of MS on your enjoyment of sex?**
30. *Feeling unattractive to others?*
31. *Being unable to wear the clothes you want to?*
32. Needing to go to the toilet urgently?
33. **Needing to go to the toilet frequently?**
34. Problems with constipation?
35. *Problems emptying your bladder?*
36. *Bladder accidents (incontinence)?*
37. *Bowel accidents (incontinence)?*
38. *Using a catheter? (Circle "1" if you do not use a catheter)*
39. Problems sleeping?
40. *Disturbing other people's sleep?*
41. Feeling mentally fatigued?
42. **Feeling physically fatigued?**
43. Getting tired when you do things?
44. **Having to limit what you do because of tiredness?**
45. Feeling unwell?
46. Your body not doing what you want it to do?
47. *The effect of MS on the appearance of your body?*
48. Feeling depressed?
49. Worries related to your MS?
50. *Tearfulness?*
51. Feeling anxious or tense?
52. *Feeling angry?*
53. Feeling irritable, impatient, or short tempered?
54. Feeling frustrated?
55. Lack of confidence?
56. *Feeling bored?*
57. *Feeling afraid?*
58. *Feeling worthless?*
59. *Feeling lonely or isolated?*
60. *Feeling ignored by people because you have MS?*
61. *Feeling embarrassed by your MS?*
62. *Difficulties with your roles in your family (e.g. as a parent or partner)?*
63. The effect of MS on your spouse/partner or family?
64. Having to depend on others to do things for you?
65. **Having to depend on others being around to enable you to do things?**
66. *People doing things for you when you don't need or want them to?*
67. *A lack of* **emotional** *support from people close to you?*
68. **A lack of** *practical* **support from people close to you?**
69. *Problems communicating with family and friends?*
70. *Difficulties* **making** *new relationships?*
71. *Difficulties* **keeping** *close relationships?*
72. *Avoiding public places because of your MS?*
73. *Feeling that people treat you differently because of your MS?*

74. Limitations in your social and leisure activities **at home**?
75. Limitations in your social and leisure activities *outside your home***?**
76. Being stuck at home more than you would like to be?
77. Difficulties using your hands in everyday tasks?
78. Difficulties with self-care activities (e.g. washing, dressing)?
79. Having to cut down the amount of time you spent on work or other daily activities?
80. Not getting as much done as you would like at work or during your other daily activities?
81. Limitations in the type of work or other daily activities you can do?
82. Having to change your long-term work plans?
83. The effect of MS on your driving? (Circle '1' if you do not have a driving licence)
84. Problems using transport (e.g. car, bus, train, taxi)?
85. Taking longer to do things?
86. Having to plan your life around your MS?
87. Difficulties planning things on a day-to-day basis?
88. Difficulties planning things in the future (e.g. going away for weekends or holidays)?
89. Uncertainty about what the future holds for you?
90. Difficulty doing things spontaneously (e.g. going out on the spur of the moment)?
91. Having to use a wheelchair? (Circle "1" if you do not use a wheelchair)
92. Having to make changes to your home because of your MS (e.g. bath rails, wheelchair ramp)?
93. Side-effects of drugs you take for MS? (Circle '1' if you do not take any drugs for your MS)
94. Financial difficulties because of your MS?
95. Difficulties getting answers from your doctors about your MS?
96. MS stopping you achieving what you want from life?
97. Feeling that you have missed out on things because of your MS?
98. Do physically demanding tasks?
99. Walk indoors?
100. Walk outdoors?
101. Run?
102. Climb up and down stairs?
103. Stand when doing things?
104. Stay sitting up straight?
105. Get up from sitting?
106. Sit up from lying?
107. Transfer into and out of a chair or a wheelchair?
108. Keep your balance when standing or walking?
109. Move or turn over in bed?
110. Get in or out of a car or other similar vehicle?
111. Bend down?
112. Get in or out of the bath or shower, whichever you use the most often?
113. Do up buttons?
114. Cut up food?
115. Swallow some types of food?
116. Drink without coughing?
117. Wash your hair?
118. Brush your teeth, put on make-up, shave, etc.?
119. Wash the *top half* **of your body?**
120. Wash the *bottom half* **of your body?**
121. Dress the **top half** *of your body?*
122. Dress the **bottom half** *of your body?*
123. Go to the toilet (bladder)?
124. Go to the toilet (bowel)?
125. Write?
126. Grip things tightly (e.g. turning on taps)?
127. Use both hands together when doing things?
128. Hold things?
129. Carry things?

