Development of a preference-based measure for multiple sclerosis: The Preference-Based Multiple Sclerosis Index (PBMSI)

Ayse Kuspinar, B.Sc. (Physical Therapy), M.Sc. (Rehabilitation Sciences)

School of Physical and Occupational Therapy

Faculty of Medicine

McGill University

Montreal, Quebec, Canada

December, 2014

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy (Rehabilitation Sciences)

Table of Contents

LIST OF TABLES

Supp: Supplementary.

LIST OF FIGURES

Supp: Supplementary.

ABSTRACT

Assessing health-related quality of life (HRQL) has moved to the forefront of clinical research and is considered a crucial endpoint of clinical interventions. One approach to assessing HRQL is through the use of health profiles. Health profiles are analyzed by sub-scale, where each sub-scale represents a domain of health. These measures do not provide information on the relative importance attached to each domain. As a result, the domains cannot be combined into an overall score, and a trade-off cannot be made between domains when evaluating the effectiveness of interventions. Another approach to measuring HRQL is through the use of preference-based measures. Not only do these measures provide descriptive information on the various dimensions of health, but also provide a value for each. They have the advantage of leading to a single number that balances gains in one domain against losses in another. When linked to life-expectancy, they provide measures of quality adjusted life years (QALY) and are used to make decisions about the cost-effectiveness of interventions. The best known preference-based measures are the Health Utilities Index (HUI), the EuroQol-5D (EQ-5D) and the Short Form-6D (SF-6D). However, the challenge of using such generic preference-based measures in people with Multiple Sclerosis (MS) is that they may not capture all domains of health relevant to the disease and the domain weighting is based on the values from the naive general population.

Therefore, the overall objective of this PhD thesis is to take important steps towards developing a Preference-Based Multiple Sclerosis Index (PBMSI) for use as a global outcome in clinical and cost-effectiveness studies for MS.

To do this, a systematic review of HRQL outcomes in MS interventions was carried out and identified that an imporant source of heterogeneity in the literature arises from the many different measures used and domains evaluated (Manuscript 1). As preference-based measures reduce some of the heterogeneity by yielding one value across mutliple domains of health, the content of generic preference-based measures was assessed in light of the domains identified as being important to people with MS (Manuscript 2), and a review of their psychometric properties was carried out (Manucript 3). Results revealed that these generic measures were missing several domains that were affected by MS, such as walking, fatigue and cognition, identifying a measurement gap. Making use of a rich data source (that I had previously collected as part of my MSc), optimally performing items targeting the important MS domains were identified and tested

for their discriminatory capacity with respect to known groups with differing disability (Manuscript 4). This study yielded a set of 5 bilingual items (English and French) ready for testing for comprehension and wording using cognitive interviewing with a sample of 22 people with MS (Manuscript 5). An item met criteria for acceptability after 3 to 4 rounds of interviews.

The final step in this thesis was to elict preferences for different health states generated through combinations of items, using two different standard methods of preference elicitation which are known to have conceptual and practical differences (Standard Gamble and Rating Scale). Manuscript 6 presents the results of this preliminary investigation in a sample of 61 patients with MS. The results indicate that the Standard Gamble is difficult for patients to understand and produces higher values than the Rating Scale. The scoring algorithm developed based on each of the methods yielded vastly different results. Although the Standard Gamble is a classical technique of measuring preferences using decision making, it was not practical in this patient population. On the other hand, the Rating Scale is more suitable for the population but the values are not choice based potentially limiting their use for economic evaluation of interventions.

RÉSUMÉ

Évaluer la qualité de vie liée à la santé (QVLS) est devenue une préoccupation de premier plan en recherche clinique et est considéré comme un critère d'évaluation crucial des interventions cliniques. Une des approches utilisées pour évaluer la QVLS est le profil de santé. Les profils de santé sont analysés par des sous-échelles, où chaque sous-échelle représente un domaine de la santé. Cependant, les profils de santé ne fournissent aucune information sur l'importance relative de chaque domaine. En conséquence, les domaines ne peuvent être combinés en un score global et un compromis entre les domaines ne peut être fait lors d'évaluation de l'efficacité d'une intervention. Une autre approche pour mesurer la QVLS est par l'utilisation des mesures basées sur les préférences. Ces mesures fournissent une valeur pour chacune des différentes dimensions de la santé en plus de donner une description sur celles-ci. Elles ont l'avantage d'offrir un nombre unique équilibrant les gains dans un domaine avec les pertes dans un autre. Lorsqu'elles sont associées à l'espérance de vie, elles fournissent une mesure des années de vie pondérées en fonction de leur qualité (QALY – *quality-adjusted life year*) et sont utilisées pour prendre des décisions sur le rapport coût-efficacité des interventions. Les mesures basées sur les préférences les plus connues sont le Health Utilities Index (HUI), le EuroQol-5D (EQ-5D) et le Short Form-6D (SF-6D). Par contre, un des défis avec l'utilisation de ces mesures basées sur les préférences génériques chez les personnes atteintes de sclérose en plaque (SP) est qu'elles ne capturent pas nécessairement tous les domaines de la santé pertinents à cette maladie et la pondération de chaque domaine est basée sur les valeurs de la population en générale, naïve à la maladie.

Par conséquent, l'objectif global de cette thèse de doctorat est d'aller de l'avant dans le développement d'une mesure basée sur les préférences spécifique à la SP, le Preference-Based Multiple Sclerosis Index (PBMSI), afin de l'utiliser dans les études cliniques à titre d'indicateur et de rapport coût-efficacité pour la SP.

Pour y parvenir, une revue systématique des indicateurs de la QVLS ciblant les études d'interventions sur la SP a été réalisée et a identifiée qu'une source importante d'hétérogénéité dans la littérature provient de la grande diversité de mesures utilisées et des domaines évalués (Manuscrit 1). Puisque les mesures basées sur les préférences réduisent en partie cette hétérogénéité en offrant une valeur résumant plusieurs domaines de la santé, le contenu des mesures basées sur les préférences génériques a été évalué en fonction des domaines identifiés

comme étant important pour les gens atteint de SP (Manuscrit 2). De plus, une revue de leur propriété psychométriques a également été effectuée (Manuscrit 3). Les résultats ont révélé qu'il manquait plusieurs domaines affectés par la SP dans ces mesures génériques, tels que la marche, la fatigue et les facultés cognitives, identifiant ainsi une lacune au niveau de la mesure. Faisant usage d'une banque de donnée riche en information (données collectées dans le cadre de ma maîtrise), les items ciblant les domaines importants de la SP et ayant une performance optimale ont été identifiés et testés sur leur capacité discriminatoire en fonction de groupes connus ayant différentes incapacités (Manuscrit 4). De cette étude est ressortie 5 items bilingues (anglais et français) prêts à être testés pour leur compréhension et formulation à l'aide d'entrevues cognitives sur un échantillon de 24 personnes atteintes de SP (Manuscrit 5). Les items ont atteint les critères d'acceptabilité après 3 ou 4 tours d'entrevues.

La dernière étape de cette thèse a été d'établir les préférences pour différents états de santé. Elles ont été générées à l'aide de combinaisons d'items utilisant deux méthodes standards et différentes pour éliciter les préférences (pari standard et échelle d'évaluation). Ces méthodes sont reconnues pour avoir des différences conceptuelles et pratiques. Le Manuscrit 6 présente les résultats de cette étude préliminaire avec un échantillon de 61 personnes atteintes de SP. Les résultats indiquent que le pari standard est très difficile à comprendre pour les patients et produit des valeurs plus élevées que l'échelle d'évaluation. Les algorithmes de pointage développés selon les deux méthodes produisent des résultats grandement différents. Malgré le fait que le pari standard soit une technique classique pour mesurer les préférences en utilisant la prise de décision, il s'est avéré que celle-ci n'était pas pratique avec cette population. En contre partie, l'échelle d'évaluation s'est avérée plus appropriée pour cette population. Cependant, le fait que les valeurs ne sont pas basées sur les choix pourrait potentiellement limiter leur utilisation pour l'évaluation économique des interventions.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the contributions of all the investigators in the ‘Gender and Life Impact of Multiple Sclerosis Study’, namely Dr. Nancy Mayo, Dr. Pierre Duquette, Dr. Yves Lapierre, Dr. Francois Grand’Maison, and Dr. Lisa Koski. This would not have been possible also without the participation of the patients who gave their valuable time for the assessment of multiple tests and measures, and of my colleague Lily Teng who collected all of the data with me. The Gender and Life Impact of Multiple Sclerosis Study has provided the rich data source for furthering my MSc study into my present doctoral work.

I would like to express my sincere gratitude to my supervisor, Dr. Nancy Mayo, for her invaluable guidance and support throughout my PhD studies. She has always been very generous in supervising my dissertation and helping me understand the methodological and statistical aspects of this research. I am immensely indebted for her willingness to meet with me regularly, for her readiness to answer my questions happily, and for her careful attentiveness in the development of my thesis.

I would like to also extend my thanks to the other esteemed members of my supervisory committee, Dr. Pierre Duquette, Dr. Eric Latimer and Dr. Simon Pickard for their professional inputs and valuable contributions in the development of my work.

I wish to thank Dr. Lois Finch for her timely given answers to my queries on Rasch Analysis, Dr. Susan Bartlett for her kind support throughout my research, Carolina Moriello for her generous help with the recruitment of subjects, and Susan Scott for her guidance in the statistical analysis of the data.

My gratitude goes further to my colleagues; Dr. Shahnaz Sharhbanian for her continuous remarkable moral support throughout the past years, Dr. Ana Maria Rodriguez for her thoughtful assistance with the first manuscript of my work along with her genuine friendship throughout my studies, Vanessa Bouchard for her tremendous help and assistance during both the translations and the cognitive interviews in French, and finally to the rest of my friends and colleagues at the Division of Clinical Epidemiology: Sabrina Figuerido, Marie-Eve Letellier, Christiane Lourenco, and Stanley Hum for their kind enthusiastic support and friendship.

There comes in the end my dear family who deserves to be duly included in the list of my humble acknowledgement. Firstly, my deepest gratitude goes to my mom and dad, who have always been inspirational role models in my life, and who have kindly and patiently supported me throughout my entire education, especially in the past year by looking after their little grandson while I worked on my thesis. I will ever remain grateful for their unconditional love, care and guidance. The next in the list of my gratitude is my sister who has always been my best friend all my life. I love her dearly and thank her for all her constant encouragement. Last but not the least, no words suffice to express my indebtedness to my husband for his unfailing moral support, love and patience during my PhD studies. Also, the new addition to our family, our son Emre has been a bundle of joy and has made the past year ever memorable, especially for sharing the late nights with me while I wrote my thesis.

PREFACE

Statement of Originality

In this thesis we describe a step-by-step process towards developing a preference-based measure for MS, titled the Preference-Based Multiple Sclerosis Index (PBMSI). The topic arose out of my experience in the Gender and Life Impact of Multiple Sclerosis Study when personally collecting health and disability outcome data on 189 people with MS. After that experience, it was evident to me that a different approach to the measurement of important health outcomes was needed. I identified that a MS specific preference-based measure would fill this gap. This thesis presents key developmental steps towards this goal.

The approach taken closely followed the guidelines and recommendations set by the Food and Drug Administration (FDA) for the development of patient-reported outcomes. As demonstrated in this thesis, considerable conceptual work using the published literature, expert experience, and patient input was carried out to develop the domains of the MS specific preference-based measure. Furthermore, modern psychometric methods were used to select items and verify their response levels. These selected items were then further refined using cognitive interviews in both English and French to ensure their readability and understanding by patients. Last, patients were asked to rate their preferences for each item using two different methods, the standard gamble and the rating scale. To my knowledge, there is no MS-specific preference-based measure that has been developed on patient input concerning the domains, items and preference weights. Therefore, the overall objective of this PhD thesis was to take important steps towards developing a Preference-Based Multiple Sclerosis Index (PBMSI) for use as a global outcome in clinical and cost-effectiveness studies for MS.

Contribution of authors

This thesis builds upon work from the Gender Life Impact of Multiple Sclerosis Study (PI Nancy Mayo) for which I assessed 189 MS patients on a series of performance-based and self-reported tests.

The manuscripts included in this thesis are the work of Ayse Kuspinar with extensive editing and feedback from Dr. Nancy Mayo and Dr. Simon Pickard. For all of the six manuscripts, data collection, statistical analysis and write up were conducted by the doctoral candidate under the direct supervision of Dr. Nancy Mayo. As supervisor, Dr. Nancy Mayo oversaw all aspects of the thesis and provided expertise regarding research methodology and statistics.

Dr. Ana Maria Rodriguez was a co-author in the first manuscript, as she helped extract data from articles during the systematic review process. Dr. Lois Finch was a co-author on the fourth manuscript for providing statistical guidance on Rasch Analysis. Vanessa Bouchard co-authored the fifth manuscript for her assistance with the cognitive interview process in French. Dr. Simon Pickard was a co-author on the fourth and sixth manuscripts, for his expertise in health economics and for providing editorial feedback.

Thesis Organization and Overview

The thesis consists of six manuscripts, four of which have already been published in recognized scientific journals. In order to follow the regulations of the Graduate and Postdoctoral Studies (GPS), additional chapters have been incorporated in this thesis. As requested by the GPS, an introduction and conclusion independent of the manuscripts have been included. We must admit that duplications are inevitable in this thesis.

A brief outline of the thesis is as follows. *Chapter 1* is a literature review on Multiple Sclerosis (MS), preference-based measures and the Quality Adjusted Life Year (QALY).

Chapter 2 presents the rationale for developing the PBMSI and outlines the main objectives of the manuscripts.

Chapter 3 consists of the first manuscript entitled “The effects of clinical interventions on health-related quality of life in multiple sclerosis: a meta-analysis”. The study’s objective is to estimate the extent to which existing health care interventions designed specifically to target health-related quality of life (HRQL) in persons with MS achieve this aim. This study identified all randomized clinical trials in MS that used HRQL as an outcome, and therefore provided me with the foundational knowledge in the area of HRQL measurement. This work has been published in *Multiple Sclerosis Journal*.

Chapter 4 links the first manuscript to the second manuscript.

Chapter 5 consists of the second manuscript entitled “Do generic utility measures capture what is important to the quality of life of people with MS?”. The objective of this study was to estimate the extent to which generic utility measures captured important domains that are affected by MS. This study determined the domains of the Preference-Based Multiple Sclerosis Index (PBMSI) and critiqued the content validity of generic preference-based measures in MS. This work has been published in *Health and Quality of Life Outcomes*.

Chapter 6 links the second manuscript to the third manuscript.

Chapter 7 consists of the third manuscript entitled “A review of the psychometric properties of generic utility measures in multiple sclerosis”. This study was a structured review of the psychometric properties of generic preference-based measures in MS. This study summarized not only the published literature on the topic, but also included original data that was collected in our unit. This work has been published in *PharmacoEconomics*.

Chapter 8 links the third manuscript to the fourth manuscript.

Chapter 9 consists of the fourth manuscript entitled “Using existing data to identify candidate items for a health state classification in multiple sclerosis”. The main aim of this paper is to describe the development of the *prototype* Preference-Based MS Index (P-PBMSI). This paper identified items best reflecting the domains of quality of life important to people with MS; and provided evidence for the discriminative capacity of the response options by cross walking onto a visual analogue scale (VAS) of health rating. This work has been published in *Quality of Life Research*.

Chapter 10 links the fourth manuscript to the fifth manuscript.

Chapter 11 consists of the fifth manuscript titled “The development of a bilingual MS-specific health classification system: the Preference-Based Multiple Sclerosis Index (PBMSI).” The objective of this study was to qualitatively revise the PBMSI items using expert and patient feedback.

Chapter 12 links the fifth manuscript to the sixth manuscript.

Chapter 13 consists of the sixth manuscript titled “Developing a valuation function for a multiple sclerosis specific classification system: comparison of standard gamble and rating scale”. In this study we elicited patient preferences for the different items in the PBMSI using the Standard

Gamble and the Rating Scale. The purpose of this study was to contribute preliminary evidence towards the similarities and differences in the Standard Gamble and the Rating Scale to reflect patient preferences for the different items in the PBMSI, where contrasts were on absolute and utility values, level of difficulty, and discriminative ability.

Chapter 14 is a summary of the findings and conclusions of the six manuscripts, as well as the implications for future research.

Corresponding figures, tables and references are presented at the end of each manuscript. Reference styles were based on each journal's requirements. The appendices include information that were not presented in the manuscripts, but were important to include in the thesis.

Ethics approval for the studies was obtained from the Research Ethics Board of the McGill University Health Center.

CHAPTER 1

Overview of Multiple Sclerosis

What is MS?

Multiple sclerosis (MS) is an inflammatory, demyelinating disease of the central nervous system (CNS) that can lead to the manifestation of a range of symptoms.[1] The prevalence rate in Canada is one of the highest in the world at 240 per 100,000.[2] The patho-physiology involves damage to the nervous system by the body's own immune system.[3] Cells attack myelin sheath and underlying fibres, leading to disruption of signal transmission from the brain to the body.[4] The aetiology of MS is unknown, however, there is evidence that both genetic and environmental factors are involved in triggering the disease.[1;5]

Disease course in MS

In 1996 the United States National Multiple Sclerosis Society defined the disease course in MS into 4 types.[6] The most common type is relapsing-remitting MS (RRMS), characterized by acute attacks followed by full or partial recovery. Fifty percent of patients with RRMS develop secondary progressive MS (SPMS), described by a steady increase in disability with or without acute relapses. Primary progressive MS (PPMS) is distinguished by disease progression from onset and represents approximately 10% of MS patients. The least common known is progressive relapsing MS (PRMS) which is characterized by constant progression of disease from onset with superimposed relapses.[6;7]

In 2013, this classification was revised[8] as there is now an increased understanding of the disease and its pathology. For example, clinically isolated syndrome (CIS), which describes individuals with an initial episode of neurologic symptoms that could be MS but have yet to fulfill diagnostic criteria, has been added. Moreover, the new classification system categorizes all types of MS as active or non-active. Active MS is defined as the occurrence of clinical relapse or the presence of new lesions in the brain over a specified period of time, preferably at least one year.[8]

Measuring Disease Severity in MS

The most widely used outcome measure of disease severity and progression in MS is the Expanded Disability Status Scale (EDSS).[9] It is a classification scheme extending from 0 (normal neurological examination) to 10 (death due to MS). Scores 1.0 to 3.5 of the EDSS are scored using the Functional Systems (FS) component of the scale. The FS consists of the eight major systems of the central nervous system (CNS), which are pyramidal, cerebellar, brainstem, mental, spasticity, sensory, visual, and bowel and bladder. Scores 4.0 to 9.5 are scored primarily by the person's ability to ambulate. EDSS score of 6 and 6.5 refer to people who require an assistive device for ambulation, and scores 7.0 or greater consist of persons with severe disability, such as those requiring a wheelchair. It is administered by a neurologist and takes approximately 10 to 20 minutes to complete.[9]

Medical Treatment in MS

Disease Modifying Agents (DMAs) have played a critical role in the advancement of MS management. In 1993, the first immunomudulating agent was approved by the U.S. Food and Drug Administration (FDA) called interferon beta-1b (INF- β1b) for RRMS. Shortly after, interferon beta-1a (INF- β1a) and glatiramer acetate (GA) were also approved.[10] These drugs are referred to as first-line DMAs. Clinical trials demonstrate that these therapeutic agents decrease relapse rates by approximately 30%.[11-13] Side-effects include flu-like symptoms and injection-site reactions. Both INF and GA require regular, long-term, self-injection administration, which raises issues of tolerance and adherence to treatment.[14]

Second-line therapies that have been approved and that are more effective than INF and GA, are Natalizumab and Mitoxantrone.[14] These agents are administered only once a month or every 3 months intravenously, and have been shown to reduce relapse rates by 68%.[15] However, they have also been shown to have severe side effects such as progressive multifocal leukoencephalopathy[16] (a viral disease characterized by inflammation of white matter in the brain), cardiotoxicity,[17] and acute leukemia.[18]

More recently, oral DMAs have been developed to tackle the issue of adherence and tolerance that occurs with injectable DMAs. Currently there are four oral DMAs, three of which have already been approved (Fingolimod,[19;20] Teriflunomide,[21-23] Dimethyl Fumarate[24;25]) and one that is under investigation (Laquinimod[26]). Clinical trials have demonstrated that these agents are able to reduce relapse rates with the same efficiency as first-line (injectable) DMAs.[19-23] However, serious adverse events have been reported with these agents, including progressive multifocal leukoencephalopathy and cardiac complications. Unfortunately, the risk-benefit profile of these new oral therapies have restricted their use in clinical practice, and appear to require careful consideration in patient selection and monitoring.[14]

Cost of MS

In Canada, the mean total cost per MS patient per year is Can $37,672. The cost of treatment with MS therapies represents 33% of the total costs. Furthermore, the cost due to patients' sick leave and retirement due to MS comprises 32% of total costs. Direct and indirect costs increase with disease severity, due to patients' need for increased medical and non-medical services.[27;28] For patients with mild disability (EDSS score 0-3), the mean cost per patient per year is estimated at Can $30,836, for patients with moderate disability (EDSS 4-6.5) it is estimated at Can $46,622, and for patients with severe disability (EDSS score 7-9) it is estimated at Can $77,981. Relapses contribute an additional economic burden of Can $10,512 per patient per year.[28]

Measurement of health-related quality of life

Assessing health-related quality of life (HRQL) has moved to the forefront of clinical research and is considered a crucial endpoint of clinical interventions. HRQL, in effect, reveals the patients' perspective on health and well-being. It fits well with the World Health Organization's definition of health, which states that health is "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity".[29] Published papers often use the terms quality of life (QOL) and HRQL interchangeably, despite certain differences between the two.[5] HRQL is distinguished from global QOL by those aspects of life that are most likely to be affected by health.[30-32] Domains that are outside of the purview of the health care system such as job satisfaction, quality of housing, and the neighborhood in which one lives, are not included in HRQL.[30] Physical function, social engagement, and emotional/mental health are all domains of HRQL.[30-32]

One approach to assessing HRQL is through the use of health profiles.[33;34] Health profiles are analyzed by sub-scale, where each sub-scale represents a domain of HRQL. The most widely used existing health profile is the SF-36 Health Survey.[35;36] The SF-36 is comprised of 36 items that can be divided into 8 domains. Each domain is scored on a scale from 0 to 100, with higher scores being representative of better functioning and well-being. Health profiles do not provide information on the relative importance attached to each domain.[37;38] As a result, the domains cannot be combined into an overall score, and a trade-off cannot be made between them when evaluating the effectiveness of interventions. For example, if a treatment has a positive effect on physical health but a negative one on mental health, unless we know the relative importance attached to each domain, it is impossible to determine whether the intervention resulted in a net improvement or decline in HRQL.[38;39]

Preference-based measures, on the other hand, do attach a value to each health state described.[34] Not only do these measures provide descriptive information on the various dimensions of health, but also provide a value for each one. They have the advantage of leading to a single number that balances gains in one domain against losses in another. When linked to life-expectancy, they provide measures of quality adjusted life years (QALY) and are used to make decisions about the cost-effectiveness of interventions.[40]

The Quality Adjusted Life Year (QALY)

The Quality-Adjusted Life Year (QALY) is a single comprehensive measure of health improvement that captures the effect of an intervention on both mortality (quantity of life) and morbidity (quality of life).[33] The QALY is a generic measure that can be used to compare the effectiveness of different interventions.[38] Furthermore, if the costs of the interventions are known QALYs can be used to calculate cost-utility ratios.[41] A cost-utility ratio is the difference between the costs of two interventions divided by the difference in the QALYs they produce. The assumption with cost-utility analysis is that, all else being equal, the program with the lowest cost per QALY is favored, because it produces the greatest health benefit to the community for the lowest cost.[42]

There are generally two methods of obtaining a value for the 'Q' in the QALY: direct and indirect.[40] Direct methods involve asking patients or the public to value health states using a standard valuation or preference elicitation technique (e.g. the standard gamble) whereas, the indirect method, involves asking patients to complete a preference-based measure (e.g. the EQ-5D). In the next section, we will first review the direct methods of estimating the 'Q' in the QALY, followed by a review of the indirect methods (i.e. preference-based measures).