Stage 2

Reduction from 42 items to 39 items

Three items removed as they did not load ≥ 0.40 on either of the two factors:

1. Numbness or loss of sensation
19. Problems with vision when reading
34. Constipation

Reduction from 39 items to 34 items

Five items considered conceptually difficult to retain within their respective factors:

Physical impact scale:
63. Effect of MS on your spouse/partner or family
82. Having to change long-term work plans
97. Feeling that you have missed out on things because of your MS

Psychological impact scale:
2. Pins and needles, tingling or burning sensations
5. Pain

Reduction from 34 items to 30 items

Two items consistently registered probable scaling successes (rather than definite scaling successes) on testing item convergent and discriminant validity:

4. Hot or cold temperatures making your MS worse (physical impact scale)
54. Feeling frustrated (psychological impact scale)

Two other items were eliminated:

99. Walk indoors
removed as content considered well covered by item 12 (difficulties moving about indoors), which was also considered more applicable to people who could not walk

87. Difficulties planning things on a day-to-day basis
removed as it did not fit the conceptual definition of physical impact.

Reduction from 30 to 29 items

One item consistently registered probable scaling successes (rather than definite scaling successes) on testing item convergent and discriminant validity:

43. Getting tired when you do things

Appendix 2

Multiple Sclerosis Impact Scale (MSIS-29)[a]

- The following questions ask for your views about the impact of MS on your day-to-day life **during the past two weeks**
- For each statement, please **circle** the **one** number that **best** describes your situation
- Please answer **all** questions

In the past two weeks, how much has your MS limited your ability to …	Not at all	A little	Moderately	Quite a bit	Extremely
1. Do physically demanding tasks?	1	2	3	4	5
2. Grip things tightly (e.g. turning on taps)?	1	2	3	4	5
3. Carry things?	1	2	3	4	5

In the past two weeks, how much have you been bothered by …	Not at all	A little	Moderately	Quite a bit	Extremely
4. Problems with your balance?	1	2	3	4	5
5. Difficulties moving about indoors?	1	2	3	4	5
6. Being clumsy?	1	2	3	4	5
7. Stiffness?	1	2	3	4	5
8. Heavy arms and/or legs?	1	2	3	4	5
9. Tremor of your arms or legs?	1	2	3	4	5
10. Spasms in your limbs?	1	2	3	4	5
11. Your body not doing what you want it to do?	1	2	3	4	5
12. Having to depend on others to do things for you?	1	2	3	4	5

[a] © Neurological Outcome Measures Unit, Institute of Neurology, University College London, WC1N 3BG, 2001.

In the past two weeks, how much have you been bothered by …	Not at all	A little	Moderately	Quite a bit	Extremely
13. Limitations in your social and leisure activities at home?	1	2	3	4	5
14. Being stuck at home more than you would like to be?	1	2	3	4	5
15. Difficulties using your hands in everyday tasks?	1	2	3	4	5
16. Having to cut down the amount of time you spent on work or other daily activities?	1	2	3	4	5
17. Problems using transport (e.g. car, bus, train, taxi, etc.)?	1	2	3	4	5
18. Taking longer to do things?	1	2	3	4	5
19. Difficulty doing things spontaneously (e.g. going out on the spur of the moment)?	1	2	3	4	5
20. Needing to go to the toilet urgently?	1	2	3	4	5
21. Feeling unwell?	1	2	3	4	5
22. Problems sleeping?	1	2	3	4	5
23. Feeling mentally fatigued?	1	2	3	4	5
24. Worries related to your MS?	1	2	3	4	5
25. Feeling anxious or tense?	1	2	3	4	5
26. Feeling irritable, impatient, or short tempered?	1	2	3	4	5
27. Problems concentrating?	1	2	3	4	5
28. Lack of confidence?	1	2	3	4	5
29. Feeling depressed?	1	2	3	4	5

Appendix 3

Instructions for administration and scoring the MSIS-29

Administration

This section outlines the administration procedures of the MSIS-29. The MSIS-29 is simple to administer and takes only a few minutes to complete. It took an average of 2 minutes 44 seconds to complete in a sample of ten people studied (range = 1 minute 45 seconds to 4 minutes 26 seconds). It can be handed to people directly in a clinical or research setting for completion, or administered via postal survey. For a more detailed description of the administering instructions, please see *Multiple Sclerosis Impact Scale (MSIS-29): user manual*. Please contact the lead author of this report for a copy of the manual: Dr Jeremy Hobart, Consultant Neurologist and Honorary Senior Clinical Research Fellow in the Peninsula Medical School, Department of Clinical Neurosciences, Derriford Hospital, Plymouth PL6 8DH.

Where the MSIS-29 is handed to patient for self-completion

People may ask for clarification on a particular question. It is important not to influence the patients' responses as the questionnaire is designed to be a self-report measure, and the individual's perspective is of interest. Repeat the question and the questionnaire instruction again. If the respondent asks what a particular question means, ask them to answer the questions according to what they *think* it means. If the respondent has difficulty choosing from the response options, gently guide them to the response option that *most closely* resembles how they think or feel. If they say they don't like a particular question, or thinks it is unnecessary or inappropriate, emphasise that all the questions are included in the questionnaire for a reason. Do encourage respondents to fill in all the questions.

Some people with MS are unable to fill in the questionnaire due to their disability. In these cases, the following steps should be taken. If a respondent has visual difficulties, there are two options. First, a large-type version of the questionnaire could be prepared and offered. The second option is to administer the questionnaire by interview. If the questions are to be read aloud to patients, make sure that you do not influence their response. Read the questions and the response options exactly as they are written in the questionnaire. The questionnaire is designed as a self-report measure, and clinicians must make sure not to influence the responses. If a respondent has difficulties in circling the answers because of weakness, tremor or poor coordination, the administrator should circle the answers for the respondent.

If the respondent refuses to complete a question or the questionnaire, tell them that the completion of the questionnaire is voluntary, but that it will provide important information on how people with MS manage their illness. However, never force patients to answer a particular question or the questionnaire. Remember that participation is voluntary. If the respondent still refuses, ask them whether there were any particular reasons for refusal, record this, and thank the respondent.

Where the MSIS-29 is administered via postal survey

Clear written instructions, preferably in a covering letter or an information sheet, should be given to respondents who are completing the questionnaire at home. Explain how to fill in the questionnaire, i.e. circle one response to each question, and that it is easy to complete. Remind them to read the instructions at the top of the questionnaire carefully. Emphasise that there are no right or wrong answers and to choose one answer which is closest to the way that they are feeling. Explain that it is important to get their views for a better understanding of how people with MS manage their illness. Stress the need to answer all the questions even if some questions seem similar to others or not applicable to them. Explain that this is because MS affects people in many different ways. If they need help in filling out the questionnaire (reading or writing), they can get someone to help, but emphasise that their answers should be their own. Make sure to give them a contact number for any queries they have with regards to the completion of the questionnaire, and thank them for their time and effort.

Scoring

Physical impact score

The physical impact score is computed by summing items number 1–20 inclusive.

This score can then by transformed to a score on a scale of 0–100 using the formula below:

$$\frac{100 \times (\text{observed score} - 20)}{100 - 20}$$

Psychological impact score

The psychological score is computed by summing items number 21–29 inclusive.

This score can then by transformed to a score on a scale of 0–100 using the formula below:

$$\frac{100 \times (\text{observed score} - 9)}{45 - 9}$$

For more detailed description of the scoring instructions, please see *Multiple Sclerosis Impact Scale (MSIS-29): user manual*. Please contact the lead author of this report for a copy of the manual: Dr Jeremy Hobart, Consultant Neurologist and Honorary Senior Clinical Research Fellow in the Peninsula Medical School, Department of Clinical Neurosciences, Derriford Hospital, Plymouth PL6 8DH.