Measuring the 'Q' in the QALY: Direct Methods

Standard gamble

The standard gamble (SG) is a classical method of measuring preferences, based on the axioms of expected utility theory.[43] It is the only available technique that measures preferences under conditions of both risk and uncertainty.[38] With the SG, respondents are presented with a given health state, and are asked to consider whether they would prefer to remain in that health state for the rest of their life or take a chance with a new (imaginary) treatment. They are told that the new treatment has the ability to return them to perfect health immediately but also has the ability to cause instant death. The probability of returning to full health on taking the new treatment is gradually decreased (and the chance of death increased) until the patient decides to remain in their current health state. The indifference point represents the value that the patient places on that health state.[43;44]

The SG has been shown to provide higher preference values than the other two commonly used techniques, the Time Trade Off (TTO) and the Rating Scale (RS).[45] This is probably because SG scores embody risk preferences, whereas the TTO and VAS do not.[38;46-48] As risk of death is highly undesirable respondents may stop the gambling sooner, resulting in a high utility value.[38;48]

Time trade-off

The time trade-off (TTO) is a choice-based technique developed specifically for use in health care[40] as a less complex alternative to the SG.[38] Similar to the SG, the TTO method presents the subject with a choice.[49] However, contrary to the SG where the subject is asked to choose between a certain outcome and a gamble, the TTO asks the subject to choose between two alternatives of certainty. The subject is presented with two alternatives – alternative 1: living for period t in a specified but less than perfect health state; or alternative 2: perfect health for time period x where $x<t$. The length of time in perfect health (x) is varied until the subject is indifferent between the two alternatives. The preference value for the less than perfect health state is determined by: x/t.[38;49]

An underlying assumption of the TTO technique is that individuals' trade-off a constant proportion of their remaining life years to improve their health status, regardless of the number of years that remain.[38] However, studies have shown that this assumption may not always hold true, as a person's decision to trade-off time may be influenced by his/her life expectancy.[50]

Rating scale

The rating scale (RS) is based on psychometric or measurement theory.[51] It consists of a vertical line with numerical and verbal descriptors at each end. The RS is intended to have interval properties and is labeled from 0 to 100. The endpoints can be “most desirable” or “best imaginable health state” at the top end, and “death” or “worst imaginable health state” at the lower end. The subject is provided with a set of health states to value and is asked to place each health state on the RS. The distance between the placement of health states should correspond to the subject’s understanding about the relative differences between the health states.[44] If “death” is identified as the worst state and is placed at the 0 end of the scale, then preferences are simply equal to the scale value given to each health state. If death is not identified as the worst state but is placed on some intermediate point on the scale (*d*), then preferences are measured as: $(x-d)/(1-d)$, where x is the rating given to a health state and d is the rating given to death.[43]

Research has demonstrated that the RS is simple and easy to use.[38;40] In surveys, it has demonstrated high response rate and high levels of completion.[40] Also, the RS is cheaper and less time-consuming than the other health state valuation techniques (i.e. the SG and TTO).[38] Studies that have compared the three techniques (SG, TTO and RS) have reported that, for the same health state, scores obtained using the RS are lower than those from the SG and TTO.[38;40]

Measuring the ‘Q’ in the QALY: Indirect Methods

The indirect approach, involves patients completing a preference-based measure (also known as a utility measure) of HRQL.[40] Following the completion of the preference-based measure, responses are converted to health indices using preference weights that have been previously obtained from a random sample of the general population.[33] In clinical trials, indirect methods have the advantage of avoiding the laborious work of valuing a series of health states each time a study is carried out.[38;40]

Existing preference-based measures are all generic in nature. The following section will provide a review of these preference-based measures in order of their development.

Quality of well-being scale

The Quality of Well Being (QWB) is comprised of three dimensions (mobility, physical activity, social activity), with 3-5 levels each and a list of 27 symptoms, describing a total of 1215 states.[52] The dimension of mental health is not included in the scale. The values for the QWB have been elicited using the RS on a random sample of the general population in San Diego, CA, USA.[52;53] The scoring function is linear additive, as the three dimensions are assumed to be independent.[52] The QWB requires interviewer administration[53] and takes between 15-35 minutes.[54] A newer version that can be self-administered has been developed.[38;42]

Health Utilities Index

There are three versions of the Health Utilities Index (HUI). The first HUI (HUI1) was developed by Torrance et al.[55] in 1982 for use in neonatal intensive care. This version was later modified to produce the HUI2 for use in survivors of childhood cancer.[56] HUI2 describes 24000 health states and consists of 7 dimensions: sensation (vision, hearing and speech), mobility, emotion, cognition, self-care, pain and fertility.[57] Health state preferences were elicited using the RS and SG from a random sample of parents in the general population in Hamilton, ON, Canada. The HUI2 scoring function uses a multiplicative functional form. Later, the HUI3 was developed for use in population health surveys in Canada.[58] The HUI3 includes 8 dimensions: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. Each dimension has 5 or 6 levels, describing a total of 972 000 different health states. Health state values were obtained using the RS with four marker multi-attribute states obtained using the SG. The RS scores were then transformed to SG scores, based on the best fitting values of the corner health states. Similar to the HUI2, a multiplicative function combines the dimensions into an index.

15D

The 15D was developed in Finland based on an evaluation of official Finnish health documents and the World Health Organization's definition of health.[59;60] It consists of 15 dimensions: mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental function, discomfort and symptoms, depression, distress, vitality and sexual function. Each dimension has 5 levels. Valuations were elicited from the Finnish population using the RS and combined using an additive model. The 15D is self-administered and takes about 5 to 10 minutes to complete.[59;60]

EQ-5D

The EQ-5D was developed by a multidisciplinary team of researchers in Europe.[61;62] The EQ-5D is comprised of 2 components. The first consists of 5 dimensions: mobility, self-care, usual activities, pain and anxiety/depression. Each dimension has 3 response options, providing a total of 243 unique health states. The second consists of a RS on which respondents provide a rating of their current health status. The EQ-5D is self-administered and takes 1-2 minutes to complete. Health state values were first obtained in the United Kingdom from a nationally representative community sample, using the TTO and RS techniques for 42 marker health states. Because some states in the EQ-5D were considered by respondents to be worse than dead, the lower boundary of the scale is -0.59. The EQ-5D uses statistical modeling (a modified additive model) to combine item responses into an index.[61;62] Health state valuations for the EQ-5D have been conducted in other countries as well,[63] with the most recent one being in the USA.[64;65] Values obtained from the US survey were found to be different from those obtained from the United Kingdom, indicating that country-specific values should be used when possible.[65]

The Assessment of Quality of Life

The first version of the Assessment of Quality of Life (AQOL Mark 1) was developed in 1997 by Richardson and Hawthorne.[66] A newer and revised version (AQOL Mark 2) was published by the same authors in 2004.[67] The AQOL2 has 6 dimensions: independent living, social relationships, mental health, coping, pain, and sensory perception. Each dimension has a number of items, with each item having four or more response levels. The measure is self-administered and takes 5 to 10 minutes to complete. Health state valuations were obtained using the TTO technique in a random sample of the population in Victoria, Australia. A two-stage multiplicative model is used: the first to combine items within each dimension, and the second to combine each dimension into the utility index.

SF-6D

The SF-6D was derived from the SF-36 by Brazier et al.[68] and includes 6 dimensions: physical function, role limitation (a combination of role physical and role emotional), social function, bodily pain, mental health and vitality. Each dimension has 4 to 6 response levels, describing a total of 18 000 health states. Health state values were obtained from a random sample of the UK general population, using the SG technique. A total of 249 states were valued, where each respondent was asked to value 6 states. Random effects regression methods were used to combine scores into a single index.

Modeling health state valuation data

In order to develop a preference-based measure, one must assign a value to each health state described by the measure using one or more of the preference elicitation techniques explained earlier (SG, TTO or VAS).

Preference-based measures can generate hundreds and often thousands of health states.[38] This is because a preference-based measure will provide n^i unique health states (n = number of response levels and i = number of items). In other words, the number of potential health states grows rapidly with an increase in the number of items or response levels e.g. an instrument with 2 response levels in each of the 3 items/dimensions generates 8 (2^3) health states, while one with 6 items, each with 4 levels generates 4096 (4^6) states.[69] It is simply not practical to elicit direct valuations for all of the health states described by a preference-based measure.[38] As a result, the typical procedure used when developing a preference-based measure is to value a *subset* of health states, and then combine them in a multi-attribute utility function (i.e. scoring algorithm) to calculate a value for all possible health states in the classification system.

Multi-attribute utility theory

Multi-attribute utility theory (MAUT) is an approach used to estimate values for all possible health states in a classification system.[38] The HUI2 and HUI3 used MAUT to identify the appropriate multi-attribute utility function (MAUF).[70] The MAUF can be additive, multiplicative or multi-linear.[58] First, each dimension is valued separately to estimate single dimension utility functions. An example is, being ‘Able to walk around the neighborhood with walking equipment, but without the help of the other person’ (a single dimensional health state on the ambulation dimension).[38] Second, corner states are valued - a corner state is a multidimensional health state in which all items are described by their best level while one item is set at its worst level. For example, the corner state for the speech dimension may be: I can hardly be understood by anyone when I speak but I can walk in the community as I desire, go up and down several flights of stairs, and drive a car anywhere.[39] Using one of the MAUFs, weights are calculated for each possible health state described by the classification system.

The other approach uses statistical modeling to estimate a function (i.e. scoring algorithm) for all possible health states in a classification system. The EQ-5D and the SF-6D have used this approach to model health states. A difference between MAUT and statistical modeling is that the former has a strong theoretical foundation in decision theory, whereas the latter does not.[38] The absence of a theory means that there is little guidance when selecting the health states that need to be directly valued with statistical modeling. Therefore, the PBMSI will be developed using MAUT, as it is based on a strong theoretical foundation, and provides explicit guidance on the selection of health states that need to be directly valued.

Whose preferences?

Generic preference-based measures, such as the EQ-5D and the HUI, have been developed using preferences obtained from the general population. However, in recent years there has been debate as to whether preferences should be obtained from the public or from patients.[38;71] The challenge with using generic preference-based measures in clinical practice and research is that the valuations represent social preferences of the general population rather than representing patients with the disease.[72] The main argument for the use of general population values is that it is society that pays for the service, and thus they should be the ones involved in health care decision making.[38] Advocates for the use of patient preferences argue that patients know their health states better than anyone trying to imagine it.[73] Contrary to the National Institute for Health and Clinical Excellence (NICE) guidelines (for economic evaluations) that require the use of the general public for valuing health states, the U.S. Food and Drug Administration (FDA) guidelines for the use of patient-reported outcomes require the direct involvement of patients.[74] A number of studies have demonstrated that patients tend to value health states higher than members of the general population.[75-79] A recent study compared health valuations between self-ratings and ratings of corresponding health state profiles by members of the general population not experiencing those states.[80] The author pooled data from several different UK sources yielding a total of 23,679 useable observations. 139 unique EQ-5D health states were identified in the dataset and the mean RS rating was calculated for each of the states. When he compared these self-rated health states with the standard UK TTO utility weights, the self-rated values were significantly higher than those based on social preferences.[80] These results are consistent with a previous study by Insinga and Fryback,[81] where the authors found differences ranging from 73-275% between respondents' own self-rated values and those estimated from the general population. McPherson et al.[82]

compared the level of agreement between the general public's rating of health states against patients with a chronic disabling disease (namely rheumatoid arthritis, stroke or multiple sclerosis). The authors found that there were significant discrepancies in ratings between patients and the public, suggesting that "there is a fundamental difference in how people with disability experience life (and health) as compared with nondisabled people."[82]

In 2010, Peeters and Stiggelbout[83] conducted a meta-analysis of all studies (n=30) that compared patient and non-patient (defined as general public, professionals, or proxies) preference values on the SG, TTO and RS. Their results revealed that patients gave higher valuations than non-patients on the TTO (difference = 0.05 points, $p<0.05$) and RS (difference = 0.04 points, $p<0.01$), but not on the SG (difference = 0.01, $p>0.05$).

CHAPTER 2: RATIONALE AND OBJECTIVES

Rationale of the thesis

Multiple Sclerosis (MS) is an inflammatory, demyelinating disease of the central nervous system (CNS), with wide ranging effects on function, health, and quality of life. From a clinical perspective, the most widely used measure is the Expanded Disability Status Scale (EDSS), which is a single-item disability classification scale used by MS neurologists to quantify disability. It is known to have a number of psychometric limitations.[84] From the patient's perspective, a number of MS specific and generic health indices have been used with the most common being the generic SF-36.[36;85] The only measures which yield one value for quantifying the overall health impact of MS are generic preference-based measures. However, the challenge with using such generic preference-based measures in people with MS is that they may not capture all domains of health relevant to the disease either as benefits or harms.

Objectives

Therefore, the overall objective of this PhD thesis is to take important steps towards developing a Preference-Based Multiple Sclerosis Index (PBMSI) for use as a global outcome in clinical and cost-effectiveness studies for MS.

To operationalize this global objective, a series of specific objectives were developed towards the manuscripts that formed this thesis.

1. A systematic review to estimate the extent to which existing interventions improve health-related quality of life in persons with MS.

   Manuscript 1: *The effects of clinical interventions on health-related quality of life in multiple sclerosis: a meta-analysis*

2. To estimate the extent to which generic preference-based measures capture domains that are important to the quality of life of people with MS.

   Manuscript 2: *Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis?*

3. To summarize the evidence from published literature on the psychometric properties of generic utility measures in MS.

Manuscript 3: *A review of the psychometric properties of generic utility measures in multiple sclerosis*.

4. To describe the development of a prototype Preference-Based Multiple Sclerosis Index (P-PBMSI). The specific objectives were: (i) to identify items best reflecting the domains of quality of life important to people with MS; and (ii) to provide evidence for the discriminative capacity of the response options by cross walking onto a visual analogue scale (VAS) of health rating.

   Manuscript 4: *Using existing data to identify candidate items for a health state classification system in multiple sclerosis.*

5. Using expert and patient feedback, to qualitatively revise the items selected for inclusion in the Preference-Based Multiple Sclerosis Index (PBMSI).

   Manuscript 5: *The development of a bilingual MS-specific health classification system: the Preference-Based Multiple Sclerosis Index (PBMSI).*

6. To contribute preliminary evidence towards the similarities and differences in the Standard Gamble and the Rating Scale to reflect patient preferences for the different items in the PBMSI, where contrasts were on absolute and utility values, level of difficulty, and discriminative ability. ***Manuscript 6:*** *Developing a valuation function for a multiple sclerosis specific classification system: comparison of standard gamble and rating scale.*

CHAPTER 3 (MANUSCRIPT 1)

The effects of clinical interventions on health-related quality of life in multiple sclerosis: a meta-analysis

Ayse Kuspinar[1], Ana Maria Rodriguez[1] and Nancy E. Mayo[1,2]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada

[2]Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

Published in *Multiple Sclerosis Journal* (2012) 18:1686-1704.

Communication addressed to:

Ayse Kuspinar, M.Sc., Ph.D. Candidate
School of Physical & Occupational Therapy
Faculty of Medicine, McGill University
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
Tel: 514-934-1934 ext 31564
E-mail: ayse.kuspinar@mail.mcgill.ca

MULTIPLE SCLEROSIS JOURNAL | MSJ

Multiple Sclerosis Journal
18(12) 1686–1704
© The Author(s) 2012
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1352458512445201
msj.sagepub.com
SAGE

The effects of clinical interventions on health-related quality of life in multiple sclerosis: a meta-analysis

Ayse Kuspinar[1], Ana Maria Rodriguez[1] and Nancy E Mayo[1,2]

Abstract

The objective is to estimate the extent to which existing health care interventions designed specifically to target health-related quality of life (HRQL) in persons with multiple sclerosis (MS) achieve this aim. The structured literature search was conducted using multiple electronic databases including Ovid MEDLINE, EMBASE, Cumulative Index to Nursing and Allied Health Literature and the Cochrane Central Register of Controlled Trial, for the years 1960 to 2011. The methodological quality of selected randomized controlled trials (RCTs) was assessed using the Cochrane Collaboration's recommended domain-based method. Effect size (ES) was used to measure the effect of each intervention on HRQL. The studies were combined using a random-effects model to account for inter-study variation. Heterogeneity was tested for using the *I*-test and publication bias was assessed using funnel plots and the Egger weighted regression statistic. Thirty-nine RCTs met the criteria, all with acceptable methodological quality. Six major types of interventions were identified through the search. The smallest effect was observed for self-management and complementary and alternative medicine (ES=0.2), followed by medication (ES=0.3) then cognitive training and exercise (ES=0.4), and psychological interventions to improve mood (ES=0.7). The magnitude of positive effect on HRQL varied between the different types of interventions. The extent to which interventions are able to improve HRQL depends on delivering a potent intervention to those persons who have the potential to benefit.

Keywords

multiple sclerosis, quality of life, review, meta-analysis, treatment outcome, rehabilitation

Date received: 4th January 2012; revised: 12th March 2012; accepted: 21st March 2012

Introduction

Multiple sclerosis (MS) is an unpredictable, inflammatory, demyelinating disease of the central nervous system (CNS).[1] It is the leading cause of neurological disability in young adults,[2] affecting three times as many females as males.[3] The etiology of MS is unknown; however, there is evidence that both genetic and environmental factors are involved in triggering the disease.[1,4]

Assessing health-related quality of life (HRQL) has moved to the forefront of clinical research and is considered the ultimate endpoint of clinical interventions. This is especially true for chronic conditions like MS, as the management of these diseases is rehabilitative or palliative in nature, rather than curative. It is now well-established that persons with MS have significantly lower levels of HRQL as compared with the healthy population.[5] This is also the case even in individuals with mild disease.[6] Persons with MS have reported lower HRQL than that of patients affected by other chronic diseases, such as rheumatoid arthritis[7] and Parkinson's disease.[8] HRQL, in effect, reveals the patients' perspective on health and well-being. It fits well with the World Health Organization's definition of health, which states that health is 'a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity'.[9]

HRQL measurement can be in the form of a single question that simply asks the patient 'How is your quality of life?' However, more commonly it is in the form of a questionnaire

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Canada.
[2]Division of Clinical Epidemiology, Royal Victoria Hospital, Canada.

Corresponding author:
Ayse Kuspinar, School of Physical and Occupational Therapy, McGill University, 3654 Promenade Sir-William-Osler, Montreal, Quebec, H3G 1Y5 Canada
Email: ayse.kuspinar@mail.mcgill.ca

made up of a series of items or questions. There are two types of HRQL questionnaires: (i) health profiles and (ii) utility/preference-based measures.[10] Health profiles yield separate values for the different domains measured.[11] For example, the well-known and widely used Short Form 36 Health Survey (SF-36) has two component summary scales: physical (PCS) and mental health (MCS). Health profiles can be generic, such as the SF-36 and the Sickness Impact Profile (SIP), or they can be disease-specific like the MS Quality of Life-54 (MSQOL-54) and the Multiple Sclerosis Impact Scale-29 (MSIS-29). Utility or preference-based measures, on the other hand, are created using utility theory and reflect the preferences of patients for the different domains of health.[10] Utility or preference-based measures have the advantage of leading to a single number that balances gains in one domain against losses in another.[12] The best known utility or preference-based measures are the Canadian Health Utilities Index (HUI) and the EQ-5D index.[12]

MS produces a range of unpleasant and debilitating symptoms, including fatigue, muscle weakness, sensory problems, loss of memory and concentration, bowel and bladder problems, to name a few. These can have a profound impact on daily functioning, relationships, social and leisure activities, which in turn may lead to reduced HRQL.

A vast range of interventions have been developed in the field of MS that are targeted to improve HRQL, including exercise, cognitive therapy, and complementary and alternative medicine. Disease modifying therapies (DMTs) also play an essential part in the treatment of MS; however, these interventions target the disease process itself including disability and relapses, rather than HRQL specifically. HRQL is measured to monitor side effects of DMTs and make decisions about undertaking or prescribing the medication.[2,5] Therefore, this review will not include the evaluation of DMTs on HRQL, but rather will focus on the evaluation of interventions that are targeted to improve HRQL.

Among the vast range of existing treatment options, we have not yet determined which of these treatments may really work for people with MS. The objective of this systematic review is to estimate the extent to which existing health care interventions designed specifically to target HRQL in persons with MS achieve this aim.

Methods

The *Cochrane handbook for systematic reviews of interventions*[13] has been used as a guide to write the systematic review.

Type of study design used

Only parallel group randomized controlled trials (RCTs) were included. Cross-over RCTs were excluded because of the difficulty in separating the intervention effect. Studies published in languages other than English or French and unpublished or grey literature were excluded.

Types of participants

Persons with clinically definite MS were included in the review without restrictions for disease severity, sex, type of MS or the presence of medical co-morbidities. Trials were excluded if participants (1) were younger than 18 years old; (2) had an MS attack one month prior to study entry, or (3) were part of a study with different types of populations (e.g. MS and other neurological disorders).

Types of interventions

All interventions, *except* for DMTs and corticosteroids, that used a HRQL measure as a primary or secondary outcome, were included. Interventions that were of DMTs or corticosteroids were excluded as they target the disease process itself including disability and relapses, rather than HRQL specifically. Furthermore, with such interventions, HRQL outcome measures are administered to monitor the side effects of the drug. No restrictions were made on dose, frequency, intensity or duration of treatment.