Health Technology Assessment Programme

Prioritisation Strategy Group

Members

Chair,
Professor Tom Walley, Director, NHS HTA Programme & Professor of Clinical Pharmacology, University of Liverpool

Professor Bruce Campbell, Consultant Vascular & General Surgeon, Royal Devon & Exeter Hospital

Professor Shah Ebrahim, Professor in Epidemiology of Ageing, University of Bristol

Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Radcliffe Hospital, Oxford

Dr Ron Zimmern, Director, Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

HTA Commissioning Board

Members

Programme Director,
Professor Tom Walley, Director, NHS HTA Programme & Professor of Clinical Pharmacology, University of Liverpool

Chair,
Professor Shah Ebrahim, Professor in Epidemiology of Ageing, Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol

Deputy Chair,
Professor Jenny Hewison, Professor of Health Care Psychology, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds School of Medicine, Leeds

Professor Douglas Altman, Professor of Statistics in Medicine, Centre for Statistics in Medicine, Oxford University, Institute of Health Sciences, Cancer Research UK Medical Statistics Group, Headington, Oxford

Professor John Bond, Professor of Health Services Research, Centre for Health Services Research, University of Newcastle, School of Health Sciences, Newcastle upon Tyne

Professor John Brazier, Director of Health Economics, Sheffield Health Economics Group, School of Health & Related Research, University of Sheffield, ScHARR Regent Court, Sheffield

Dr Andrew Briggs, Public Health Career Scientist, Health Economics Research Centre, University of Oxford, Institute of Health Sciences, Oxford

Dr Christine Clark, Medical Writer & Consultant Pharmacist, Cloudside, Rossendale, Lancs
and
Principal Research Fellow, Clinical Therapeutics in the School of Pharmacy, Bradford University, Bradford

Professor Nicky Cullum, Director of Centre for Evidence Based Nursing, Department of Health Sciences, University of York, Research Section, Seebohm Rowntree Building, Heslington, York

Dr Andrew Farmer, Senior Lecturer in General Practice, Department of Primary Health Care, University of Oxford, Institute of Health Sciences, Headington, Oxford

Professor Fiona J Gilbert, Professor of Radiology, Department of Radiology, University of Aberdeen, Lilian Sutton Building, Foresterhill, Aberdeen

Professor Adrian Grant, Director, Health Services Research Unit, University of Aberdeen, Drew Kay Wing, Polwarth Building, Foresterhill, Aberdeen

Professor Alastair Gray, Director, Health Economics Research Centre, University of Oxford, Institute of Health Sciences, Headington, Oxford

Professor Mark Haggard, Director, MRC ESS Team, CBU Elsworth House, Addenbrooke's Hospital, Cambridge

Professor F D Richard Hobbs, Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham, Primary Care and Clinical Sciences Building, Edgbaston, Birmingham

Professor Peter Jones, Head of Department, University Department of Psychiatry, University of Cambridge, Addenbrooke's Hospital, Cambridge

Professor Sallie Lamb, Research Professor in Physiotherapy/Co-Director, Interdisciplinary Research Centre in Health, Coventry University, Coventry

Dr Donna Lamping, Senior Lecturer, Health Services Research Unit, Public Health and Policy, London School of Hygiene and Tropical Medicine, London

Professor David Neal, Professor of Surgical Oncology, Oncology Centre, Addenbrooke's Hospital, Cambridge

Professor Tim Peters, Professor of Primary Care Health Services Research, Division of Primary Health Care, University of Bristol, Cotham House, Cotham Hill, Bristol

Professor Ian Roberts, Professor of Epidemiology & Public Health, Intervention Research Unit, London School of Hygiene and Tropical Medicine, London

Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh, Western General Hospital NHS Trust, Bramwell Dott Building, Edinburgh

Professor Martin Severs, Professor in Elderly Health Care, Portsmouth Institute of Medicine, Health & Social Care, St George's Building, Portsmouth

Dr Jonathan Shapiro, Senior Fellow, Health Services Management Centre, Park House, Birmingham

Current and past membership details of all HTA 'committees' are available from the HTA website (www.ncchta.org)

Diagnostic Technologies & Screening Panel

Members

Chair,
Dr Ron Zimmern, Director of the Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

Dr Paul Cockcroft, Consultant Medical Microbiologist/ Laboratory Director, Public Health Laboratory, St Mary's Hospital, Portsmouth

Professor Adrian K Dixon, Professor of Radiology, Addenbrooke's Hospital, Cambridge

Dr David Elliman, Consultant in Community Child Health, London

Dr Andrew Farmer, Senior Lecturer in General Practice, Institute of Health Sciences, University of Oxford

Dr Karen N Foster, Clinical Lecturer, Dept of General Practice & Primary Care, University of Aberdeen

Professor Jane Franklyn, Professor of Medicine, University of Birmingham

Professor Antony J Franks, Deputy Medical Director, The Leeds Teaching Hospitals NHS Trust

Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London

Dr Susanne M Ludgate, Medical Director, Medical Devices Agency, London

Dr William Rosenberg, Senior Lecturer and Consultant in Medicine, University of Southampton

Dr Susan Schonfield, CPHM Specialised Services Commissioning, Croydon Primary Care Trust

Dr Margaret Somerville, Director of Public Health, Teignbridge Primary Care Trust, Devon

Mr Tony Tester, Chief Officer, South Bedfordshire Community Health Council, Luton

Dr Andrew Walker, Senior Lecturer in Health Economics, University of Glasgow

Professor Martin J Whittle, Head of Division of Reproductive & Child Health, University of Birmingham

Dr Dennis Wright, Consultant Biochemist & Clinical Director, Pathology & The Kennedy Galton Centre, Northwick Park & St Mark's Hospitals, Harrow

Pharmaceuticals Panel

Members

Chair,
Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Oxford Radcliffe Hospital

Professor Tony Avery, Professor of Primary Health Care, University of Nottingham

Professor Iain T Cameron, Professor of Obstetrics & Gynaecology, University of Southampton

Mr Peter Cardy, Chief Executive, Macmillan Cancer Relief, London

Dr Christopher Cates, GP and Cochrane Editor, Bushey Health Centre, Bushey, Herts.

Mr Charles Dobson, Special Projects Adviser, Department of Health

Dr Robin Ferner, Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Dr Karen A Fitzgerald, Pharmaceutical Adviser, Bro Taf Health Authority, Cardiff

Professor Alastair Gray, Professor of Health Economics, Institute of Health Sciences, University of Oxford

Mrs Sharon Hart, Managing Editor, *Drug & Therapeutics Bulletin*, London

Dr Christine Hine, Consultant in Public Health Medicine, Bristol South & West Primary Care Trust

Professor Robert Peveler, Professor of Liaison Psychiatry, Royal South Hants Hospital, Southampton

Dr Frances Rotblat, CPMP Delegate, Medicines Control Agency, London

Mrs Katrina Simister, New Products Manager, National Prescribing Centre, Liverpool

Dr Ken Stein, Senior Lecturer in Public Health, University of Exeter

Professor Terence Stephenson, Professor of Child Health, University of Nottingham

Dr Richard Tiner, Medical Director, Association of the British Pharmaceutical Industry, London

Professor Dame Jenifer Wilson-Barnett, Head of Florence Nightingale School of Nursing & Midwifery, King's College, London

Therapeutic Procedures Panel

Members

Chair,
Professor Bruce Campbell,
Consultant Vascular and General Surgeon, Royal Devon & Exeter Hospital

Dr Mahmood Adil, Head of Clinical Support & Health Protection, Directorate of Health and Social Care (North), Department of Health, Manchester

Professor John Bond, Head of Centre for Health Services Research, University of Newcastle upon Tyne

Mr Michael Clancy, Consultant in A & E Medicine, Southampton General Hospital

Dr Carl E Counsell, Senior Lecturer in Neurology, University of Aberdeen

Dr Keith Dodd, Consultant Paediatrician, Derbyshire Children's Hospital, Derby

Professor Gene Feder, Professor of Primary Care R&D, Barts & the London, Queen Mary's School of Medicine and Dentistry, University of London

Ms Bec Hanley, Freelance Consumer Advocate, Hurstpierpoint, West Sussex

Professor Alan Horwich, Director of Clinical R&D, The Institute of Cancer Research, London