No restrictions were made on the type of control group used. Control groups could be inactive (e.g. placebo, no treatment, usual care or a waiting list control), or active (e.g. a different drug or a different kind of therapy).[14]

Types of outcome measures

Studies that measured HRQL as an outcome, ideally using an accepted and validated instrument, were reviewed. The following criteria were set forth:

> Only HRQL instruments for which a single index was available were included. These could be single-items (e.g. global QOL question); a utility/preference-based measure (e.g. EQ-5D, HUI); or a health-profile measure for which a single-domain index was available (e.g. the PCS and MCS scores of the SF-36 health profile or the physical or psychological dimensions of the MSIS-29).
>
> Outcome measures that were developed with a primary purpose of evaluating symptom impact or severity (e.g. Fatigue Severity Index, Hospital Anxiety and Depression Scale) were excluded.
>
> HRQL assessments had to be made by the patient. Assessments that were made by a physician, caregiver or proxy were excluded.
>
> When more than one HRQL measure was used in a study, the one with the largest effect size (ES) was included in the forest plot.
>
> If HRQL was assessed on more than one occasion, only the first occasion after the intervention had been completed was chosen. Follow-up assessments were not included in the analysis.

Search methods for identification of studies

A systematic search of all published literature in scientific journals that used a HRQL measure as an outcome in persons with MS was carried out. Trials were identified by searching the following databases: Ovid MEDLINE (1948 to September 2011), EMBASE (1980 to September 2011), Cumulative Index to Nursing and Allied Health Literature (1960 to September 2011) and the Cochrane Central Register of Controlled Trials (CENTRAL). The following terms were used: Multiple Sclerosis AND (quality of life OR health status OR health-related quality of life OR well-being OR utility OR preference-based OR patient-reported OR health profile OR health survey) AND Limit to Randomized Controlled Trials.

Data collection and extraction

Two authors independently screened all citations and abstracts that were identified in the search. Based on their titles and abstracts, irrelevant articles were disregarded. Afterwards, each author independently evaluated the full texts of potentially eligible studies. Articles that did not meet inclusion criteria were excluded, and the reasons for exclusion documented. Any disagreements on the eligibility of a study were resolved by discussion. By using a data extraction form, each author independently extracted information from the final set of articles that were included in the meta-analysis. Data was extracted on the study population (age, sex, type of MS, etc.), study characteristics (where did the study take place, recruitment period, etc.), intervention characteristics (type, dose, duration, etc.) and outcomes (description of HRQL outcome and estimates of intervention effect).

Assessment of study quality

The quality of the included studies was assessed using the Cochrane risk of bias tool.[15] The tool consists of seven domains: sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective outcome reporting and any other potential sources of bias. Sequence generation refers to the method of randomization used such as random number table or a computer random number generator. Allocation concealment, on the other hand, refers to whether the authors prevented foreknowledge of treatment allocation by, for example, using opaque sealed envelopes.[15] As HRQL is measured by self-report, the outcome assessment is by definition un-blinded and bias can be introduced if participants are aware of the specific hypothesis being tested. As this systematic review was of outcome effects (not intervention effects), the criteria of selective outcome reporting was not considered in the quality rating. Typically, systematic reviews assess the effect of a specific intervention on a few different outcomes. However, in the case of this review, we assessed the effect of a variety of interventions on one specific outcome: HRQL. In other words, we only included studies that reported findings on HRQL, regardless of what the intervention was. Risk of bias from selective outcome reporting occurs when an outcome is mentioned in the publication, but findings are not reported due to lack of clinical or statistical significance.[15] As we only included trials that reported findings on HRQL, risk of bias due to selective outcome reporting was not relevant in the context of this systematic review. Two reviewers independently judged the adequacy of each study in relation to each of the six domains. A judgment of ‘Yes’ indicated low risk of bias, ‘No’ indicated high risk of bias and ‘Unclear’ indicated unclear or unknown risk of bias.[15]

Data analysis

ES was used to measure the effect of each intervention on HRQL, and was calculated using *Hedges' adjusted g*.[16] This involved taking the difference in the mean change in the outcome (pre- and post-intervention) between an intervention and a control group, and dividing it by the initial pooled standard deviation (SD).[17,18] In cases where baseline values were not reported, post-intervention values were used instead. If the SD was not given, the primary authors were contacted for further information. If the authors could not be reached, the SD was estimated from the *p*-value or 95% confidence interval (CI).[19]

In this systematic review, we included only HRQL instruments for which a single index was available. These could be single-items (e.g. global QOL question); a utility/preference-based measure (e.g. EQ-5D, HUI); or a health-profile measure for which a single-domain index was available (e.g. the PCS and MCS scores of the SF-36). If the eight sub-scale scores were provided for the SF-36, SAS9.2 was used to estimate the PCS and MCS scores through the standardized three-step procedure. First, z-scores were calculated by subtracting subscale means for the general US population from the mean subscale scores of the study and dividing the difference by the standard deviation of the US population. Second, the product of the z-scores and the subscale factor score coefficients were taken and summed together. Third, t-scores were calculated by multiplying the PCS and MCS scores by 10 and adding 50 to the product.[20] SD values for the PCS and MCS scores were taken from other articles that used the SF-36 and had a similar sample size.[13]

The first post-intervention time point was used for the systematic review, as it best reflected the effect of the intervention on the outcome. If a study had two intervention groups that were similar, data from both groups was combined to create a single pair-wise comparison against the control group.[19] If the intervention groups were different, then only one group was kept in the analysis. Unadjusted mean values were used in ES calculations, but in instances where only the adjusted values were provided, these were used instead.

The studies were combined using a random-effects model to account for inter-study variation.[21] Statistical analysis was carried out using MIX1.7.

Procedure for contacting study authors

Primary authors of articles were contacted if insufficient data was provided to calculate an ES. The authors were first provided with a description of the systematic review and its specific objectives, followed by questions regarding the missing information. A table was attached in the email for them to fill out and send back to us. Authors were given 10 days to respond. If study authors did not respond to our first email, a subsequent email was sent a week later.

Measure of effect

A positive ES indicated an improvement in the intervention group, and a negative ES indicated an improvement in the control group. Cohen's criteria was used for interpreting magnitude of ES, where an ES of ~0.2 is small, ~0.5 is moderate and ~0.8 is large.[22] An ES was statistically significant if its CI excluded 0 and clinically significant if it was ≥0.50.[23]

Heterogeneity and publication bias

Heterogeneity was tested for using the I^2 statistic. This statistic is the percentage of the total variation across studies that are due to between study heterogeneity, rather than chance[21]. There are two types of heterogeneity: clinical and methodological. Clinical heterogeneity is due to variability in participants, interventions and outcomes. Methodological heterogeneity is due to variability in study design and risk of bias.[21] A p-value of less than 0.10 and $I^2 > 50\%$ was considered as evidence for substantial heterogeneity.[21] Publication bias was assessed through the Egger weighted regression statistic and visual inspection of funnel plots.[24]

Results

Trial flow and study characteristics

A total of 552 potentially relevant articles were identified through the initial search. Duplicates were removed, leaving 335 records for more detailed evaluation. The abstracts and titles of the 335 articles were screened, and 241 studies that did not meet the inclusion criteria were removed. A total of 94 articles were left for full-text reviews. After each study underwent a full text review, a total of 50 studies were excluded for the following reasons: a) no HRQL instrument used in study (n=26), b) study design was not a parallel group RCT (n=10), c) population not exclusive to MS (n=3), d) duplicate data (n=3), e) study did not target HRQL (n=2), f) trial protocol (n=2), g) study involved patients with relapses (n=1), h) results not presented by group (n=1), i) cross-over design (n=1) and j) incorrect scoring of HRQL measure (n=1). Figure 1 provides details of the study selection process.

Forty-four articles were left for data extraction, of which 16 studies needed further clarification in regards to the mean and SD values. Primary authors were contacted, and several responded that the data would be difficult to retrieve as the study was conducted several years ago.

Among the 16 trials, original data was obtained from primary authors for three of them. For two studies, the authors reported that there was no between group difference, and therefore were given an ES of 0. For six studies, the PCS and MCS scores were estimated from the eight subscale scores using the standard scoring algorithm for the SF-36 (see data analysis for details). Five studies were excluded from the systematic review because an ES could not be calculated. Reasons for exclusion and descriptive information regarding these studies are provided in the supplementary material (Supplementary Tables 1 and 2).

Thirty-nine trials were left for inclusion in the systematic review. Thirteen studies were published in North America, 25 were published in Europe and one in Australia. Table 1 provides a detailed description of the 39 studies included in the systematic review.[25–63]

Collectively, the 39 trials included a total of 2952 persons with MS. Sample size of studies ranged from five to 133 per group, and mean age of participants ranged from 33 to 51 years. The Expanded Disability Status Scale (EDSS) scores of subjects varied from 0 (no disability) to 9.5 (bedridden).

Outcomes

HRQL was a primary outcome for 13[25-36] of the 39 trials. Therefore, there were 26 trials where HRQL was measured as a secondary outcome. To evaluate consistency between the primary outcome findings and HRQL, concordance between the two outcomes was evaluated. Whether or not an intervention had a clinically significant effect (ES≥0.5) on the primary outcome measure and whether this same effect was observed for HRQL was investigated. There was 73% (n=19) agreement between the primary findings and HRQL.

Table 2 lists the HRQL measures that were used in the included trials, whether or not they were rescored so that higher numbers were indicative of better HRQL, and the number of times they were used.

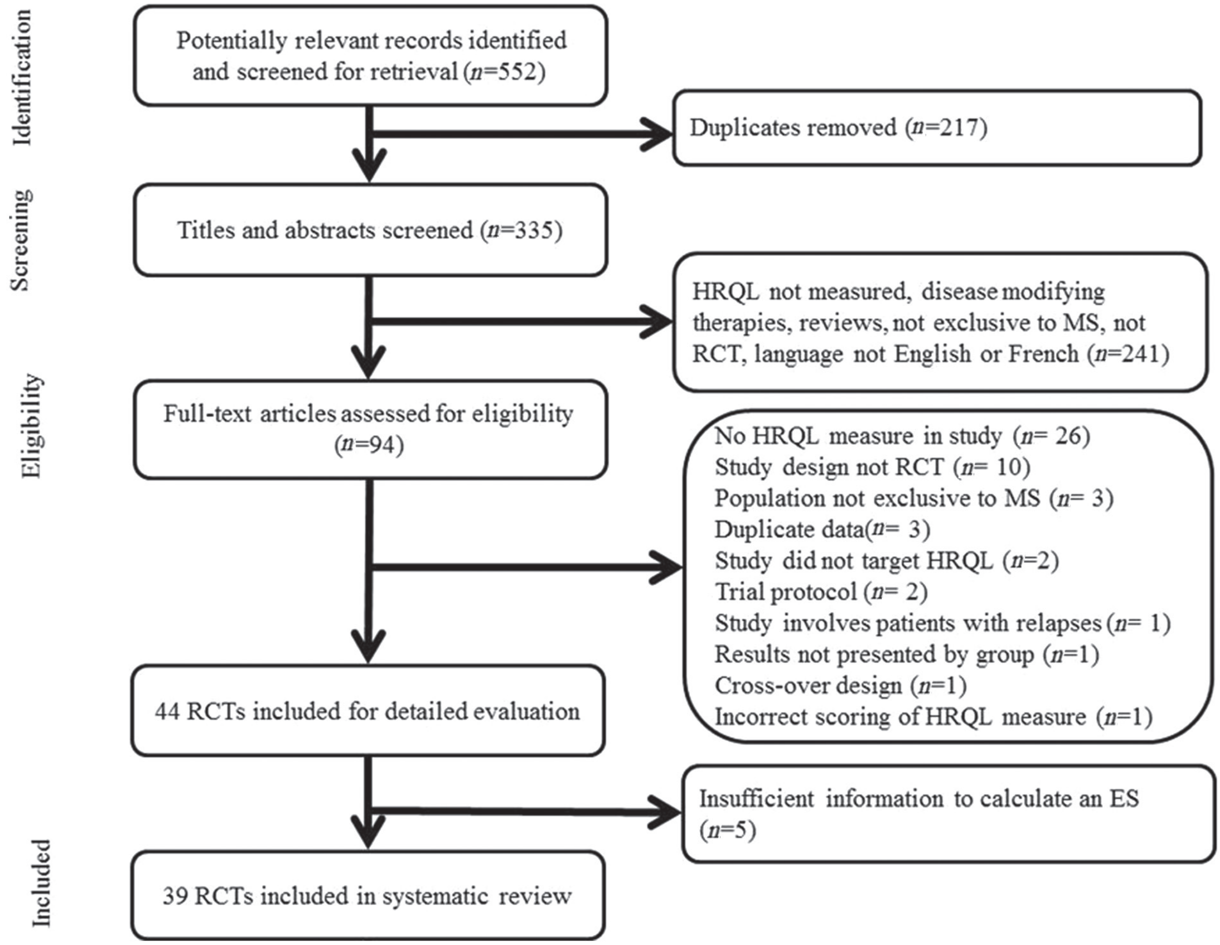

Identification
Screening
Eligibility
Included
Potentially relevant records identified and screened for retrieval (n=552)
Duplicates removed (n=217)
Titles and abstracts screened (n=335)
HRQL not measured, disease modifying therapies, reviews, not exclusive to MS, not RCT, language not English or French (n=241)
Full-text articles assessed for eligibility (n=94)
No HRQL measure in study (n= 26)
Study design not RCT (n= 10)
Population not exclusive to MS (n= 3)
Duplicate data(n= 3)
Study did not target HRQL (n=2)
Trial protocol (n= 2)
Study involves patients with relapses (n= 1)
Results not presented by group (n=1)
Cross-over design (n=1)
Incorrect scoring of HRQL measure (n=1)
44 RCTs included for detailed evaluation
Insufficient information to calculate an ES (n=5)
39 RCTs included in systematic review

Figure 1. Flow diagram to illustrate process of study selection.

Interventions

Six major types of interventions were identified through the search: a) complementary and alternative medicine (*n*=7), b) self-management or self-efficacy (*n*=7), c) exercise/rehabilitation (*n*=13), d) cognitive training (*n*=3), e) medication for symptom management (*n*=6), and f) psychological interventions for mood (*n*=3).

Effect of complementary and alternative medicine on HRQL. There were seven studies[26,27,29,37–40] that involved complementary and alternative medicine, and the pooled ES was 0.16 (95% CI -0.06 to 0.40, *p*=0.19). There was no heterogeneity among the studies (I^2=0%, *p*=0.97) (Figure 2(a)). Risk of bias due to sequence generation and concealment of allocation was either low[29,37–40] or unclear.[26,27,39,40] There was blinding of participants and outcome assessment in all of the trials except for two[27,40] (20%). There were no trials (0%) that were at high risk of bias due to incomplete outcome data (Table 3). Visual inspection of the funnel plots and the Egger weighted regression statistic indicated that there was no publication bias (*p*=0.50).

Effect of self-management on HRQL. There were seven studies[25,33,34,41–44] that evaluated the effect of a self-management and self-efficacy program on HRQL. The pooled ES was 0.24 (95% CI 0.10 to 0.38, *p*<0.01) (Figure 2(b)). Heterogeneity among the studies was null and non-significant (I^2=0.0%, *p*=0.46). As it is very difficult to blind participants in behavioral interventions, all six studies (100%) were at high risk of bias from blinding. As HRQL was measured by self-report, blinding of outcome assessment was also judged at high risk of bias in all studies (100%). Two studies[25,34] (28%) were found to be at high risk of bias for incomplete outcome data (Table 3). The funnel plots and the Egger weighted regression statistic indicated the absence of publication bias (*p*=0.20).

Effect of medication on HRQL. There were six different types of medication targeting symptom management in MS: levetiracetam for neuropathic pain,[45] dextromethorphan/quinidine for pseudobulbar affect,[46] paroxetine for depression,[47] sativex for overactive bladder,[48] modafinil for fatigue[49] and memantine for cognitive impairment.[50] The

Table 1. Descriptive information for each study included in the systematic review (grouped according to type of intervention).

	First author (year)	Sample size (I) Intervention group (C) Control group	Age Mean ± SD	EDSS Mean ± SD	Intervention	Duration and frequency	Control	Primary outcome of study (Effect size)	HRQL measure used in study (Effect size)	Data extracted
Complementary and alternative medicine	McClurg (2011)[40]	15 (I) 15 (C)	52.4 ± 12.3 59.3 ± 14.7	2.0 ± 1.0 3.0 ± 1.0	Abdominal massage	15-min, 1×/ week, 4 weeks[a] Follow-up at 8 weeks.	Bowel management advice	Constipation questionnaire (ES 1.16)	MSIS-29 Physical and Psychological (ES 0)	No significant change in scores at any time point
	Hughes (2009)[39]	35 (I) 36 (C)	50.0 ± 11.1 53.0 ± 11.0	5.8 ± 0.95 6.2 ± 0.8	Reflexology	45-min, 1×/ week, 10 weeks[e] Follow-up at 16 and 22 weeks	Placebo	Pain (ES 0)	MSIS-29 Physical and Psychological[a](ES 0.27)[g]	Median and IQR
	Shinto (2008)[27]	15 (I) 15 (C1) 15 (C2)	43.5 ± 9.2 (I and C)	2.5 ± 1.1 (I and C)	Naturopathic medicine (vitamins, fish-oil, diet intervention)	Eight visits with the naturopath Assessment at 6 months	C1: usual care C2: education sessions about MS[f]	HRQL (ES 0.15)	SF-36 PCS and MCS (ES 0.15)	Mean ± SD (of the change)
	Johnson (2006)[29]	12 (I) 11 (C)	48.7 ± 11 52.8 ± 9.5	3.5 ± 2.1 3.9 ± 2.6	Ginkgo extract (EGb 761)	Four 60 mg tablets per day, 4 weeks	Placebo	HRQL (ES 0.06)	FAMS (ES 0.06)	Mean ± SD
	Warke (2006)[38]	30 (I1) 30 (I2) 30 (C)	45.6 ± 9.3 47.8 ± 12.0 48.7 ± 14.2	Not presented	I1: low-frequency TENS I2: high-frequency TENS[f]	45-min, minimum 2×/ day, 6 weeks[e] Follow-up at weeks 10 and 32	Placebo	Pain (ES 0.58)	MSQOL-54 PCS and MCS[a] (ES 0.05)[g]	Mean ± SEM
	Weinstock (2005)[26]	13 (I) 14 (C)	39.9 ± 10.0 45.1 ± 7.7	2.0 ± 1.3 1.9 ± 0.6	Fish oil (ω-3) and low-fat diet (<15%)	6 capsules of 1 g fish oil/ day. Regular meetings with dietician for 12 months	Olive oil (placebo) and low-fat diet (<30%)	HRQL (ES 0.22)	SF-36 PCS (ES 0.22)	Mean ± SD (from graph)
	Al-Smadi (2003)[37]	5 (I1) 5 (I2) 5 (C)	34–65 (range only)	Not presented	I1: low-frequency TENS I2: high-frequency TENS[f]	45-min, minimum 3×/ day, 6 weeks[e] Follow-up at week 10	Placebo TENS	Pain (ES 0.99)	SF-36 PCS and MCS Leeds MSQOL[a] (ES 0.60)[g]	Mean ± SEM
	Miller (2011)[25]	102 (I) 104 (C)	48.1 ± 9.1 48.1 ± 9.7	Not presented	Secure web-based messaging and self-management of MS symptoms	12 months	Secure web-based messaging	HRQL (ES -0.04)	SIP[a] EQ-5D index (ES -0.04)[g]	Mean ± SD

Table 1. (Continued)

	First author (year)	Sample size (I) Intervention group (C) Control group	Age Mean ± SD	EDSS Mean ± SD	Intervention	Duration and frequency	Control	Primary outcome of study (Effect size)	HRQL measure used in study (Effect size)	Data extracted
Self-management/ self-efficacy	Barlow (2009)[42]	78 (I) 64 (C)	48.2 ± 10.1 50.7 ± 11.7	Not presented	Chronic Disease Self-Management Course	Weekly 2-h session, 6 sessions, 6 weeks. Assessment at 4-months[e] Follow-up at 12-months	Wait-list	Self-efficacy (ES 0.37) Depression (ES 0.23)	MSIS-29 Physical[a] and Psychological (ES 0.25)[g]	Mean ± SD
	Bombardier (2008)[43]	70 (I) 60 (C)	47.5 (41–54) 45.0 (40.5–52.0)	Not presented	Telephone counseling for health promotion	60 to 90-min initial interview, and 30-min telephone counseling at weeks 1,2,4,8, and 12	Waitlist	Health promotion behaviors (ES 0.5)	SF-36 PCS and MCS[a] (ES 0.33)[g]	Median and IQR
	McAuley (2007)[41]	13 (I) 13 (C)	43.5 ± 7.6 (I and C)	Not presented	Workshops to enhance self-efficacy relative to physical activity + exercise program	Bi-weekly workshops incorporated into 3-month exercise program	Workshop on general health-related topics + exercise program	Exercise adherence (ES 0.47)	SF-12 PCS and MCS SWLS[a] (ES 0.44)	Mean ± SD
	Ennis (2006)[44]	31 (I) 30 (C)	45.0 ± 9.0 46.0 ± 8.0	1.0–7.0 (range)	Health promotion education	3-h, 1×/week, 8 weeks	Usual care	Health promoting behaviors (ES 0.95)	SF-36 PCS and MCS[a] (ES 0.49)	PCS and MCS estimated from 8 subscale scores
	Stuifbergen (2003)[33]	56 (I) 57 (C)	45.8 ± 10.1 (I + C)	Not presented	Wellness intervention program for women	90-min, 1×/ week, 8 weeks[a] Follow-up at 3 and 8 months.	Waitlist	Health promoting behaviors (ES 0.32) HRQL (ES 0.44)	SF-36 PCS and MCS[a] (ES0.44)	PCS and MCS estimated from 8 subscale scores

(Continued)

Table 1. (Continued)

	First author (year)	Sample size (I) Intervention group (C) Control group	Age Mean ± SD	EDSS Mean ± SD	Intervention	Duration and frequency	Control	Primary outcome of study (Effect size)	HRQL measure used in study (Effect size)	Data extracted
	O'Hara (2001)[34]	73 (I) 96 (C)	52.5 ± 11.2 50.4 ± 10.4	Not presented	Self-management program	1–2h, 2×/month, 1month. Evaluation at 6 months.	No intervention.	Mobility (ES -0.03) Daily activities (ES 0.12) HRQL (ES 0.17)	SF-36 PCS and MCS[a] (ES0.17)	PCS and MCS estimated from 8 subscale scores
Medication for symptom management	Moller (2011)[49]	62 (I) 59 (C)	41.4 ± 9.5 40.8 ± 11.2	3.5 ± 1.4 3.1 ± 1.4	Modafinil (wakefulness-promoting artificial psycho-stimulant)	Up to 200 mg/day per week, 2 months	Placebo	Fatigue (ES 0.50)	HAQUAMS (ES -0.08)	Mean ± SD
	Kavia (2010)[48]	63 (I) 67 (C)	48.6 ± 9.3 46.8 ± 11.2	Not presented	Sativex (for overactive bladder)	100 ml, 8 actuations in 3-h period and 48 actuations in 24-h period, 2 months	Placebo	Number of incontinence episodes (ES 0.12)	Incontinence Quality of Life Questionnaire (ES 0.26)	Mean, *p*-value
	Lovera (2010)[50]	54 (I) 60 (C)	50.5 ± 8.2 50.4 ± 7.7	4.5 ± 2.2 4.4 ± 1.9	Memantine	10 mg, 2×/day, 12 weeks	Placebo	Cognitive function (ES 0)	SF-36 (ES 0)	No data presented, but authors stated no significant difference between groups
	Ehde (2008)[47]	42 (I and C)	45 ± 10.1 (I and C)	EDSS 0-4 (*n*=22) EDSS 4.5-6.5 (*n*=16) EDSS 7-9.5 (*n*=4)	Paroxetine (for depression)	Started at an initial dose of 10 mg/day, titrated up to 40 mg daily, 12 weeks	Placebo	Depressive symptoms (ES 0.05)	SF-36 PCS and MCS[a] (ES 0.54)[g]	Mean ± SD
	Rossi (2008)[45]	12 (I) 8 (C)	40.4 ± 6.3 33.2 ± 9.3	2.6 ± 1.5 2.3 ± 1.1	Levetiracetam (for neuropathic pain)	500 mg, 3 months	Placebo	Pain (ES 1.54)	MSQOL-54 Item on global QOL (ES 0.59)	Mean ± SD (from graph)
	Panitch (2006)[46]	76 (I) 74 (C)	46.3 ± 9.8 43.7 ± 10.0	Not presented	Dextromorphan and quinidine (for pseudobulbar affect)	30 mg 2×/day, 85 days	Placebo	Emotional lability (ES 0.87)	VAS for global QOL (ES 0.61)	Mean ± SD (adjusted mean)

Table 1. (Continued)

	First author (year)	Sample size (I) Intervention group (C) Control group	Age Mean ± SD	EDSS Mean ± SD	Intervention	Duration and frequency	Control	Primary outcome of study (Effect size)	HRQL measure used in study (Effect size)	Data extracted
Cognitive training	Hildebrandt (2007)[52]	17 (I) 25 (C)	42.4 (25–55) 36.5 (23–63)	2.9 (Range 1-7) 2.7 (Range 0-7)	Computer based cognitive training	30 min/day, 5×/ week, 6 weeks	No intervention	Memory (ES 0.48)	SF-12 PCS and MCS[a] (ES 0.27)[g]	Mean ± SD
	Solari (2004)[53]	40 (I) 37 (C)	46.2 ± 9.2 41.2 ± 10.6	3.0 (1.5–7.0)[d] 4.0 (1.5–6.5)[d]	Computer-aided memory and attention retraining	45-min, 2×/ week, 8 weeks[e] Follow-up at 16 weeks.	Computer-aided visuo-constructional and visuo-motor retraining (sham)	Cognitive function (ES 0.65 on word list generation)	MSQOL-54 PCS and MCS[a] (ES 0.95)[g]	Mean ± SD
	Lincoln (2002)[51]	77 (I1) 71 (I2) 77 (C)	43.0 43.0 40.5 (Median only)	18.0 16.0 16.0[c]	I1: cognitive rehabilitation[a] I2: cognitive assessment (results sent to patients' doctor)	Up to 6 months	No intervention (results of screening assessments not given to patients or their general practitioner)	Activities of daily living (ES 0.32) Mood (ES 0.17)	SF-36 PCS[a] and MCS (ES -0.04)[g]	Mean ± SD (post-intervention)
Exercise/ rehabilitation	Dalgas (2010)[32]	15 (I) 16 (C)	47.7 ± 10.4 49.1 ± 8.4	3.7 ± 0.9 3.9 ± 0.9	Progressive resistance training of lower extremities	2×/week, 12 weeks[e] Follow-up at 24 weeks	Waitlist	Fatigue (ES 0.84) Mood (ES 0.62) HRQL (0.61)	SF-36 PCS[a] and MCS (ES 0.61)[g]	Mean ± SD[b]
	Cakt (2010)[55]	15 (I1) 15 (I2) 15 (C)	36.4 ± 10.5 43.0 ± 10.2 35.5 ± 10.9	Not presented	Cycling resistance training[a] (I1) Lower-limb strengthening (I2)	2×/week, 2-months.	No intervention	Duration of exercise (ES 0.84)	SF-36 PCS[a] and MCS (ES 0.86)	PCS and MCS estimated using 8 subscale scores
	Dettmers (2009)[56]	15 (I) 15 (C)	45.8 ± 7.9 39.7 ± 9.1	2.6 ± 1.2 2.8 ± 0.7	Endurance exercises	45-min, 3×/ week, 3 weeks	Balance, stretching and coordination exercises	Walking distance (ES 0.83)	HAQUAMS (ES 0.45)	Data provided by study authors

(Continued)

Table 1. (Continued)

First author (year)	Sample size (I) Intervention group (C) Control group	Age Mean ± SD	EDSS Mean ± SD	Intervention	Duration and frequency	Control	Primary outcome of study (Effect size)	HRQL measure used in study (Effect size)	Data extracted
McCullagh (2008)[58]	17 (I) 13 (C)	40.5 ± 12.7 33.6 ± 6.1	Not presented	Walking/ running, cycling, arm-strengthening and stair-master	50-min, 2×/ week, 12 weeks[e] Follow-up at 6 months	Usual care	HRQL (ES 0.86) Fatigue (ES 1.06)	FAMS[a] MSIS-29 (total score) (ES 0.86)[g]	Median and IQR
Khan (2008)[62]	49 (I) 52 (C)	49.5 ± 8.6 51.1 ± 9.7	Range 0 to 6.5+	Multidisciplinary inpatient (IP) or outpatient (OP) rehabilitation	IP: 5 days/week, 3 h/day, 3–6 weeks OP: 2–3×/week, up to 6 weeks	Wait-list	Disability (ES 0.37)	MSIS-29 Physical[a] and Psychological (ES 0.08)[g]	Mean ± SD
Bjarnadottir (2007)[59]	6 (I) 10 (C)	38.7 36.1	2.1 1.8	Aerobic and strength training	60-min, 3×/ week, 5 weeks	Usual activity	Peak oxygen consumption (ES 1.52)	SF-36 PCS and MCS[a] (ES 0.65)	PCS and MCS estimated from 8 subscale scores
Storr (2006)[31]	38 (I) 52 (C)	53.0 ± 8.9 50.1 ± 9.9	6.5 (3.5–8.0)[d] 6.5 (1.5–8.0)[d]	Multi-disciplinary inpatient rehabilitation	3–5×/week, 3–5 weeks	Waitlist	HRQL (ES 0.28)	FAMS[a] LASQ (no data provided) (ES 0.28)	Mean ± SD
Romberg (2005)[57]	47 (I) 48 (C)	43.8 ± 6.3 43.9 ± 7.9	2.0 (1.5–3.5)[d] 2.5 (2.0–3.5)[d]	Resistance and aerobic training	4–5×/week, 26 weeks (weeks 1–3: inpatient rehabilitation, weeks 4–26: home program)	Waitlist	Functional impairment (ES 0.67)	MSQOL-54 Item on global QOL and PCS and MCS[a] (ES 0.10)[g]	Mean ± SD
Oken (2004)[60]	22 (I) 20 (C1) 15 (C2)	48.8 ± 10.4 49.8 ± 7.4 48.4 ± 9.8	2.9 ± 1.7 3.2 ± 1.7 3.1 ± 2.1	Yoga	90-min, 1×/ week, 6 months	Waitlist (C1)[a] Aerobic exercises (C2)	Cognitive function (ES 0.21)	SF-36 PCS and MCS[a] (ES 0.39)	Authors provided data
Pozzilli (2002)[28]	133 (I) 68 (C)	47.0 ± 10.3 46.7 ± 13.3	6.0 ± 2.0 5.8 ± 2.2	Home based multi-disciplinary care	12 months	Usual care (followed in MS clinic)	HRQL (ES 0.59) Cost (ES N/A)	SF-36 PCS[a] and MCS (ES 0.59)[g]	Mean, *p*-value
Patti (2002)[35]	58 (I) 53 (C)	45.2 ± 12.0 46.1 ± 6.0	6.2 ± 1.2 6.1 ± 1.2	Outpatient multi-disciplinary rehabilitation	60-min, 6×/ week, 6 weeks	Home exercise program	HRQL (ES 0.53)	SF-36 8 subscales (ES 0.53)	Data provided by study authors
Solari (1999)[61]	27 (I) 23 (C)	44.6 ± 10.2 44.9 ± 10.6	5.5 (3.0–6.5)[d] 5.5 (3.5–7.0)[d]	Inpatient rehabilitation	2×/day, each 45-min long, daily for 3 weeks[e] Followed up at weeks 9 and 15	Home exercise program	Disability (ES 0.90)	SF-36 PCS & MCS[a] (ES 0.60)[g]	Mean ± SD

Table 1. (Continued)

	First author (year)	Sample size (I) Intervention group (C) Control group	Age Mean ± SD	EDSS Mean ± SD	Intervention	Duration and frequency	Control	Primary outcome of study (Effect size)	HRQL measure used in study (Effect size)	Data extracted
	Petajan (1996)[54]	21 (I) 25 (C)	41.1 ± 2.0 39.0 ± 1.7	3.8 ± 0.3 2.9 ± 0.3	Aerobic training	30-min training, 3×/week, 15 weeks	No intervention	Exercise capacity (ES 3.83)	SIP Total, Physical[a] and Psychosocial (ES 0.61)[g]	Mean ± SE
Psychological interventions for mood	Cosio (2011)[36]	62 (I) 65 (C)	Not presented	Not presented	Telephone administered cognitive behavioral therapy	50-min, 1×/ week, 16 weeks	Telephone administered supportive emotion-focused therapy	HRQL (ES 0.51)	VAS for global QOL (ES 0.51)	Data provided by authors
	Forman (2010)[63]	18 (I) 18 (C)	47.3 ± 10.3 47.7 ± 9.8	18 (16–25)[c] 28 (19–31)[c]	Cognitive behavior therapy	Six sessions, 2-h/session, 3 months[e] Follow-up at 6 months	Usual care	Anxiety (ES 0.07) Depression (ES 0.94)	MSIS-29 Physical and Psychological dimension SF-36 PCS[a] and MCS (ES 0.62)[g]	Mean ± SD (post-intervention)
	Grossman (2010)[30]	76 (I) 74 (C)	45.9 ± 10.0 48.7 ± 10.6	3.03 ± 1.1 2.9 ± 0.8	Mindfulness training	Weekly 2.5-h classes, 2 months	Usual care	HRQL (ES 0.86)	PQOLC[a] HAQUAMS (ES 0.86)[g]	Mean ± SD

[a] Used in analysis.

[b] Obtained from primary author.

[c] Median (and inter-quartile range) measured using Guy's Neurological Disability Scale.

[d] Median (range).

[e] Time point used in analysis.

[f] Two groups combined to create a single pair-wise comparison with the intervention group.

[g] When more than one measure or domain was present, the one with the larger effect size was kept.

ES: Effect size; SD: standard deviation; IQR: inter-quartile range; SE: standard error; SEM: standard error of the mean; TENS: transcutaneous electrical nerve stimulation; EDSS: Expanded Disability Status Scale; QOL: quality of life; Leeds MSQOL: Leeds MS Quality of Life Scale; HAQUAMS: Hamburg Quality of Life Questionnaire in Multiple Sclerosis; SIP: Sickness Impact Profile; EQ-5D: EuroQol-5D; SF-12: Short Form-12; SF-36: Short Form-36; PCS: physical component summary; MCS: mental component summary; PQOLC: German-language Profile of Health-Related Quality of Life in Chronic Disorders; MSIS-29: Multiple Sclerosis Impact Scale-29; MSQOL-54: Multiple Sclerosis Quality of Life-54; SWLS: Satisfaction with Life Scale; FAMS: Functional Assessment of Multiple Sclerosis, LASQ: Life Appreciation and Satisfaction Questionnaire; VAS: visual analog scale; IP: inpatient; OP: outpatient

Table 2. List of HRQL measures used in the included studies.

Name	Abbreviation	Rescored (Yes/no)	Number of times used
Short Form-36	SF-36	No	18
Multiple Sclerosis Quality of Life-54	MSQOL-54	No	3
Multiple Sclerosis Impact Scale-29	MSIS-29	Yes	6
Single-item Quality of Life	Single-item QOL	No	4
Short Form-12	SF-12	No	2
Sickness Impact Profile	SIP	Yes	3
Functional Assessment of Multiple Sclerosis	FAMS	No	3
Hamburg Quality of Life Questionnaire in MS	HAQUAMS	Yes	3
Leeds MS Quality of Life Scale	LMSQOL	Yes	1
German-language Profile of Health-Related Quality of Life in Chronic Disorders	PQOLC	No	1
Satisfaction with Life Scale	SWLS	No	1
Life Appreciation and Satisfaction Questionnaire	LASQ	Yes	1
EuroQol-5D	EQ-5D	No	1
Incontinence Quality of Life Questionnaire	IQOL	No	1

mean estimate of effect of all studies combined was 0.35 (95% CI 0.02 to 0.68) with substantial heterogeneity among the studies (I^2=70%, p=0.007) (Figure 2(c)). There were no studies (0%) that were at high risk of bias for the domains evaluated. The Egger weighted regression statistic for publication bias was non-significant (p=0.31).

Effect of cognitive training on HRQL. There were three trials[51-53] that evaluated the impact of cognitive training on HRQL. The pooled ES of these studies was 0.38 (95% CI -0.26 to 1.02, p=0.24), with substantial heterogeneity among the studies (I^2=82.6%, p=0.003) (Figure 2(d)). Risk of bias from sequence generation and concealment of allocation was evaluated to be high for one study[52] (33%). Participants were not blinded to treatment arm in two (66%) of the trials.[51,52] There were no trials (0%) that were at high risk of bias for incomplete outcome data (Table 3). There were insufficient data points to assess publication bias.

Effect of exercise or rehabilitation on HRQL. There were 13 studies that evaluated the effects of exercise therapy or rehabilitation on HRQL. Interventions consisted of aerobic training,[54] resistance training,[32,55,56] aerobic combined with resistance training,[57-59] yoga,[60] physical therapy[61] and interdisciplinary rehabilitation.[28,31,35,62] The mean estimate of effect of all the studies combined was 0.43 (95% CI 0.29 to 0.57, p<0.001) (Figure 2(e)). The I^2 statistic for heterogeneity was null and non-significant (I^2=0%, p=0.53). Risk of bias from blinding of participants was assessed as high in all of the studies (100%), because given the nature of the intervention participants could not be blinded to treatment arms. Furthermore, blinding of outcome assessment was considered to be at high risk of bias in all studies (100%), as HRQL was measured via self-report. There were three studies (23%) that were at high risk of bias for incomplete outcome data (Table 3). The Egger weighted regression statistic for publication bias was non-significant (p=0.69).

Effect of psychological interventions targeting mood on HRQL. There were three trials that involved cognitive behavioral interventions to improve depression, anxiety and well-being. The pooled ES of the studies was 0.68 (95% CI 0.45 to 0.91) with no heterogeneity (I^2=0.9%, p=0.36) (Figure 2(f)). Risk of bias from sequence generation and incomplete outcome data was low or unclear for all three studies. As patients were not blinded to study hypothesis and HRQL was measured via self-report, all studies (100%) were at high risk of bias for blinding of participants and outcome assessment (Table 3). Publication bias could not be assessed due to the small number of studies.

Discussion

This study reported the results of a meta-analysis of an outcome rather than what is typically done, a meta-analysis of the effects of an intervention. In our study, the interventions varied but the outcome was constant. The magnitude of positive effect on HRQL varied between the different types of interventions. The smallest effect was observed for self-management and complementary and alternative medicine (ES=0.2), followed by medication (ES=0.3) and exercise and cognitive training (ES=0.4), followed by exercise, cognitive training and medication (ES=0.4), followed by psychological interventions to improve mood (ES=0.7).

Interventions regarding complementary and alternative medicine included reflexology, abdominal massage, transcutaneous electrical nerve stimulation (TENS), ginkgo extract and dietary interventions with essential fatty acids and vitamins/minerals. HRQL was measured as a primary outcome for three

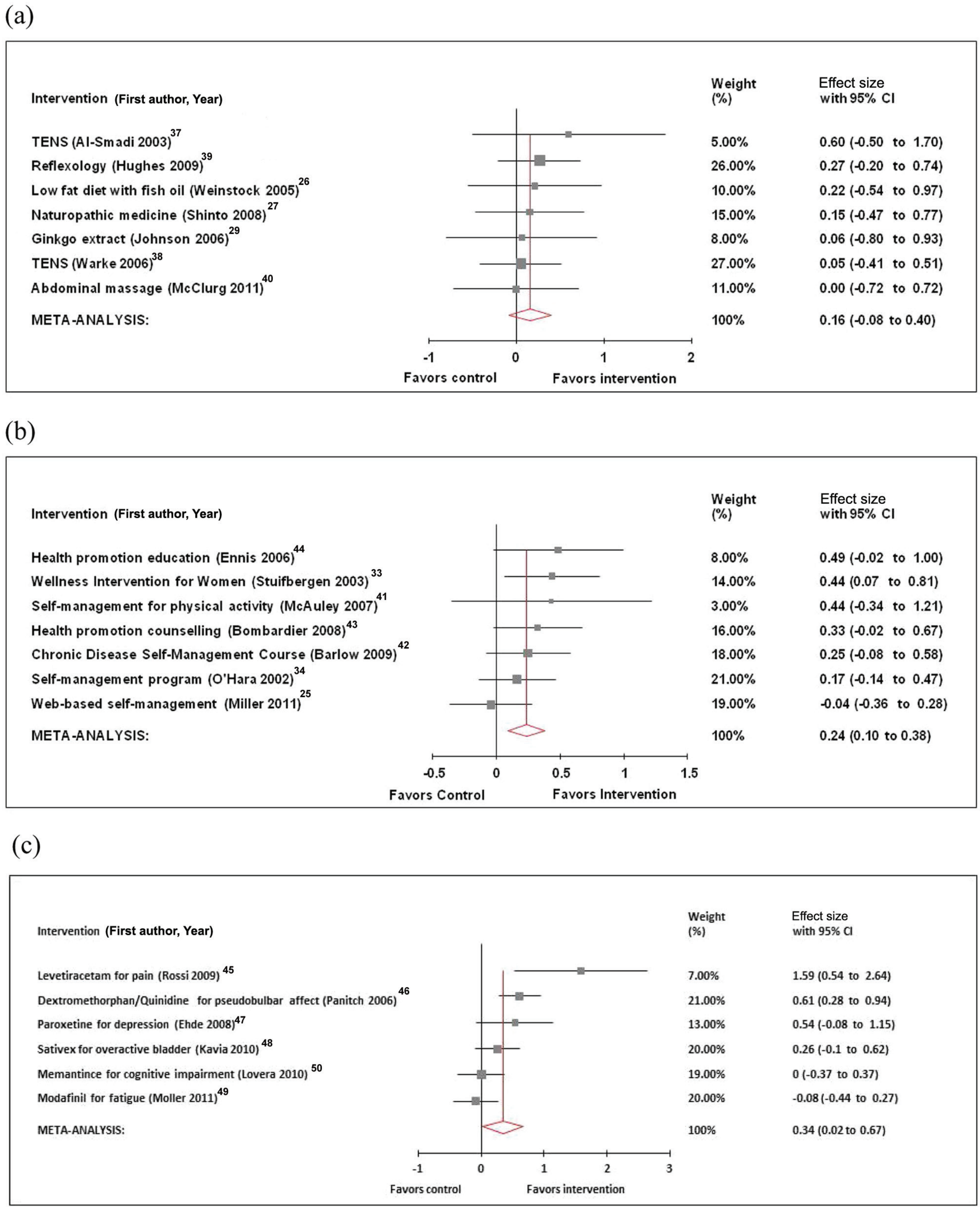


Figure 2. (Continued)

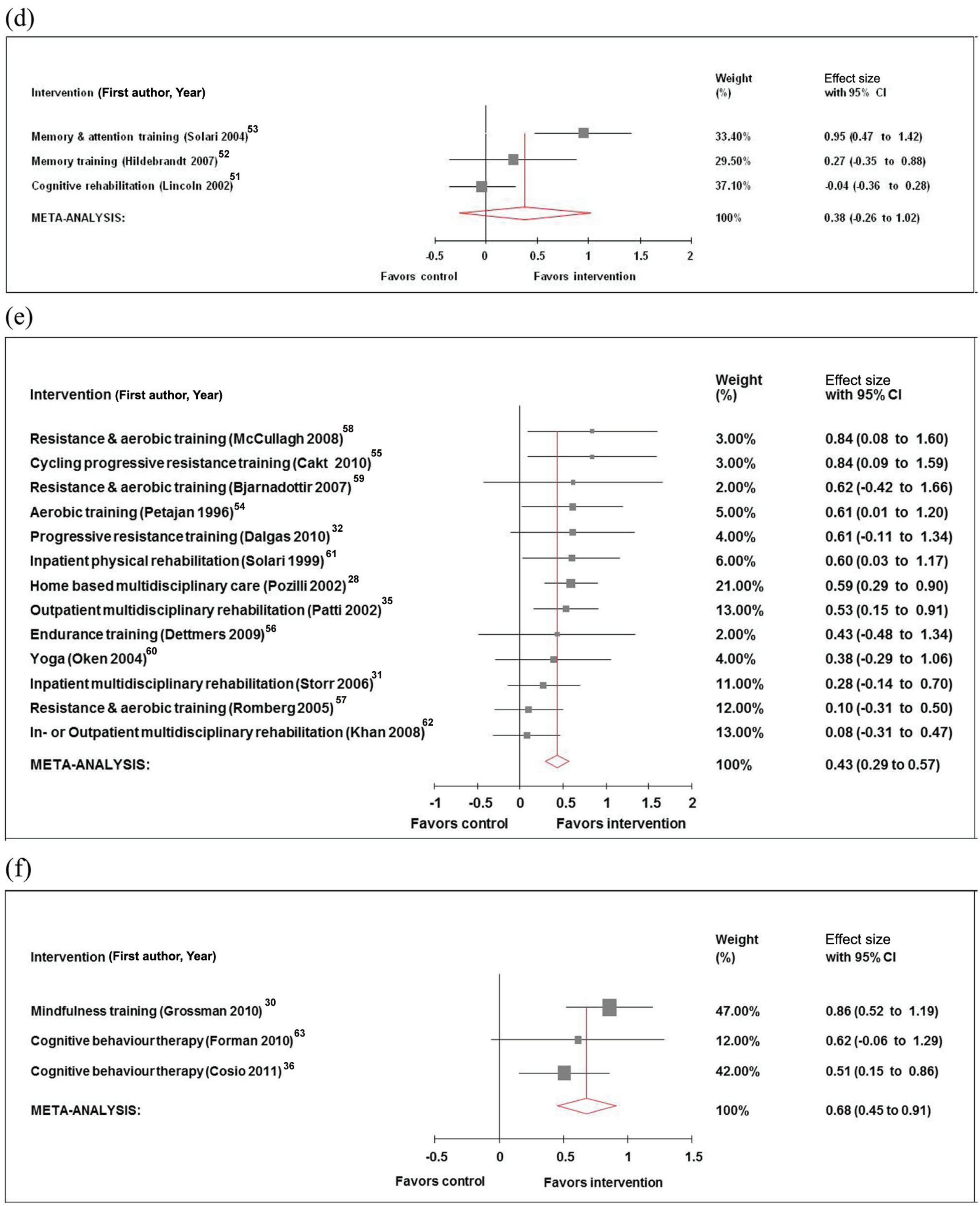


Figure 2. Random effects meta-analysis of (a) seven studies that examine the effects of complementary and alternative medicine on HRQL; (b) seven studies that examine the effects of self-management and self-efficacy on HRQL; (c) six studies that examine the effect of medications for symptom management on HRQL; (d) three studies that examine the effects of cognitive training on HRQL; (e) 13 studies that examine the effects of exercise training or rehabilitation on HRQL; (f) three studies that examine the effects of psychological interventions for mood on HRQL.