Dr Phillip Leech, Principal Medical Officer for Primary Care, Department of Health, London

Mr George Levvy, Chief Executive, Motor Neurone Disease Association, Northampton

Professor James Lindesay, Professor of Psychiatry for the Elderly, University of Leicester

Dr Mike McGovern, Senior Medical Officer, Heart Team, Department of Health, London

Dr John C Pounsford, Consultant Physician, North Bristol NHS Trust

Professor Mark Sculpher, Professor of Health Economics, Institute for Research in the Social Services, University of York

Dr L David Smith, Consultant Cardiologist, Royal Devon & Exeter Hospital

Professor Norman Waugh, Professor of Public Health, University of Aberdeen

Expert Advisory Network

Members

Mr Gordon Aylward,
Chief Executive,
Association of British Health-Care Industries, London

Ms Judith Brodie,
Head of Cancer Support Service, Cancer BACUP, London

Mr Shaun Brogan,
Chief Executive, Ridgeway Primary Care Group, Aylesbury, Bucks

Ms Tracy Bury,
Project Manager, World Confederation for Physical Therapy, London

Mr John A Cairns,
Professor of Health Economics, Health Economics Research Unit, University of Aberdeen

Professor Howard Stephen Cuckle,
Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Professor Nicky Cullum,
Director of Centre for Evidence Based Nursing, University of York

Dr Katherine Darton,
Information Unit, MIND – The Mental Health Charity, London

Professor Carol Dezateux,
Professor of Paediatric Epidemiology, London

Professor Martin Eccles,
Professor of Clinical Effectiveness, Centre for Health Services Research, University of Newcastle upon Tyne

Professor Pam Enderby,
Professor of Community Rehabilitation, Institute of General Practice and Primary Care, University of Sheffield

Mr Leonard R Fenwick,
Chief Executive, Newcastle upon Tyne Hospitals NHS Trust

Professor David Field,
Professor of Neonatal Medicine, Child Health, The Leicester Royal Infirmary NHS Trust

Mrs Gillian Fletcher,
Antenatal Teacher & Tutor and President, National Childbirth Trust, Henfield, West Sussex

Ms Grace Gibbs,
Deputy Chief Executive, Director for Nursing, Midwifery & Clinical Support Servs., West Middlesex University Hospital, Isleworth, Middlesex

Dr Neville Goodman,
Consultant Anaesthetist, Southmead Hospital, Bristol

Professor Robert E Hawkins,
CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor F D Richard Hobbs,
Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham

Professor Allen Hutchinson,
Director of Public Health & Deputy Dean of ScHARR, Department of Public Health, University of Sheffield

Professor Rajan Madhok,
Medical Director & Director of Public Health, Directorate of Clinical Strategy & Public Health, North & East Yorkshire & Northern Lincolnshire Health Authority, York

Professor David Mant,
Professor of General Practice, Department of Primary Care, University of Oxford

Professor Alexander Markham,
Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Dr Chris McCall,
General Practitioner, The Hadleigh Practice, Castle Mullen, Dorset

Professor Alistair McGuire,
Professor of Health Economics, London School of Economics

Dr Peter Moore,
Freelance Science Writer, Ashtead, Surrey

Dr Andrew Mortimore,
Consultant in Public Health Medicine, Southampton City Primary Care Trust

Dr Sue Moss,
Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton, Surrey

Professor Jon Nicholl,
Director of Medical Care Research Unit, School of Health and Related Research, University of Sheffield

Mrs Julietta Patnick,
National Co-ordinator, NHS Cancer Screening Programmes, Sheffield

Professor Chris Price,
Visiting Chair – Oxford, Clinical Research, Bayer Diagnostics Europe, Cirencester

Ms Marianne Rigge,
Director, College of Health, London

Professor Sarah Stewart-Brown,
Director HSRU/Honorary Consultant in PH Medicine, Department of Public Health, University of Oxford

Professor Ala Szczepura,
Professor of Health Service Research, Centre for Health Services Studies, University of Warwick

Dr Ross Taylor,
Senior Lecturer, Department of General Practice and Primary Care, University of Aberdeen

Mrs Joan Webster,
Consumer member, HTA – Expert Advisory Network

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278