Table 3. Risk of bias in included studies.

	First author (year)	Domain				
		Adequate sequence generation	Adequate allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Incomplete outcome data addressed
Self-management	Miller (2011)[25]	Y	Y	N	N	N
	Barlow (2009)[42]	Y	Y	N	N	Y
	Bombardier (2008)[43]	Y	Y	N	N	Y
	McAuley (2007)[41]	U	U	N	N	Y
	Ennis (2006)[44]	Y	U	N	N	Y
	Stuifbergen (2003)[33]	Y	U	N	N	Y
	O'Hara (2002)[34]	Y	Y	N	N	N
Alternative medicine	McClurg (2011)[40]	Y	U	N	N	Y
	Hughes (2009)[39]	Y	U	Y	Y	Y
	Shinto (2008)[27]	U	U	N	N	Y
	Warke (2006)[38]	Y	Y	Y	Y	Y
	Johnson (2006)[29]	Y	Y	Y	Y	Y
	Weinstock (2005)[26]	U	U	Y	Y	Y
	Al-Smadi (2003)[37]	Y	Y	Y	Y	Y
Exercise/rehabilitation	Dalgas (2010)[32]	U	Y	N	N	Y
	Cakt (2010)[55]	Y	U	N	N	N
	Dettmers (2009)[56]	Y	U	N	N	N
	McCullagh (2008)[58]	Y	N	N	N	N
	Khan (2008)[62]	Y	U	N	N	Y
	Bjarnadottir (2007)[59]	Y	Y	N	N	N
	Storr (2006)[31]	N	N	N	N	Y
	Romberg (2005)[57]	Y	U	N	N	Y
	Oken (2004)[60]	Y	Y	N	N	Y
	Patti (2002)[35]	Y	Y	N	N	Y
	Pozilli (2002)[28]	Y	U	N	N	Y
	Solari (1999)[61]	Y	Y	N	N	Y
	Petajan (1996)[54]	U	U	N	N	Y
Cognitive training	Hildebrandt (2007)[52]	N	N	N	N	Y
	Solari (2004)[53]	Y	Y	Y	Y	Y
	Lincoln (2002)[51]	Y	Y	N	N	Y
Medication for symptoms	Moller (2011)[49]	U	U	Y	Y	Y
	Kavia (2010)[48]	Y	Y	Y	Y	Y
	Lovera (2010)[50]	Y	Y	Y	Y	Y
	Rossi (2008)[45]	Y	U	Y	Y	Y
	Ehde (2008)[47]	Y	Y	Y	Y	Y
	Panitch (2006)[46]	Y	Y	Y	Y	Y
Interventions for mood	Grossman (2010)[30]	Y	U	N	N	Y
	Forman (2010)[63]	Y	Y	N	N	Y
	Cosio (2010)[36]	Y	U	N	N	Y

Y: low risk of bias; N: high risk of bias; U: unclear risk of bias

of the trials,[26,27,29] and as a secondary outcome in four trials.[37-40] The pooled estimate of effect for the seven studies was small and did not reach statistical significance. Only one intervention[37] (which was a pilot study), which involved the use of TENS for low back pain, demonstrated a clinically significant effect on HRQL; however, this effect did not reach statistical significance. Due to these encouraging results Warke et al.[38] conducted a larger RCT a few years later, but did not observe a significant treatment effect on HRQL.

The effect of a self-management program on HRQL was assessed in seven studies, and their combined effect was, by Cohen's criteria, small. Two studies involved health promotion counseling,[43,44] one study involved enhancing self-efficacy related to physical activity,[41] one was an internet-based self-management system[25] and the remaining were self-management programs in the community.[33,34,42] The pooled estimate of effect was statistically significant as the 95% CI excluded the null value. Three of the studies[33,41,44] had a moderate effect on HRQL; however, only one[33] reached statistical significance. The quality of the trials was moderate; suggesting that more research regarding self-management in MS is required.

Among the different types of medication targeting symptom management, levetiracetam for central neuropathic pain,[45] dextromethorphan/quinidine for pesudobulbar affect[46] and paroxetine for major depressive disorder[47] had a clinically significant effect on HRQL. The former two medications also demonstrated a statistically significant effect on HRQL, whereas the latter did not. The pooled ES of all of the interventions combined was of moderate magnitude and statistically significant. However, there was considerable heterogeneity among the studies so the combined estimate of effect must be interpreted with caution.

The effect of cognitive training on HRQL was evaluated in three trials.[51-53] Two trials involved using a computer-aided program targeting memory[52] or memory and attention,[53] while the third trial[51] involved cognitive rehabilitation of any deficits identified during initial evaluation. The primary endpoint for all of the studies was cognitive performance. HRQL was measured as a secondary endpoint. Out of the three studies, the one by Solari et al.[53] had, by Cohen's criteria, a large effect on HRQL. This effect was both clinically and statistically significant. However, the other two included studies did not observe a clinically or statistically significant effect on HRQL. Solari et al.[53] only included people with cognitive impairments, whereas the other two studies included people with and without cognitive impairments. This important difference in sampling strategy may explain why the former had a large effect while the latter had a small or no effect on HRQL. The mean estimate of effect of the studies combined was of moderate magnitude but did not reach statistical significance. However, the pooled estimate of effect must be inferred with caution because of the large degree of clinical heterogeneity among the studies.

For interventions involving exercise or rehabilitation, aerobic training,[54] progressive resistance training,[32,55] aerobic combined with resistance training,[53,59] inpatient physical rehabilitation[61] and interdisciplinary rehabilitation[28,35] had clinically significant effects on HRQL. Out of these eight interventions, six of them[28,35,53-55,61] reached statistical significance. There was a low level of heterogeneity among the studies, so there were probably no major clinical (e.g. participants and outcomes) or methodological (e.g. study design or risk of bias) differences between them.[13] The combined estimate of effect was of moderate magnitude and statistically significant, as the CI excluded the null value.

The largest pooled effect was observed for psychological interventions targeting emotional well-being (i.e. mood). There were three studies in this area: one was mindfulness training[30] and the other two were cognitive behavioral therapy.[37,63] All interventions had a clinically significant effect on HRQL; however, only two[30,37] reached statistical significance. HRQL was a primary endpoint for the latter two studies; hence they were probably powered to detect an effect of that magnitude. The combined effect of the three interventions was, by Cohen's criteria, large and statistically significant.

Methodological quality of the included studies

Most of the included studies were of moderate or high quality, with low risk of bias. Behavioral interventions such as self-management, exercise and psychological interventions for mood were at high risk of bias for blinding of patients and outcome assessment. However, the feasibility of blinding patients in such studies is often very difficult or impossible.[40] Incomplete outcome data was adequately addressed in many of the trials by using intention to treat analysis. For the studies that used per protocol analysis, reasons for missing data were explained and follow-up response rates were greater than 80%.

Consistency between primary outcome findings and HRQL

There were 26 trials where HRQL was not a primary endpoint (was measured as a secondary outcome). There was 73% (*n*=19) agreement between the primary and the secondary outcome measures (HRQL). In six out of 26 studies (23%)[38,40,43,49,52,57] the intervention had a clinically significant effect on the primary outcome measure, but not on HRQL. The primary outcome measures in these studies were symptoms (e.g. pain, fatigue, memory, etc.) or functional status (e.g. walking ability). These findings suggest that, although an intervention may have an effect on symptoms or function, this effect does not always carry over to HRQL.

Effect size and sample size

Out of the 39 included trials, 18 found an ES of 0.5 or greater, but only 11 (61%) were powered to detect this ES. The remaining 21 studies found ESs smaller than 0.5 and, with the exception of one,[33] none were powered for these ESs. Whether a study is able to statistically detect a difference depends on the magnitude of the effect to detect and the sample size.[64,65] The effect of sample size on significance can best be visualized using the 95% CI, as wide intervals arise from small studies and the effect does not reach statistical significance when the interval includes the null value.

The trials included in this systematic review had sample sizes ranging from five to 133 per group. In fact, for a trial with two independent samples, with alpha set to 0.05 and 80% power, the sample size required to detect a large ES

(0.8) is 26 persons per group, a moderate ES (0.5) is 64 per group and a small ES (0.2) is 394 per group. Most of the included studies may not have had sufficient sample size because HRQL was a secondary endpoint for them. Hence, sample size calculations were not based on the HRQL measure administered, but rather on the study's primary outcome measure. However, in order for us to accurately assess the effects of existing health care interventions on HRQL, we need more studies that are targeting HRQL as a primary endpoint. Or if HRQL is assessed as a secondary endpoint, we need studies that are adequately powered to detect a significant effect. This meta-analysis provides estimates of ESs for sample size considerations in future trials. Studies that involve psychological interventions for mood (ES=0.69) require 35 persons per group. Those that involve exercise (ES=0.43), cognitive training (ES=0.38) and medication (ES=0.35) require 86, 110 and 130 persons per group, respectively. Furthermore, trials that are concerned with self-management (ES=0.24) and complementary and alternative medicine (ES=0.16) need 274 and 615 people per group, respectively.

Limitations

There were several limitations that need to be addressed. First, we included only studies that were published in peer reviewed journals, and excluded unpublished or grey literature. However, the funnel plots indicated that the exclusion of unpublished data did not have an effect on publication bias. Second, we included all types of control groups (active and inactive) in the review. An intervention that is compared with an inactive control group may demonstrate a larger effect than one compared with an active control group. However, we had only six trials where the control group was given an active intervention.[25,27,35,36,41,61] Third, the magnitude of ES observed for an intervention may depend on whether the HRQL outcome was disease specific or generic. This is because disease-specific measures may be more responsive to change and thus yield larger ESs than generic measures.[13] Last, we cannot rule out the presence of response shift in a trial. When individuals experience a change in their health state, they may alter their internal standards, values or conceptualization of HRQL.[66,67] At randomization of a clinical trial, both groups will likely start with the same conceptualization of the outcome (HRQL) and internal standard of measurement. However, through the intervention, the treatment group may obtain new information and knowledge about MS and ways of coping with the disease, therefore altering the evaluation of their HRQL.[68] Furthermore, we cannot conclude that change was solely due to an intervention effect and that it was not affected by response shift, unless response shift is evaluated and ruled out using design and statistical approaches. As we were not able to measure it in the context of this review, we cannot rule out the presence of response shift.[68]

Conclusion

The extent to which interventions are able to improve outcomes depends on delivering a potent intervention to those persons who have the potential to benefit. Therefore, interventions targeting specific outcomes will be more effective for those people with the targeted problem (e.g. pain, spasticity, incontinence, or memory and attention deficits). These targeted interventions are often included in good clinical care and are relatively easy to implement. However, interventions such as exercise or self-management, which are likely to potentially benefit all, are in contrast difficult to implement, particularly in a highly medicalized clinical environment. It is also important that interventions be designed optimally using theory and/or evidence to guide their components. While exercise interventions have a strong empirical base, there is now a strong theoretical basis for components of self-management interventions,[39] but it is not clear how many of these elements of effective self-management were incorporated into the included studies. Therefore, future areas of research should include not only knowledge generation to develop and target needed interventions, but also research in knowledge translation. A common challenge in studies of knowledge generation and knowledge translation is designing adequately powered studies of potent and meaningful interventions.

Acknowledgements

This work was supported in part by a scholarship from Fonds de la Recherche en Santé du Québec, Multiple Sclerosis Society of Canada, and the Physiotherapy Foundation of Canada.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

References

1. Noseworthy JH, Lucchinetti C, Rodriguez M, et al. Multiple sclerosis. *N Engl J Med* 2000; 343: 938–952.
2. Freeman JA, Thompson AJ, Fitzpatrick R, et al. Interferon-beta1b in the treatment of secondary progressive MS: Impact on quality of life. *Neurology* 2001; 57: 1870–1875.
3. Orton SM, Herrera BM, Yee IM, et al. Sex ratio of multiple sclerosis in Canada: A longitudinal study. *Lancet Neurol* 2006; 5: 932–936.
4. Wingerchuk DM, Noseworthy JH and Lucchinetti CF. Multiple sclerosis: Current pathophysiological concepts. *Lab Invest* 2001; 81: 263–281.
5. Nortvedt MW, Riise T, Myhr KM, et al. Quality of life in multiple sclerosis: Measuring the disease effects more broadly. *Neurology* 1999; 53: 1098–1103.
6. Pugliatti M, Riise T, Nortvedt MW, et al. Self-perceived physical functioning and health status among fully ambulatory multiple sclerosis patients. *J Neurol* 2008; 255: 157–162.
7. Rudick RA, Miller D, Clough JD, et al. Quality of life in multiple sclerosis. Comparison with inflammatory bowel disease and rheumatoid arthritis. *Arch Neurol* 1992; 49: 1237–1242.

8. Riazi A, Hobart JC, Lamping DL, et al. Using the SF-36 measure to compare the health impact of multiple sclerosis and Parkinson's disease with normal population health profiles. *J Neurol Neurosurg Psychiatry* 2003; 74: 710–714.
9. Breslow L. A quantitative approach to the World Health Organization definition of health: Physical, mental and social well-being. *Int J Epidemiol* 1972; 1: 347–355.
10. Guyatt GH, Feeny DH and Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993; 118: 622–629.
11. Hays RD. Generic versus disease-targeted instruments. In: Fayers P and Hays D (eds) *Assessing quality of life in clinical trials*. 2nd ed. New York: Oxford University Press Inc., 2005, pp.3–8.
12. Feeny D. Preference-based measures: utility and quality-adjusted life years. In: Fayers P and Hays D (eds) *Assessing quality of life in clinical trials*. 2nd ed. New York: Oxford University Press Inc., 2005, pp.405–429.
13. Higgins JP and Green S. *Cochrane handbook for systematic reviews of interventions*. Chichester: John Wiley & Sons Ltd, 2008.
14. O'Connor D, Green S and Higgins JP. Defining the review question and developing criteria for including studies. In: Higgins JP and Green S (eds) *Cochrane handbook for systematic reviews of interventions*. Chichester: John Wiley & Sons Ltd, 2008, pp.83–94.
15. Higgins JP, Altman DG and Sterne JAC. Assessing risk of bias in included studies. In: Higgins JP and Green S (eds) *Cochrane handbook for systematic reviews of interventions*. Chichester: John Wiley & Sons Ltd, 2008, pp.187–241.
16. Egger M, Smith GD, and Altman DG. *Systematic reviews in health care: Meta-analysis in context*. 2nd ed. London: BMJ Publishing Group, 2007.
17. Kazis LE, Anderson JJ and Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989; 27: S178–S189.
18. Asano M, Dawes DJ, Arafah A, et al. What does a structured review of the effectiveness of exercise interventions for persons with multiple sclerosis tell us about the challenges of designing trials? *Mult Scler* 2009; 15: 412–421.
19. Higgins JP and Deeks JJ. Selecting studies and collecting data. In: Higgins JP and Green S (eds) *Cochrane handbook for systematic reviews of interventions*. Chichester: John Wiley & Sons Ltd, 2008, pp.151–185.
20. Ware JE Jr, Kosinski M and Keller SD. *SF-36 physical and mental health summary scales: A user's manual*. Boston: New England Medical Center, 1994.
21. Deeks JJ, Higgins JP and Altman DG. Analyzing data and undertaking meta-analysis. In: Higgins JP and Green S (eds) *Cochrane handbook for systematic reviews of interventions*. Chichester: John Wiley & Sons Ltd, 2008, pp.243–293.
22. Cohen J. A power primer. *Psychol Bull* 1992; 112: 155-159.
23. Sloan JA, Cella D and Hays RD. Clinical significance of patient-reported questionnaire data: Another step toward consensus. *J Clin Epidemiol* 2005; 58: 1217–1219.
24. Sterne JAC, Egger M and Moher D. Addressing reporting biases. In: Higgins JP and Green S (eds) *Cochrane handbook for systematic reviews of interventions*. Chichester: John Wiley & Sons Ltd, 2008, pp.297–334.
25. Miller DM, Moore SM, Fox RJ, et al. Web-based self-management for patients with multiple sclerosis: A practical, randomized trial. *Telemed J E-Health* 2011; 17: 5–13.
26. Weinstock-Guttman B, Baier M, Park Y, et al. Low fat dietary intervention with omega-3 fatty acid supplementation in multiple sclerosis patients. *Prostaglandins Leukot Essent Fatty Acids* 2005; 73: 397–404.
27. Shinto L, Calabrese C, Morris C, et al. A randomized pilot study of naturopathic medicine in multiple sclerosis. *J Altern Complement Med* 2008; 14: 489-496. (Erratum appears in *J Altern Complement Med* 2008; 14: 793).
28. Pozzilli C, Brunetti M, Amicosante AM, et al. Home based management in multiple sclerosis: Results of a randomised controlled trial. *J Neurol Neurosurg Psychiatry* 2002; 73: 250–255.
29. Johnson SK, Diamond BJ, Rausch S, et al. The effect of Ginkgo biloba on functional measures in multiple sclerosis: A pilot randomized controlled trial. *Explore (NY)* 2006; 2: 19–24.
30. Grossman P, Kappos L, Gensicke H, et al. MS quality of life, depression, and fatigue improve after mindfulness training: A randomized trial. *Neurology* 2010; 75: 1141–1149.
31. Storr LK, Sorensen PS and Ravnborg M. The efficacy of multidisciplinary rehabilitation in stable multiple sclerosis patients. *Mult Scler* 2006; 12: 235–242.
32. Dalgas U, Stenager E, Jakobsen J, et al. Fatigue, mood and quality of life improve in MS patients after progressive resistance training. *Mult Scler* 2010; 16: 480–490.
33. Stuifbergen AK, Becker H, Blozis S, Timmerman G, Kullberg V. A randomized clinical trial of a wellness intervention for women with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation 84 (4) (pp 467-476), 2003 Date of Publication: 01 Apr 2003* 2003; 01.
34. O'Hara L, Cadbury H, De SL, et al. Evaluation of the effectiveness of professionally guided self-care for people with multiple sclerosis living in the community: A randomized controlled trial. *Clin Rehabil* 2002; 16: 119–128.
35. Patti F, Ciancio MR, Reggio E, et al. The impact of outpatient rehabilitation on quality of life in multiple sclerosis. *J Neurol* 2002; 249: 1027–1033.
36. Cosio D, Jin L, Siddique J, et al. The effect of telephone-administered cognitive-behavioral therapy on quality of life among patients with multiple sclerosis. *Ann Behav Med* 2011; 41: 227–234.
37. Al-Smadi J, Warke K, Wilson I, et al. A pilot investigation of the hypoalgesic effects of transcutaneous electrical nerve stimulation upon low back pain in people with multiple sclerosis. *Clin Rehabil* 2003; 17: 742–749.
38. Warke K, Al-Smadi J, Baxter D, et al. Efficacy of transcutaneous electrical nerve stimulation (tens) for chronic low-back pain in a multiple sclerosis population: A randomized, placebo-controlled clinical trial. *Clin J Pain* 2006; 22: 812–819.
39. Hughes CM, Smyth S and Lowe-Strong AS. Reflexology for the treatment of pain in people with multiple sclerosis: A double-blind randomised sham-controlled clinical trial. *Mult Scler* 2009; 15: 1329–1338.
40. McClurg D, Hagen S, Hawkins S, et al. Abdominal massage for the alleviation of constipation symptoms in people with multiple sclerosis: A randomized controlled feasibility study. *Mult Scler* 2011; 17: 223–233.
41. McAuley E, Motl RW, Morris KS, et al. Enhancing physical activity adherence and well-being in multiple sclerosis: A randomised controlled trial. *Mult Scler* 2007; 13: 652–659.

42. Barlow J, Turner A, Edwards R, et al. A randomised controlled trial of lay-led self-management for people with multiple sclerosis. *Patient Educ Couns* 2009; 77: 81–89.
43. Bombardier CH, Cunniffe M, Wadhwani R, et al. The efficacy of telephone counseling for health promotion in people with multiple sclerosis: A randomized controlled trial. *Arch Phys Med Rehabil* 2008; 89: 1849–1856.
44. Ennis M, Thain J, Boggild M, et al. A randomized controlled trial of a health promotion education programme for people with multiple sclerosis. *Clin Rehabil* 2006; 20: 783–792.
45. Rossi S, Mataluni G, Codeca C, et al. Effects of levetiracetam on chronic pain in multiple sclerosis: Results of a pilot, randomized, placebo-controlled study. *Eur J Neurol* 2009; 16: 360–366.
46. Panitch HS, Thisted RA, Smith RA, et al. Randomized, controlled trial of dextromethorphan/quinidine for pseudobulbar affect in multiple sclerosis. *Ann Neurol* 2006; 59: 780–787.
47. Ehde DM, Kraft GH, Chwastiak L, et al. Efficacy of paroxetine in treating major depressive disorder in persons with multiple sclerosis. *Gen Hosp Psychiatry* 2008; 30: 40–48.
48. Kavia RB, De RD, Constantinescu CS, et al. Randomized controlled trial of Sativex to treat detrusor overactivity in multiple sclerosis. *Mult Scler* 2010; 16: 1349–1359.
49. Moller F, Poettgen J, Broemel F, et al. HAGIL (Hamburg Vigil Study): A randomized placebo-controlled double-blind study with modafinil for treatment of fatigue in patients with multiple sclerosis. *Mult Scler* 2011; 17: 1002–1009.
50. Lovera JF, Frohman E, Brown TR, et al. Memantine for cognitive impairment in multiple sclerosis: A randomized placebo-controlled trial. *Mult Scler* 2010; 16: 715–723.
51. Lincoln NB, Dent A, Harding J, et al. Evaluation of cognitive assessment and cognitive intervention for people with multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2002; 72: 93–98.
52. Hildebrandt H, Lanz M, Hahn HK, et al. Cognitive training in MS: Effects and relation to brain atrophy. *Restor Neurol Neurosci* 2007; 25: 33–43.
53. Solari A, Motta A, Mendozzi L, et al. Computer-aided retraining of memory and attention in people with multiple sclerosis: A randomized, double-blind controlled trial. *J Neurol Sci* 2004; 222: 99-104. (Erratum appears in *J Neurol Sci* 2004; 224: 113.)
54. Petajan JH, Gappmaier E, White AT, et al. Impact of aerobic training on fitness and quality of life in multiple sclerosis. *Ann Neurol* 1996; 39: 432–441.
55. Cakt BD, Nacir B, Genc H, et al. Cycling progressive resistance training for people with multiple sclerosis: A randomized controlled study. *Am J Phys Med Rehabil* 2010; 89: 446–457.
56. Dettmers C, Sulzmann M, Ruchay-Plossl A, et al. Endurance exercise improves walking distance in MS patients with fatigue. *Acta Neurol Scand*2009; 120: 251–257.
57. Romberg A, Virtanen A and Ruutiainen J. Long-term exercise improves functional impairment but not quality of life in multiple sclerosis. *J Neurol* 2005; 252: 839–845.
58. McCullagh R, Fitzgerald AP, Murphy RP, et al. Long-term benefits of exercising on quality of life and fatigue in multiple sclerosis patients with mild disability: A pilot study. *Clin Rehabil* 2008; 22: 206–214.
59. Bjarnadottir OH, Konradsdottir AD, Reynisdottir K, et al. Multiple sclerosis and brief moderate exercise. A randomised study. *Mult Scler* 2007; 13: 776–782.
60. Oken BS, Kishiyama S, Zajdel D, et al. Randomized controlled trial of yoga and exercise in multiple sclerosis. *Neurology* 2004; 62: 2058–2064.
61. Solari A, Filippini G, Gasco P, et al. Physical rehabilitation has a positive effect on disability in multiple sclerosis patients. *Neurology* 1999; 52: 57–62.
62. Khan F, Pallant JF, Brand C, et al. Effectiveness of rehabilitation intervention in persons with multiple sclerosis: A randomised controlled trial. *J Neurol Neurosurg Psychiatry* 2008; 79: 1230–1235.
63. Forman AC and Lincoln NB. Evaluation of an adjustment group for people with multiple sclerosis: A pilot randomized controlled trial. *Clin Rehabil* 2010; 24: 211–221.
64. Noordzij M, Tripepi G, Dekker FW, et al. Sample size calculations: Basic principles and common pitfalls. *Nephrol Dial Transplant* 2010; 25: 1388–1393.
65. Moher D, Dulberg CS and Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994; 272: 122–124.
66. Schwartz CE, Bode R, Repucci N, et al. The clinical significance of adaptation to changing health: A meta-analysis of response shift. *Qual Life Res* 2006; 15: 1533–1550.
67. Schwartz CE and Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med* 1999; 48: 1531–1548.
68. Mayo NE, Scott SC and Ahmed S. Case management post-stroke did not induce response shift: The value of residuals. *J Clin Epidemiol* 2009; 62: 1148–1156.

Supplementary Table 1 Descriptive information for excluded studies (in chronological order)

First Author (Year)	Sample size (I) Intervention group (C) control group	Age Mean ± SD	EDSS Mean ± SD	Intervention	Duration and frequency	Control	Primary outcome of study	HRQL measure used in study	Reasons for exclusion
Schwartz (1999)	64 (I) 68 (C)	43.0 ± 9.0 (I+C)	4.6 ± 1.7 4.7 ± 1.8	Coping skills intervention	2h, 1x/week, 8 weeks	Peer telephone support	HRQL	SIP Satisfaction subscale of the QOL index	Mean & SD not presented
Mostert (2002)	13 (I) 13 (C)	45.2 ± 8.7 43.9 ± 13.9	4.6 ± 1.2 4.5 ± 1.9	Aerobic training	30-min, 5x/week, 3-4 weeks	No intervention	Maximal aerobic capacity	SF-36 social function and vitality subscale	Data presented for only 2/8 subscales.
Fowler (2005)	104 (I) 113 (C)	45.0 (26.0-73.0) 47.0 (23.0-65.0)	4.1 (1.5-6.0) 3.9 (1.0-6.0)	Sildenafil citrate	25-100mg, 12 weeks	Placebo	Erectile function index	Life Satisfaction Checklist	Checklist, scored by item, no total score, mean & SD not presented
Sutherland (2005)	11 (I) 11 (C)	43.6 ± 9.5 40.8 ± 6.1	Not presented	Autogenic training	1x/week, 10 weeks	No intervention	HRQL	MSQOL-54	Transformed subscale scores from 0 to 100, or composite scores not presented.

First Author (Year)	Sample size (I) Intervention group (C) control group	Age Mean ± SD	EDSS Mean ± SD	Intervention	Duration and frequency	Control	Primary outcome of study	HRQL measure used in study	Reasons for exclusion
McClurg (2006)	10 (I1) 10 (I2) 10 (C)	52.1 ± 11.5 49.9 ± 11.6 49.5 ± 8.7	5.9 ± 1.3 5.7 ± 1.0 5.4 ± 1.3	PFTA with EMG (I1) PFTA, EMG, and NMES (I2)	1x/week, 9 weeks at the clinic with home exercises Follow-up at 16 and 24 weeks.	PFTA only	Leakage episodes per 24h	MSQOL-54	Mean & SD not presented

SD: Standard Deviation; PFTA: Pelvic floor training and advice; EMG: electromyography; NMES: neuromuscular electrical stimulation.

Supplementary Table 2 Risk of bias in excluded studies

	Domain				
Author (Year)	**Adequate sequence generation**	**Adequate allocation concealment**	**Blinding of participants & personnel**	**Blinding of outcome assessment**	**Incomplete outcome data addressed**
Schwartz (1999)	Y	U	N	N	Y
Mostert (2002)	U	U	N	N	N
Fowler (2005)	Y	Y	Y	Y	Y
Sutherland (2005)	U	U	N	N	Y
McClurg (2006)	Y	U	Y	Y	Y

Schwartz CE. Teaching coping skills enhances quality of life more than peer support: Results of a randomized trial with multiple sclerosis patients. *Health Psychol* 1999; 18: 211–220.

Mostert S and Kesselring J. Effects of a short-term exercise training program on aerobic fitness, fatigue, health perception and activity level of subjects with multiple sclerosis. *Mult Scler* 2002; 8: 161–168.

Fowler CJ, Miller JR, Sharief MK, et al. A double blind, randomized study of sildenafil citrate for erectile dysfunction in men with multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2005; 76: 700–705.

Sutherland G, Andersen MB and Morris T. Relaxation and health-related quality of life in multiple sclerosis: The example of autogenic training. *J Behav Med* 2005; 28: 249–256.

McClurg D, Ashe RG, Marshall K, et al. Comparison of pelvic floor muscle training, electromyography biofeedback, and neuromuscular electrical stimulation for bladder dysfunction in people with multiple sclerosis: A randomized pilot study. *Neurourol Urodyn* 2006; 25: 337–348.

CHAPTER 4: Integration of manuscripts 1 and 2

Research questions of manuscript 1 and 2

Manuscript 1:

The effects of clinical interventions on health-related quality of life in multiple sclerosis: a meta-analysis

Manuscript 2:

Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis?

Integration of manuscript 1 and 2

The first manuscript was a systematic review and meta-analysis on the effects of clinical interventions on health-related quality of life (HRQL) in people with multiple sclerosis (MS). Studies that measured HRQL as an outcome, ideally using an accepted and validated instrument, were reviewed. Among the 39 randomized clinical trials that were included, health profiles were the most commonly used outcome measures in MS. However, the challenge with using health profiles in clinical research is that they do not provide a single value on the net effect of an intervention on patients' HRQL. At the end of this review, we identified the need for a more harmonized approach to the measurement of HRQL, particularly if we wanted to compare across interventions.

The overall objective of this thesis is to take important steps towards developing a preference-based measure for MS. Therefore, in the next manuscript we identified the domains that were most important to the quality of life of people with MS and mapped these domains onto generic preference-based measures. By doing so, we were able to recognize the domains that were missing in these generic measures and that should be included in a MS specific preference-based measure.

CHAPTER 5 (MANUSCRIPT 2)

Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis?

Ayse Kuspinar[1] and Nancy E. Mayo[1,2]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada

[2]Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

Published in *Health and Quality of Life Outcomes* (2013) 11:71.

Communication addressed to:

Ayse Kuspinar, M.Sc., Ph.D. Candidate
School of Physical & Occupational Therapy
Faculty of Medicine, McGill University
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
Tel: 514-934-1934 ext 31564
E-mail: ayse.kuspinar@mail.mcgill.ca

HEALTH AND QUALITY OF LIFE OUTCOMES

RESEARCH Open Access

Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis?

Ayse Kuspinar[1*] and Nancy E Mayo[1,2]

Abstract

Purpose: The three most widely used utility measures are the Health Utilities Index Mark 2 and 3 (HUI2 and HUI3), the EuroQol-5D (EQ-5D) and the Short-Form-6D (SF-6D). In line with guidelines for economic evaluation from agencies such as the National Institute for Health and Clinical Excellence (NICE) and the Canadian Agency for Drugs and Technologies in Health (CADTH), these measures are currently being used to evaluate the cost-effectiveness of different interventions in MS. However, the challenge of using such measures in people with a specific health condition, such as MS, is that they may not capture all of the domains that are impacted upon by the condition. If important domains are missing from the generic measures, the value derived will be higher than the real impact creating invalid comparisons across interventions and populations. Therefore, the objective of this study is to estimate the extent to which generic utility measures capture important domains that are affected by MS.

Methods: The available study population consisted of men and women who had been registered after 1994 in three participating MS clinics in Greater Montreal, Quebec, Canada. Subjects were first interviewed on an individualized measure of quality of life (QOL) called the Patient Generated Index (PGI). The domains identified with the PGI were then classified and grouped together using the World Health Organization's International Classification of Functioning, Disability and Health (ICF), and mapped onto the HUI2, HUI3, EQ-5D and SF-6D.

Results: A total of 185 persons with MS were interviewed on the PGI. The sample was relatively young (mean age 43) and predominantly female. Both men and women had mild disability with a median Expanded Disability Status Scale (EDSS) score of 2. The top 10 domains that patients identified to be the most affected by their MS were, work (62%), fatigue (48%), sports (39%), social life (28%), relationships (23%), walking/mobility (22%), cognition (21%), balance (14%), housework (12%) and mood (11%). The SF-6D included the most number of domains (6 domains) important to people with MS, followed by the EQ-5D (4 domains) and the HUI2 (4 domains) and then the HUI3 (3 domains). The mean and standard deviation (SD) for the PGI, EQ-5D and the SF-6D were 0.50 (SD 0.25), 0.69 (0.18) and 0.69 (0.13), respectively. The magnitude of difference between the PGI and the generic utility measures was large and statistically significant.

Conclusion: Although the generic utility measures included certain items that were important to people with MS, there were several that were missing. An important consequence of this mismatch was that values of QOL derived from the PGI were importantly and significantly lower than those estimated using any of the generic utility measures. This could have a substantial impact in evaluating the effect of interventions for people with MS.

Keywords: Multiple sclerosis, Quality of life, Health-related quality of life, Measurement, Utilities

* Correspondence: ayse.kuspinar@mail.mcgill.ca
[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, 3654 Promenade Sir-William-Osler, Montreal 3G 1Y5, QC, Canada
Full list of author information is available at the end of the article

Introduction

Multiple sclerosis (MS) is a chronic disease resulting from inflammation and demyelination in the central nervous system (CNS) [1] that is associated with a variety of symptoms, such as fatigue, impaired mobility and cognitive decline [2]. Several new therapies, behavioural [3-9], medical [10-14], and surgical [15-19], have been developed in the field of MS. As there are both benefits and harms from interventions, the importance of considering the patient's perspective in the evaluation of these new therapies is increasingly being emphasized. Patient-reported outcomes are used to evaluate the patient's perspective on the impact of the disease and its treatment on symptoms, function, and other aspects of quality of life (QOL). QOL is defined as an "*individuals' perception of their position in life in the context of the culture in which they live and in relation to their goals, expectations, standards and concerns* [20]." QOL is a global construct that includes domains other than health such as job satisfaction, quality of housing, and the neighborhood in which one lives [21]. Health-related quality of life (HRQL), on the other hand, is a construct that is narrower and focuses on domains within the purview of the health care system, such as normal ranges for physiological variables, physical, mental and social well-being [22,23]. Health status, a term often confused with HRQL, is a description and/or measurement of the health of an individual or population at a particular point in time against identifiable standards [24].

While there are a common set of domains that are relevant across a wide variety of health conditions, including none, these domains may be affected differentially because of the positive and negative effects of interventions. For example, a treatment may have a positive effect on one domain (e.g. mental health) but a negative one on another (e.g. physical health) and this would be condition and intervention specific.

The most widely used methodology to create an index that weighs gains in one domain against losses in another is based on utility theory. Utility measures (or preference-based measures) provide a single value for the construct (health status, HRQL, or QOL) ranging from 0 (for death or worst possible health state) to 1 (for perfect health or best possible health state) [25-29]. This value is used to calculate what is termed a "Quality-Adjusted Life Year" (QALY) which captures the effect of an intervention on quantity of life (mortality) and "quality of life" (which is conceptualized as morbidity) [30-33]. The "Q" in QALY is a misnomer given it measures only the health aspects of QOL, the other aspects, which have been elegantly identified by Flanagan, are physical and material well-being, relations with other people, social community and civic activities, personal development and fulfillment, and recreation [34].

The three most widely used utility measures, namely the Health Utilities Index Mark 2 and 3 (HUI2 and HUI3), the EuroQol-5D (EQ-5D) and the Short-Form-6D (SF-6D), label the constructs underlying these measures as health status and/or HRQL [35-39]. None list QOL as the construct being measured. Yet, for economic evaluation, the QALY is the parameter calculated and compared with cost.

In line with guidelines for economic evaluation from agencies such as the National Institute for Health and Clinical Excellence (NICE) and the Canadian Agency for Drugs and Technologies in Health (CADTH), these measures are currently being used to evaluate the cost-effectiveness of different interventions in MS. However, the challenge of using such measures in people with a specific health condition, such as MS, is that they may not capture all of the domains that are impacted upon by the health condition. If important domains are missing from the generic measures, the value derived will be higher than the real impact creating invalid comparisons across interventions and populations.

Personalized measures have been proposed as a method for identifying those aspects of a health condition that impact on QOL. While they may differ from person to person and across health conditions, the value derived from them represents QOL. The most commonly used individualized measures of QOL are the Patient Generated Index (PGI) and the Schedule for the Evaluation of Individual Quality of Life-Direct Weighting (SEIQOL-DW). Both measures capture the individual's perspective on QOL, by permitting him/her to nominate the areas of life that are most important and assign a weight to each domain. Personalized measures of QOL have been used in several clinical trials to evaluate the effectiveness of different interventions on overall QOL [40-44]. Furthermore, these measures have shown to be particularly useful in clinical settings by improving patient-physician communication and by helping prioritize treatment options [45-47].

The global aim of the study is to contribute evidence for the content validity of generic utility measures with respect to capturing the relevant domains for people with MS. The specific objective was to estimate the extent to which generic utility measures capture important domains that are affected by MS.

Methods

Subjects

The data for this study comes from a study of the life-impact of people diagnosed with MS during the era of magnetic resonance imaging (MRI) and disease modifying therapies (the New MS) [48]. The available study population consisted of both men and women who had been registered after 1994 at the three participating MS clinics in Greater Montreal, Quebec, Canada. The study was approved by all regional ethics committees. Inclusion criteria for the study were diagnosis of MS or Clinically Isolated Syndrome (CIS) after 1994. From a pool of 5000 patients, a centre-stratified random sample of 550 patients was drawn, of which 394 were contacted. From those who were contacted, the first 192 persons who responded were enrolled, 189 completed all questionnaires and 185 came for an interview. Respondents and non-respondents were compared and no clinically or statistically significant differences were found between the two groups on socio-demographic characteristics.

Measurement

Patient generated index

The PGI is an individualized measure of HRQL that was administered in three stages. In the first stage, patients were asked to identify up to five of the most important areas of their lives affected by MS. In the second stage, patients were asked to rate how badly affected they were in each of the selected areas on a scale of 0 to 10, where 0 was the worst they can imagine and 10 exactly as they would like to be. A sixth box was provided to rate all other health or non-health related areas. In the third stage, they were given twelve spending "points" or "tokens" to distribute among the areas identified. The tokens that they allocated to each area represented the relative importance of potential improvements in the chosen area. The more tokens a patient spent for an area, the more important that area was. The less tokens a patient spent, the less important that area was. The rating for each area was multiplied by the proportion of "points" for that area, which were then summed together to produce an index from 0 to 100 [49]. For ease of comparison with the utility measures, PGI scores in this study were presented on a scale from 0 to 1.

EQ-5D

The EQ-5D is a generic preference-based measure of HRQL that consists of two parts [50,51]. The first part includes 5 separate domains; mobility, capacity for self-care, conduct of usual activities, pain/discomfort and anxiety/depression. Each domain has 3 levels: no problems, some problems, extreme problems. The second part consists of a Visual Analogue Scale (EQVAS) to measure self-perceived health on a vertical scale from 0 to 100, where 0 is the worst imaginable health state, and 10 is the best imaginable health state. The EQ-5D defines 243 health states, and has a range from −0.6 to 1.0.

SF-6D

The SF-6D is a generic preference-based measure derived from the SF-36 Health Survey (or RAND-36) [23,39]. The SF-6D has 6 domains: physical functioning, role limitation, social functioning, pain, mental health and vitality. Each domain has between 4 and 6 levels. The index defines 18 000 health states, and has a range from 0.3 to 1.0.

Procedure

Figure 1 presents a flowchart of the study procedure.

Subjects were first interviewed on an individualized measure of QOL, the PGI [49]. The domains identified with the PGI were then classified and grouped together using the World Health Organization's International Classification of Functioning, Disability and Health (ICF) [52] independently by four raters. This methodology followed closely that conducted by Mayo et al [53], which evaluated the extent to which HRQL measures captured constructs beyond symptoms and function. The ICF provided a coding framework and standardized description of health related problems at the level of body structure/function (e.g. fatigue, cognition), activity (e.g. dressing, feeding, walking) and participation (e.g. school, work). These levels are also known as impairments, activity limitations and participation restrictions, respectively. Any discrepancies between raters were resolved by discussion.

Last, the domains were mapped onto the HUI2, HUI3, EQ-5D and SF-6D which had been previously mapped to the ICF [53]. The extent to which these utility measures captured domains important to patients with MS was qualitatively appraised.

Data analysis

We had data on hand for the PGI, the EQ-5D and the SF-6D (derived from the RAND-36). As all three

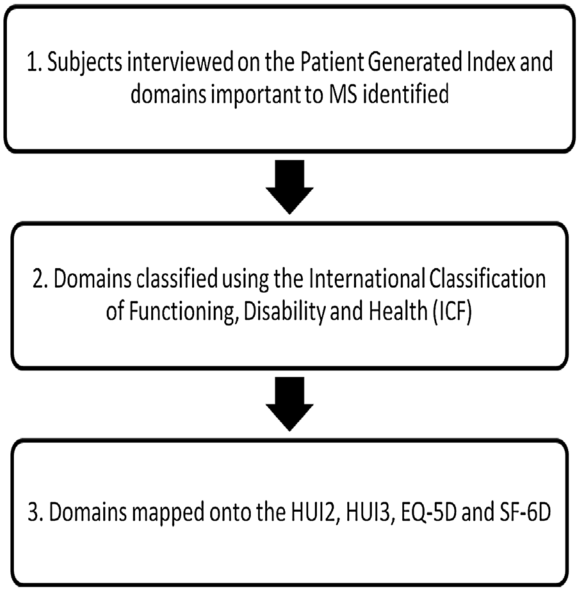

1. Subjects interviewed on the Patient Generated Index and domains important to MS identified
2. Domains classified using the International Classification of Functioning, Disability and Health (ICF)
3. Domains mapped onto the HUI2, HUI3, EQ-5D and SF-6D

Figure 1 Flowchart of the study procedure.

measures were administered on the same individual, generalized estimating equations (GEE) was used to adjust the variance for the clusters of outcome within persons. The advantage of using GEE, as opposed to the paired *t*-test, was that it allowed for simultaneous assessment and correlation among all 3 measures. The regression coefficients produced in the model were estimates of the difference between measures (with 95% CI) adjusted for the correlation among data points. An effect size (ES) was then calculated using the t-statistic, which was equal to the adjusted regression coefficient divided by its SE.

Results

A total of 185 persons with MS were interviewed on the PGI. The sample was relatively young (mean age 43) and predominantly female. Both men and women had mild disability with a median Expanded Disability Status Scale (EDSS) score of 2. The average number of years since diagnosis was 6 years, and 59% of the sample was on Disease Modifying Therapies. Demographic and clinical characteristics are presented in Table 1.

Table 2 presents the top 10 domains that patients identified to be the most affected by their MS. These areas were, work (62%), fatigue (48%), sports (39%), social life (28%), relationships (23%), walking/mobility (22%), cognition (21%), balance (14%), housework (12%) and mood (11%). The mean impact score for each domain (from 0 to 10) ranged from 3.9 to 5.0. In terms of the mean number of points spent for each domain, patients spent the most points (4.3) to improve their relationships, followed by fatigue (3.8) and then walking (mean 3.6).

Table 1 Demographic and clinical characteristics of sample (n = 185)

Characteristics	Mean (SD) or N (%)
Age (y)	42.8 (10.0)
Women/Men	137/48 (74/26)
Definite MS/CIS	170/15 (92/8)
Year since diagnosis	6.2 (3.6)
EDSS, median (IQR)	2.0 (1.0 - 3.5)
On DMT/Not on DMT/No information	110/19/56 (59/10/30)
Patient Generated Index*	0.50 (0.25)
EQ-5D**	0.69 (0.18)
SF-6D***	0.69 (0.13)

SD, standard deviation; N, number; CIS, Clinically Isolated Syndrome; EDSS, Expanded Disability Status Scale; IQR, Inter-quartile range; DMT, Disease Modifying Therapies.
*Transformed to a scale from 0 to 1, higher scores are better (1 = perfect QOL).
**Measured on a scale from −0.4 to 1, higher scores are better (1 = perfect health).
***Measured on a scale from 0.3 to 1, higher scores are better (1 = perfect health).

Table 3 presents the results for the mapping of the 10 domains identified by MS patients against the HUI2, HUI3, EQ-5D and the SF-6D. School/work was found in the EQ-5D and SF-6D but not in the HUI2 or HUI3. Fatigue was found in the SF-6D but not in the EQ-5D or the HUI measures. Sports which was the third most frequently reported domain, was only found in the SF-6D and HUI2. Social life was included in the EQ-5D and the SF-6D, but not in the HUI measures. Cognition was found in the HUI measures, but not in the EQ-5D or the SF-6D. Housework was included in the EQ-5D and the SF-6D, but not in the HUI2 or HUI3. Relationships and balance were not included in any of the utility measures. Mood was the only domain that was included in all of the measures.

The SF-6D included the most number of domains (6 domains) important to people with MS, followed by the EQ-5D (4 domains) and the HUI2 (4 domains), and then the HUI3 (3 domains).

The generic utility measures included domains that were not identified to be important by the sample, such as pain, self-care, vision, hearing, manual dexterity, speech and fertility.

The correlation between the SF-6D and the EQ-5D was 0.58. As demonstrated in Figure 2a, although the relationship between the measures was somewhat linear, discrepancies in scores between the two measures was evident. At the upper end of the scales, a number of individuals who had utility scores of 0.85 on the EQ-5D had scores as low as 0.6 on the SF-6D. A clinically meaningful difference on utility measures is 0.03, indicating that the difference in scores between the two utility measures was important. Discrepancies were also observed at the lower end of the scale, where an individual with a score of 0.12 on the EQ-5D had a score of 0.55 on the SF-6D.

The correlation between the PGI and the EQ-5D was 0.53. As presented in Figure 2b there were important discrepancies in scores between the two measures. Several individuals with very low scores on the PGI (as low as 0.1) had very high scores on the EQ-5D (as high as 0.8). For many individuals, there was also a mismatch between scores obtained using the PGI and those obtained with the EQ-5D (i.e. individuals with scores as low as 0.1 on the PGI had scores of 0.8 on the EQ-5D). Pearson's correlation between the PGI and the SF-6D was 0.53. Similar to what was observed for the EQ-5D; there were discrepancies in scores between the 2 measures, particularly towards the lower end of the scales (Figure 2c).

The impact of a mismatch between domains provided in the generic utility measures and those that are important to people with MS is illustrated by the total scores of the measures. As seen in Figure 3, the mean

Table 2 Top 10 domains identified by subjects using the Patient Generated Index

Domain	Proportion of subjects reporting problem	Degree to which subjects are affected	Number of tokens spent
	N (%)	Mean (SD)*	Mean (SD)**
School/Work	114 (62)	4.2 (3.4)	1.7 (2.0)
Fatigue	88 (48)	4.5 (2.2)	3.8 (2.7)
Sports	73 (39)	4.1 (2.6)	2.9 (2.4)
Social life	52 (28)	4.7 (2.4)	1.8 (2.6)
Relationships	43 (23)	4.8 (3.4)	4.3 (2.6)
Walking	41 (22)	3.9 (2.5)	3.6 (2.5)
Cognition	39 (21)	4.7 (2.1)	2.8 (2.2)
Balance	25 (14)	5.0 (2.3)	2.5 (3.3)
Housework	23 (12)	4.8 (2.1)	1.3 (1.0)
Mood	21 (11)	4.6 (2.4)	3.4 (2.6)

*Scored out of 10, higher is better (not affected).
**Scored out of 12, higher indicates that the domain was more important.

Table 3 The domains identified by MS subjects compared with items in generic utility measures

Measure	HUI2	HUI3	EQ-5D	SF-6D
Construct	Health status & HRQL [35,36]	Health status & HRQL [36,37]	HRQL [38]	Health status [39]
MS Domains				
School/Work	N	N	Y	Y
Fatigue	N	N	N	Y
Sports	Y	N	N	Y
Social life	N	N	N	Y
Relationships	N	N	N	N
Cognition	Y	Y	N	N
Walking	Y	Y	Y	N
Housework	N	N	Y	Y
Balance	N	N	N	N
Mood*	Y	Y	Y	Y
Total Yes (out of 10)	4	3	4	6
Not MS Domains				
Pain	Y	Y	Y	Y
Self-care	Y	N	Y	Y
Vision	Y	Y	N	N
Hearing	Y	Y	N	N
Manual dexterity	N	Y	N	N
Speech	Y	Y	N	N
Fertility	Y	N	N	N

MS Domains ordered from the largest to the smallest proportion of people with MS who identified that domain.
Y, Yes; N, No; HUI2, Health Utilities Index Mark 2; HUI3, Health Utilities Index Mark 3; SF-6D, EQ-5D, EuroQol-5D; Short-Form 6D.
*In the HUI3 this was happiness.

and standard deviation (SD) for the PGI, EQ-5D and the SF-6D were 0.50 (SD 0.25), 0.69 (SD 0.18) and 0.69 (SD 0.13), respectively. The magnitude of difference between the PGI and the 2 utility measures was 0.19 (95% CI 0.16 to 0.22) with ES equal to 12.

This mismatch was also present at the item level. A total of 41 subjects (22% of the sample) reported walking to be an important aspect of their QOL. The distribution of scores on the degree to which walking was affected for these subjects is presented in Figure 4. The impact was measured on a scale from 0 to 10 on the PGI, where 0 was the worst they could imagine and 10 was exactly as they would like to be. These scores were compared with the responses on the EQ-5D mobility item. 12 subjects out of 41 reported having no problems with walking on the EQ-5D. These people were expected to have a score of 10 on the PGI. Only 1 person reported a score of 10 on the PGI. All other subjects reported scores lower than this, scores as low as 3 (poor).

Discussion

In this study, subjects with MS were interviewed on an individualized measure to evaluate the impact of the disease on their QOL. The results of the interview generated a list of domains that were most important to the QOL of persons with MS. The domains identified were work, fatigue, sports, social life, relationships, walking, cognition, balance, housework and mood. These were then mapped onto generic utility measures to estimate the extent to which they captured domains that were important to persons with MS.

There was no one generic utility measure that captured all of the domains important to persons with MS. For example, fatigue, which affects 75 to 90% of patients with MS [54-57] was not included in the EQ-5D or the HUI measures. Walking, another commonly reported symptom was not found in the SF-6D. Cognition was not found in the EQ-5D or the SF-6D. Work, sports, and social life were not found in the HUI2 or HUI3. This was not surprising as the HUI measures were developed with the intention of evaluating 'within-the-skin' experiences that excluded social interaction [58-60]. Balance and relationships were not included in any of the utility measures.

The generic utility measures were clearly missing domains that were important to people with MS. Out of the 10 domains that persons with MS identified as being central to their QOL, only 3 of them were included in the HUI2, 4 were included in the HUI3, 4 were included in the EQ-5D and 6 were included in the SF-6D. Furthermore, the generic utility measures included several

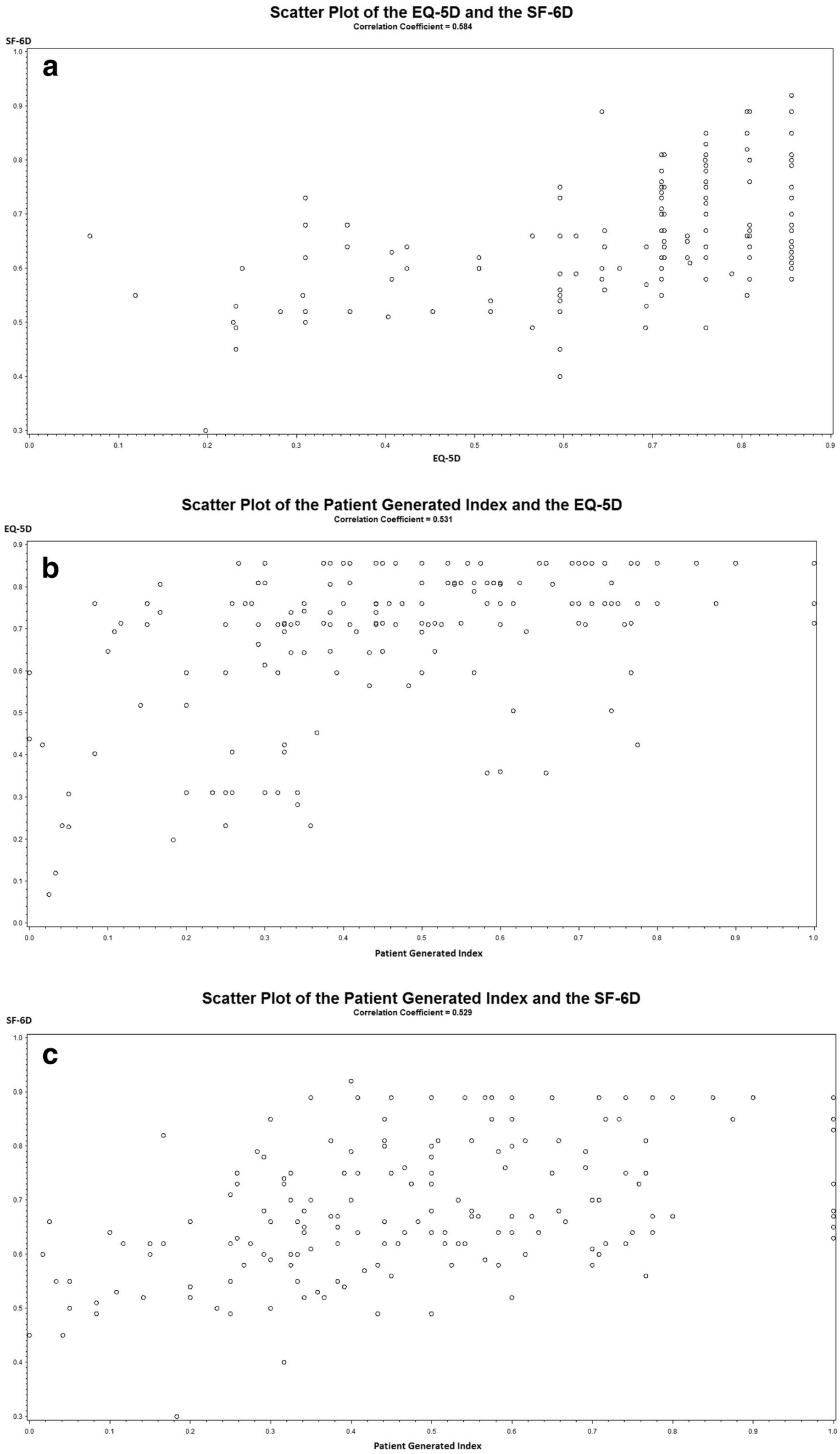


Figure 2 Relationship between the EQ-5D, the SF-6D and the Patient Generated Index. a: Scatter plot of the relationship between the EQ-5D and the SF-6D. **b**: Scatter plot of the relationship between the Patient Generated Index and the EQ-5D. **c**: Scatter plot of the relationship between the Patient Generated Index and the SF-6D.

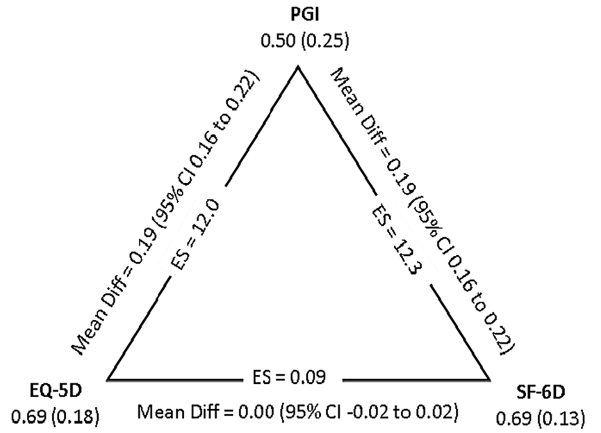

PGI
0.50 (0.25)
Mean Diff = 0.19 (95% CI 0.16 to 0.22)
ES = 12.0
Mean Diff = 0.19 (95% CI 0.16 to 0.22)
ES = 12.3
EQ-5D
0.69 (0.18)
ES = 0.09
Mean Diff = 0.00 (95% CI -0.02 to 0.02)
SF-6D
0.69 (0.13)

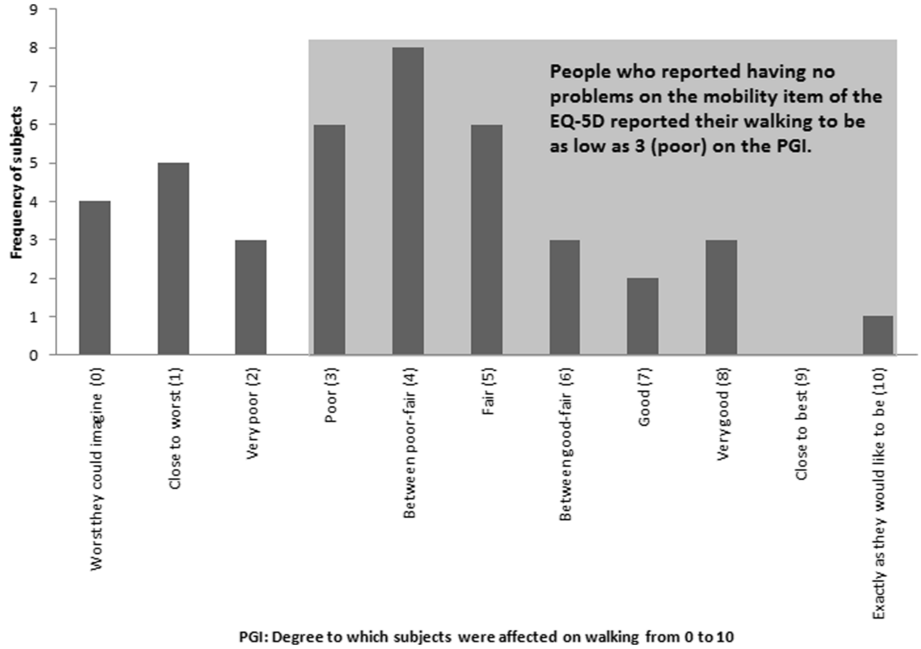
Figure 3 **Mean and standard deviation values for the PGI, EQ-5D and SF-6D, with differences and 95% CI calculated using generalized estimating equations.**

domains that were not important to persons which were sampled in the study, such as pain, self-care, hearing and manual dexterity.

To tackle the issue of lack of content validity, one emerging area of interest in the literature is the development of disease specific "bolt-ons" or dimension extensions to generic utility measures [51]. Another emerging area of interest is the development of disease-specific utility measures, which have been developed for stroke [61], pulmonary hypertension [62], asthma [63], rhinitis [64], urinary incontinence [65] and erectile dysfunction [66]. Recently, Versteegh et al. [67] derived a MS specific utility measure from the Multiple Sclerosis Impact Scale-29 (MSIS-29) using Rasch analysis. The authors selected 8 out of 29 items from the original questionnaire. Some important dimensions such as social life, work and mood were included while others such as walking, sports and physical fatigue were omitted.

There are several potential benefits to using disease specific utility measures in clinical and cost-effectiveness research. First, disease specific utility measures are designed to include domains that are specific to a disease, and therefore, are likely to be more sensitive to smaller change over time than generic measures. Second, not only do these measures provide descriptive information on the various dimensions of health, but also provide a value for each one, thus allowing trade-offs to be made between the domains. Disease-specific utility measures serve the potential to overcome one of the challenges associated with disease specific health profiles - that domains cannot be combined into a single index, which makes it difficult to conclude whether an intervention was effective or not. For example, if a treatment has a positive effect on physical health but a negative one on mental health, unless we know the relative importance attached to each domain, it is impossible to determine whether the intervention resulted in a net improvement or decline in QOL/HRQL. Furthermore, disease-specific utility measures can be used to calculate QALYs and make decisions on the cost-effectiveness of different treatments in MS.

A clinician reported outcome (ClinRO) is an assessment of the status of a patient's health condition that is



People who reported having no problems on the mobility item of the EQ-5D reported their walking to be as low as 3 (poor) on the PGI.

Figure 4 **Frequency and distribution of PGI scores on the degree to which walking was affected from 0 (worst they can imagine) to 10 (exactly as they would like to be).**

made by an observer with professional training (i.e. clinician) [69]. ClinRO are commonly used for endpoints that cannot be directly measured by the patient (e.g. EDSS to quantify level of disability in MS). An observer-reported outcome (ObsRO) is an assessment that is made by an observer without professional training (i.e. non-clinician observer such as a teacher or caregiver) [69]. This type of evaluation is typically used when the patient is unable to self-report. A patient reported outcome (PRO) is any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or other observer (e.g. symptoms, QOL, HRQL) [68,69]. PROs play a complementary role in outcome assessment by providing evidence on the benefit or harm of a treatment from the patient's perspective. Utility measures are one type of PRO. In outcome assessment, utility measures not only provide information on the benefits and harms of a treatment, but are also useful for economic applications by producing QALYS. This information can provide policy and decision makers with a means of evaluating the costs and cost-effectiveness of different treatment options for a health condition.

The first step in evaluating the validity of scores produced by a PRO is an assessment of content validity, before any other forms of validity (i.e. construct validity) are undertaken. Content validity of a PRO can be judged only by the individuals or populations being assessed (i. e. the patients themselves). The global aim of this study was to address this very question of whether generic utility measures captured domains that were important or relevant to people with MS. The results of this study suggest that many important domains in MS are not captured by generic utility measures, therefore questioning the content validity of such measures in MS. This in turn, adds doubt to the interpretability or meaningfulness of scores produced by these measures for this population.

It is important to target measures to people to ensure that the impact of a disease and its treatment are adequately and reliably captured in a clinical trial [70,71]. If a PRO includes domains that are not impacted upon by the disease or its treatment, it will not be able to capture clinically meaningful change. By targeting to the disease, measures are more likely to be sensitive to small but important clinical changes. Furthermore, the ability of PROs to detect small changes is important in determining the statistical power or the necessary sample size required for a clinical trial [72].

The results of our study revealed that the commonly used 4 generic utility measures (HUI2, HUI3, EQ-5D and SF-6D) do not capture the majority of domains important to MS. Among these generic measures, the SF-6D captured the most number of domains (6 domains) that were important to MS. Our findings suggest that the SF-6D, compared to the other generic utility measures, may be the most appropriate one to use in MS. The PGI index can be used to evaluate the clinical effectiveness of different interventions in MS. However, because the PGI was not developed using multi-attribute utility theory (hence is not a utility measure); it cannot be used for cost-utility analysis.

Ideas for future directions that build directly from this work are the use of MS specific "bolt-on" items or dimensions to generic utility measures [73]. This study has identified potential items important to MS, such as fatigue that can be used as add-ons to existing generic utility measures. Other areas of potential research that can build directly from this work are the development of an MS specific utility measure that will only include dimensions pertinent to the disease.

A particular feature of this study is that we purposely sampled people with MS diagnosed in the era of Magnetic Resonance Imaging (MRI) technology and availability of disease modifying drugs [48]. As these are the people who are faced with treatment decisions, a method of valuing changes on the most important domains of QOL affected by MS would be the most relevant for this population.

Conclusions

Generic utility measures are designed to include a common set of dimensions that most people will value highly, therefore underrepresenting those dimensions that may be specific to a particular disease. Although the generic utility measures included certain items that were important to people with MS, there were several that were missing. An important consequence of this mismatch was that values of QOL derived from the PGI were importantly and significantly lower than those estimated using any of the generic utility measures. This could have a substantial impact for evaluating the effect of interventions in people with MS. The overestimation in scores obtained with utility measures may not have an impact at the start of a clinical trial, but they will have an impact at follow-up. If scores are high at baseline, there will likely be no room for improvement on the scale, resulting in the false conclusion that the treatment group did not change post-treatment. When in reality, the treatment may have had a positive effect but the measure being administered was not able to detect this. Then the difference between the treatment and control group (assuming the control also does not change), would be zero. In addition, an intervention that is in fact beneficial to fatigue, for example, would also risk not to show change on a generic measure because this item was not included. When choosing the right outcome measure for an intervention, it is essential to choose one with items that can or should be affected by the intervention. Given that the MS specific items do impact on QOL, not including these items would result in a false estimate of QALYs and bias the evaluation of the cost-effectiveness of interventions in MS.

Abbreviations

MS: Multiple Sclerosis; HUI2: Health Utilities Index Mark 2; HUI3: Health Utilities Index Mark 3; EQ-5D: EuroQol-5D; SF-6D: Short-Form-6D; CADTH: Canadian Agency for Drugs and Technologies in Health; QOL: Quality of life; HRQL: Health-related quality of life; QALY: Quality-Adjusted Life Year; MRI: Magnetic resonance imaging; PGI: Patient Generated Index; ICF: International Classification of Functioning, Disability and Health; ES: Effect size; EDSS: Expanded Disability Status Scale.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

NM was the principal investigator of the study and AK collected the data. Both authors contributed to writing the article, data analysis and interpretation. Both authors read and approved the final manuscript.

Authors' information

AK is a physiotherapist and a PhD Candidate in Rehabilitation Sciences at the School of Physical and Occupational Therapy, McGill University. NM is a James McGill Professor in the Department of Medicine and the School of Physical and Occupational Therapy, McGill University.

Acknowledgements

We thank Dr. Sara Ahmed, Catherine Campanella, Alexander Chubunov, Alain Desir, and Olivier Mayrand-De Arteche for their assistance with the mapping of the Patient Generated Index items onto the ICF. This work was supported in part by the Canadian Institutes of Health Research and a scholarship from the Fonds de la Recherche en Santé du Quebec.

Author details

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, 3654 Promenade Sir-William-Osler, Montreal 3G 1Y5, QC, Canada. [2]Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal, QC, Canada.

Received: 14 December 2012 Accepted: 17 April 2013
Published: 25 April 2013

References

1. Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG: **Multiple sclerosis.** *N Engl J Med* 2000, **343:**938–952.
2. Burks J, Johnson K: Multiple S: *Diagnosis, Medical Management, and Rehabilitation.* New York: Demos Medical; 2000.
3. Miller DM, Moore SM, Fox RJ, Atreja A, Fu AZ, Lee JC, *et al*: **Web-based self-management for patients with multiple sclerosis: a practical, randomized trial.** *Telemed J E-Health* 2011, **17:**5–13.
4. Barlow J, Turner A, Edwards R, Gilchrist M: **A randomised controlled trial of lay-led self-management for people with multiple sclerosis.** *Patient Educ Couns* 2009, **77:**81–89.
5. Bombardier CH, Cunniffe M, Wadhwani R, Gibbons LE, Blake KD, Kraft GH: **The Efficacy of Telephone Counseling for Health Promotion in People With Multiple Sclerosis: A Randomized Controlled Trial.** *Arch Phys Med Rehabil* 2008, **89**(10):1849–1856.
6. McAuley E, Motl RW, Morris KS, Hu L, Doerksen SE, Elavsky S, *et al*: **Enhancing physical activity adherence and well-being in multiple sclerosis: a randomised controlled trial.** *Mult Scler* 2007, **13:**652–659.
7. Grossman P, Kappos L, Gensicke H, D'Souza M, Mohr DC, Penner IK, *et al*: **MS quality of life, depression, and fatigue improve after mindfulness training: a randomized trial.** *Neurology* 2010, **75:**1141–1149.
8. Forman AC, Lincoln NB: **Evaluation of an adjustment group for people with multiple sclerosis: a pilot randomized controlled trial.** *Clin Rehabil* 2010, **24:**211–221.
9. Cosio D, Jin L, Siddique J, Mohr DC: **The effect of telephone-administered cognitive-behavioral therapy on quality of life among patients with multiple sclerosis.** *Ann Behav Med* 2011, **41:**227–234.
10. Kavia RB, De RD, Constantinescu CS, Stott CG, Fowler CJ: **Randomized controlled trial of Sativex to treat detrusor overactivity in multiple sclerosis.** *Mult Scler* 2010, **16:**1349–1359.
11. Moller F, Poettgen J, Broemel F, Neuhaus A, Daumer M, Heesen C: **HAGIL (Hamburg Vigil Study): A randomized placebo-controlled double-blind study with modafinil for treatment of fatigue in patients with multiple sclerosis.** *Mult Scler* 2011, **17**(8):1002–1009.
12. Freeman JA, Thompson AJ, Fitzpatrick R, Hutchinson M, Miltenburger C, Beckmann K, *et al*: **Interferon-beta1b in the treatment of secondary progressive MS: impact on quality of life.** *Neurology* 2001, **57:**1870–1875.
13. Rudick RA, Miller D, Hass S, Hutchinson M, Calabresi PA, Confavreux C, *et al*: **Health-related quality of life in multiple sclerosis: effects of natalizumab.** *Ann Neurol* 2007, **62:**335–346.
14. Fox RJ, Miller DH, Phillips JT, Hutchinson M, Havrdova E, Kita M, *et al*: **Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis.** *N Engl J Med* 2012, **367:**1087–1097.
15. Zamboni P, Menegatti E, Galeotti R, Malagoni AM, Tacconi G, Dall'Ara S, *et al*: **The value of cerebral Doppler venous haemodynamics in the assessment of multiple sclerosis.** *J Neurol Sci* 2009, **282:**21–27.
16. Al-Omari MH, Rousan LA: **Internal jugular vein morphology and hemodynamics in patients with multiple sclerosis.** *Int Angiol* 2010, **29:**115–120.
17. Baracchini C, Perini P, Calabrese M, Causin F, Rinaldi F, Gallo P: **No evidence of chronic cerebrospinal venous insufficiency at multiple sclerosis onset.** *Ann Neurol* 2011, **69:**90–99.
18. Centonze D, Floris R, Stefanini M, Rossi S, Fabiano S, Castelli M, *et al*: **Proposed chronic cerebrospinal venous insufficiency criteria do not predict multiple sclerosis risk or severity.** *Ann Neurol* 2011, **70:**51–58.
19. Zivadinov R, Marr K, Cutter G, Ramanathan M, Benedict RH, Kennedy C, *et al*: **Prevalence, sensitivity, and specificity of chronic cerebrospinal venous insufficiency in MS.** *Neurology* 2011, **77:**138–144.
20. *Soc Sci Med***The World Health Organization Quality of Life assessment (WHOQOL): position paper from the World Health Organization.** 1995, **41:**1403–1409.
21. Ware JE Jr: **Standards for validating health measures: definition and content.** *J Chronic Dis* 1987, **40:**473–480.
22. Breslow L: **A quantitative approach to the World Health Organization definition of health: physical, mental and social well-being.** *Int J Epidemiol* 1972, **1:**347–355.
23. Brazier J, Roberts J, Deverill M: **The estimation of a preference-based measure of health from the SF-36.** *J Health Econ* 2002, **21:**271–292.
24. World Health Organization: *Glossary of Terms Used in the "Health For All" Series.* ; 1984. Ref Type: Report.
25. Kind P: **Values and valuation in the measurement of HRQoL.** In *Assessing quality of life in clinical trials.* 2nd edition. Edited by Fayers P, Hays D. New York: Oxford University Press Inc; 2005:391–404.
26. Guyatt GH, Feeny DH, Patrick DL: **Measuring health-related quality of life.** *Ann Intern Med* 1993, **118:**622–629.
27. Feeny DH, Torrance GW: **Incorporating utility-based quality-of-life assessment measures in clinical trials. Two examples.** *Med Care* 1989, **27:**S190–S204.
28. Torrance GW: **Utility approach to measuring health-related quality of life.** *J Chronic Dis* 1987, **40:**593–603.
29. Torrance GW: **Measurement of health state utilities for economic appraisal.** *J Health Econ* 1986, **5:**1–30.
30. Feeny D: **Preference-based measures: utility and quality-adjusted life years.** In *Assessing quality of life in clinical trials.* 2nd edition. Edited by Fayers P, Hays D. New York: Oxford University Press Inc; 2005:405–429.
31. Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A: *Measuring and valuing health benefits for economic evaluation.* New York: Oxford University Press Inc.; 2007.
32. Kind P, Lafata JE, Matuszewski K, Raisch D: **The use of QALYs in clinical and patient decision-making: issues and prospects.** *Value Health* 2009, **12**(Suppl 1):S27–S30.
33. Hawthorne G, Richardson J: **Measuring the value of program outcomes: a review of multiattribute utility measures.** *Expert Rev Pharmacoecon Outcomes Res* 2001, **1:**215–228.
34. Flanagan JC: **Measurement of quality of life: current state of the art.** *Arch Phys Med Rehabil* 1982, **63:**56–59.

35. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q: **Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2.** *Med Care* 1996, **34:**702–722.
36. Horsman J, Furlong W, Feeny D, Torrance G: **The Health Utilities Index (HUI): concepts, measurement properties and applications.** *Health Qual Life Outcomes* 2003, **1:**54.
37. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, *et al*: **Multiattribute and single-attribute utility functions for the health utilities index mark 3 system.** *Med Care* 2002, **40:**113–128.
38. Gudex C: **The descriptive system of the EuroQOL instrument.** In *EQ-5D concepts and methods: a developmental history.* Edited by Kind P, Brooks R, Rabin R. Dordrecht: Springer; 2005:19–33.
39. Brazier J, Usherwood T, Harper R, Thomas K: **Deriving a preference-based single index from the UK SF-36 Health Survey.** *J Clin Epidemiol* 1998, **51:**1115–1128.
40. Ahmed S, Mayo NE, Wood-Dauphinee S, Hanley JA, Cohen SR: **Using the Patient Generated Index to evaluate response shift post-stroke.** *Qual Life Res* 2005, **14:**2247–2257.
41. Goldstein RS, Gort EH, Stubbing D, Avendano MA, Guyatt GH: **Randomised controlled trial of respiratory rehabilitation.** *Lancet* 1994, **344:**1394–1397.
42. Lacasse Y, Wong E, Guatt GH: **Individualising questionnaires.** In *Individual quality of life: Approaches to conceptualization and assessment.* Edited by Joyce CR, O'Boyle CA, McGee H. Amsterdam: Harwood Academic Publishers; 1999:87–103.
43. Simpson K, Killian K, McCartney N, Stubbing DG, Jones NL: **Randomised controlled trial of weightlifting exercise in patients with chronic airflow limitation.** *Thorax* 1992, **47:**70–75.
44. Wijkstra PJ, Van AR, Kraan J, Otten V, Postma DS, Koeter GH: **Quality of life in patients with chronic obstructive pulmonary disease improves after rehabilitation at home.** *Eur Respir J* 1994, **7:**269–273.
45. Kettis-Lindblad A, Ring L, Widmark E, Bendtsen P, Glimelius B: **Patients' and doctors' views of using the schedule for individual quality of life in clinical practice.** *J Support Oncol* 2007, **5:**281–287.
46. Detmar SB, Muller MJ, Schornagel JH, Wever LD, Aaronson NK: **Health-related quality-of-life assessments and patient-physician communication: a randomized controlled trial.** *JAMA* 2002, **288:**3027–3034.
47. Patel KK, Veenstra DL, Patrick DL: **A review of selected patient-generated outcome measures and their application in clinical trials.** *Value Health* 2003, **6:**595–603.
48. Mayo N: **Setting the agenda for multiple sclerosis rehabilitation research.** *Mult Scler* 2008, **14:**1154–1156.
49. Ruta DA, Garratt AM, Leng M, Russell IT, MacDonald LM: **A new approach to the measurement of quality of life. The Patient-Generated Index.** *Med Care* 1994, **32:**1109–1126.
50. Dolan P: **Modeling valuations for EuroQol health states.** *Med Care* 1997, **35:**1095–1108.
51. Shaw JW, Johnson JA, Coons SJ: **US valuation of the EQ-5D health states: development and testing of the D1 valuation model.** *Med Care* 2005, **43:**203–220.
52. International Classification of Functioning: *Disability and Health (ICF).* Geneva: World Health Organization; 2001.
53. Mayo NE, Moriello C, Asano M, van der Spuy S, Finch L: **The extent to which common health-related quality of life indices capture constructs beyond symptoms and function.** *Qual Life Res* 2011, **20:**621–627.
54. Krupp LB, Pollina DA: **Mechanisms and management of fatigue in progressive neurological disorders.** *Curr Opin Neurol* 1996, **9:**456–460.
55. Fisk JD, Pontefract A, Ritvo PG, Archibald CJ, Murray TJ: **The impact of fatigue on patients with multiple sclerosis.** *Can J Neurol Sci* 1994, **21:**9–14.
56. Freal JE, Kraft GH, Coryell JK: **Symptomatic fatigue in multiple sclerosis.** *Arch Phys Med Rehabil* 1984, **65:**135–138.
57. Murray TJ: **Amantadine therapy for fatigue in multiple sclerosis.** *Can J Neurol Sci* 1985, **12:**251–254.
58. Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S: **A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer.** *J Clin Oncol* 1992, **10:**923–928.
59. Feeny D, Furlong W, Boyle M, Torrance GW: **Multi-attribute health status classification systems. Health Utilities Index.** *Pharmacoeconomics* 1995, **7:**490–502.
60. Feeny D, Torrance GW, Furlong W: **Health utilities index.** In *Quality of Life and Pharmaeconomics in Clinicals Trials.* 2nd edition. Edited by Spilker B. Philadelphia: Lippincott-Raven Publishers; 1996:239–252.
61. Poissant L, Mayo NE, Wood-Dauphinee S, Clarke AE: **The development and preliminary validation of a Preference-Based Stroke Index (PBSI).** *Health Qual Life Outcomes* 2003, **1:**43.
62. McKenna SP, Ratcliffe J, Meads DM, Brazier JE: **Development and validation of a preference based measure derived from the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses.** *Health Qual Life Outcomes* 2008, **6:**65.
63. Revicki DA, Leidy NK, Brennan-Diemer F, Sorensen S, Togias A: **Integrating patient preferences into health outcomes assessment: the multiattribute Asthma Symptom Utility Index.** *Chest* 1998, **114:**998–1007.
64. Revicki DA, Leidy NK, Brennan-Diemer F, Thompson C, Togias A: **Development and preliminary validation of the multiattribute Rhinitis Symptom Utility Index.** *Qual Life Res* 1998, **7:**693–702.
65. Brazier J, Czoski-Murray C, Roberts J, Brown M, Symonds T, Kelleher C: **Estimation of a preference-based index from a condition-specific measure: the King's Health Questionnaire.** *Med Decis Making* 2008, **28:**113–126.
66. Torrance GW, Keresteci MA, Casey RW, Rosner AJ, Ryan N, Breton MC: **Development and initial validation of a new preference-based disease-specific health-related quality of life instrument for erectile function.** *Qual Life Res* 2004, **13:**349–359.
67. Versteegh MM, Leunis A, Uyl-de Groot CA, Stolk EA: **Condition-specific preference-based measures: benefit or burden?** *Value Health* 2012, **15:**504–513.
68. Food US: **Drug Administration: Guidance for industry: Patient-reported outcome measures. Use in medical product development to support labeling claims.** *Fed Regist* 2009, **74:**65132–65133.
69. Velentgas P, Dreyer NA, Wu AW: **Outcome definition and measurement.** In *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. AHRQ Publication No. 12(13)-EHC099.* Edited by Velentgas P, Dreyer NA, Nourjah P, *et al.* Rockville, MD: Agency for Healthcare Research and Quality; 2013:71–92.
70. Hays RD: **Generic versus disease-targeted instruments.** In *Assessing quality of life in clinical trials.* 2nd edition. Edited by Fayers P, Hays D. New York: Oxford University Press Inc; 2005:3–8.
71. Guyatt GH, Bombardier C, Tugwell PX: **Measuring disease-specific quality of life in clinical trials.** *CMAJ* 1986, **134:**889–895.
72. Mayo NE: **Randomized trials and other parallel comparisons of treatments.** In *Medical Uses of Statistics.* 3rd edition. Edited by Bailar JC, Hoaglin DC. Hoboken: John Wiley & Sons, Inc; 2009:51–89.
73. Lin FJ, Longworth L, Pickard AS: **Evaluation of content on EQ-5D as compared to disease-specific utility measures.** *Qual Life Res.* Epub ahead of print.

doi:10.1186/1477-7525-11-71
Cite this article as: Kuspinar and Mayo: **Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis?.** *Health and Quality of Life Outcomes* 2013 **11**:71.

CHAPTER 6: Integration of Manuscripts 2 and 3

Research questions of manuscripts 2 and 3

Manuscript 2:

Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis?

Manuscript 3:

A review of the psychometric properties of generic utility measures in multiple sclerosis.

Integration of Manuscripts 2 and 3

In the second manuscript we identified the domains that were important to the quality of life of people with MS and then mapped these domains onto generic preference-based measures. Our results revealed that existing generic preference-based measures lacked content validity in MS as they were missing important domains, such as fatigue and cognition. When choosing the right outcome measure for an intervention, it is essential to choose one with items that can or should be affected by the disease and intervention.

Content validity is one type of psychometric property. In the next manuscript we will delve deeper into the topic by evaluating additional psychometric properties such as construct validity and reliability. To do this, we not only used data from the Gender and Life Impact of Multiple Sclerosis Study, but also conducted a comprehensive literature search to identify all possible studies that evaluated the validity and reliability of existing generic preference-based measures in MS.

CHAPTER 7 (MANUSCRIPT 3)

A review of the psychometric properties of generic utility measures in multiple sclerosis

Ayse Kuspinar[1] and Nancy E. Mayo[1,2]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada

[2]Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

Published in *PharmacoEconomics* (2014) 32:759-773.

Communication addressed to:

Ayse Kuspinar, M.Sc., Ph.D. Candidate
School of Physical & Occupational Therapy
Faculty of Medicine, McGill University
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
Tel: 514-934-1934 ext 31564
E-mail: ayse.kuspinar@mail.mcgill.ca

PharmacoEconomics (2014) 32:759–773
DOI 10.1007/s40273-014-0167-5

SYSTEMATIC REVIEW

A Review of the Psychometric Properties of Generic Utility Measures in Multiple Sclerosis

Ayse Kuspinar · Nancy E. Mayo

Published online: 21 May 2014

Abstract

Objective The reliability and validity of generic utility measures have not yet been summarized in people with multiple sclerosis (MS). It is important to assess the psychometric properties of these measures, to ensure that the values obtained by the scoring system are valid for interpretation and utilization by clinicians, researchers and policy makers. Therefore, the objective of this review was to summarize the evidence from published literature on the psychometric properties of generic utility measures in MS.

Methods A structured literature search was conducted by using multiple electronic databases. All potentially relevant abstracts and full-text articles were read to identify publications that may be eligible for inclusion in the review. A meta-analysis was conducted to combine correlation coefficient values for convergent validity. The Schmidt–Hunter method, a weighted mean of the correlation coefficient values, was used. Heterogeneity, the percentage of total variation across studies that is due to between-study differences rather than chance, was assessed using the I^2 statistic.

Results The following generic utility measures were identified: the EQ-5D ($n = 9$)/EQ-5D-5 Level (EQ-5D-5L) ($n = 1$), followed by the Health Utilities Index Mark 3/2 (HUI2/HUI3) ($n = 3$), the SF-6D ($n = 2$), the Assessment of Quality of Life (AQOL) ($n = 2$), and the Quality of Well-Being (QWB) scale ($n = 1$). Ceiling and floor effects were present for the EQ-5D and the SF-6D, but not for the HUI3. The EQ-5D, the SF-6D and the HUI3 demonstrated excellent reliability. In terms of discriminative ability, the SF-6D and the QWB scale were not able to differentiate between moderately and severely disabled MS patients, and the EQ-5D was not able to differentiate between those who were mildly and moderately disabled. The AQOL and the HUI3, on the other hand, demonstrated good discriminative ability, as both measures were able to differentiate between all levels of disability. As for convergent validity, the HUI2/HUI3 were highly correlated ($r = 0.7$) against measurement instruments that evaluated impairments such as disease severity, ambulation and manual dexterity. The EQ-5D, SF-6D and the QWB scale demonstrated small to moderate correlations ($r = 0.4$) against instruments evaluating impairments, and slightly stronger correlations against measures of activity limitations/participation restrictions and health-related quality of life (HRQL) ($r = 0.6$).

Conclusion To our knowledge this is the first study to review the validity and reliability of generic utility measures in MS. The HUI3 demonstrated the strongest psychometric properties when compared with other utility measures. However, the HUI3 only measures impairment and excludes important components of HRQL such as participation restrictions. The EQ-5D, the SF-6D and the QWB scale, on the other hand, do include items on participation. However, these measures demonstrated a lack of content validity in MS by missing certain domains that were important to the disease, as well as difficulty in differentiating between different levels of disability. The

Electronic supplementary material The online version of this article (doi:10.1007/s40273-014-0167-5) contains supplementary material, which is available to authorized users.

A. Kuspinar (✉) · N. E. Mayo
Faculty of Medicine, School of Physical and Occupational Therapy, McGill University, 3654 Prom Sir William Osler, Montreal, QC H3G 1Y5, Canada
e-mail: ayse.kuspinar@mail.mcgill.ca

N. E. Mayo
Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

addition of MS-specific 'bolt-ons' to generic utility measures and the development of an MS specific utility measure are possible areas of exploration for future research.

Key Points for Decision Makers

This structured review summarizing the published literature on the reliability and validity of generic utility measures in multiple sclerosis (MS) showed that each of the utility measures had their strengths and weaknesses.

The Health Utilities Index Mark 3 (HUI3) demonstrated the strongest psychometric properties when compared with other utility measures. However, the HUI3 only measures impairment and excludes important components of health-related quality of life, such as activity limitations and participation restrictions.

The EQ-5D, the SF-6D and the Quality of Well-Being (QWB) scale, on the other hand, did include items on participation in life roles. However, these measures demonstrated a lack of content validity in MS by missing certain domains that were important to the disease, as well as difficulty in differentiating between different levels of disability.

1 Introduction

Health care has a dual aim of improving quality of life and extending life expectancy. The quality-adjusted life-year (QALY) was developed to capture both of these goals. When making decisions on whether an intervention should be made available within a health care system, policy makers are often interested in the cost per QALY associated with an intervention. Generic utility measures or preference-based measures, such as the EQ-5D [1, 2] and the SF-6D [3], are usually administered on patients to capture the 'Q' in the QALY.

The assumption underlying generic utility measures is that they can make comparisons across all types of diseases and interventions. This assumption has been proven to be true for many health conditions, where these measures have passed psychometric tests of reliability and validity [4–7]. However, the validity of these measures has been questioned for other health conditions [8–11]. For example, the mobility domain of the EQ-5D consists of three response levels: 'I have no problems walking about' or 'I have some problems in walking about' or 'I am confined to bed'. The response option 'I have some problems in walking about' covers a wide range of gait disability, as it is the only level between 'no problems' and 'confined to bed'. In a study involving both patients with stroke and multiple sclerosis (MS) [12], those who reported having 'moderate' problems walking about had varying levels of function. Patients' mobility ranged from those who used a cane occasionally in public, through to those who were confined to a wheelchair most of the time but could still transfer from the wheelchair to their bed.

Furthermore, ceiling effects and floor effects have also been reported for these measures [13–15]. Brazier et al. [15] compared the SF-6D and the EQ-5D in seven different patient populations, namely low back pain, chronic obstructive pulmonary disease, irritable bowel syndrome, leg ulcer, menopausal women and osteoporosis. The EQ-5D had a larger percentage of the participants in the top category of each dimension compared with the SF-6D (i.e., 17–72 % for the EQ-5D compared with 4–35 % for the SF-6D). Conversely, the SF-6D had a larger proportion of the participants on the lowest level of physical functioning and role limitation than did the EQ-5D on mobility and usual activities (i.e., 25 and 38 % for the SF-6D vs. 0.2 and 10.5 % for the EQ-5D).

MS is a chronic, demyelinating disease of the central nervous system that has a significant impact on patients' level of functioning and disability [16]. It is associated with a variety of health-related problems such as fatigue, muscle weakness, altered sensation, limitations in carrying out daily activities and restrictions with participation in life roles. The reliability and validity of generic utility measures have not yet been summarized in this population. It is important to assess the psychometric properties of these measures, to ensure that the values obtained by the scoring system are valid for interpretation and utilization by clinicians, researchers and policy makers [17]. Generic utility measures that do not have good psychometric properties may result in a false estimate of QALYs and bias the evaluation of the cost effectiveness of different interventions in MS.

Therefore, the objective of this review was to summarize the evidence from published literature on the psychometric properties of generic utility measures in MS.

2 Methods

We conducted a structured search to identify all possible articles that provided information on the psychometric properties of generic utility measures in MS.

2.1 Search Strategy

Potentially relevant articles were identified by searching the following databases: OVID MEDLINE (1946 to October 8, 2013), EMBASE (1980 to October 8, 2013), Cumulative Index to Nursing and Allied Health Literature (1960 to October 8, 2013) and Cochrane Central Register of Controlled Trials (1960 to October 2013). These electronic databases were searched using the following terms: multiple sclerosis AND (Health Utilities Index OR HUI2 OR HUI3 OR EQ-5D OR EuroQol OR 15D OR SF-6D OR SF6D OR Assessment of Quality of Life OR AQOL OR Quality of Well-Being OR QWB). Medical subject heading (MeSH) search terms were used for all databases and a keyword search was used if the MeSH term was not available. (Please refer to the Electronic Supplementary Material for details). Utilities based on direct preference elicitation techniques such as the standard gamble, time trade-off and the Visual Analogue Scale (VAS) were not included in the search.

2.2 Study Selection

All potentially relevant abstracts were read to identify publications that could be eligible for inclusion in the review. Full-text articles of the selected abstracts were retrieved and selected based on the following inclusion/exclusion criteria:

- Type of publication: Only studies that were published in peer-reviewed journals were included. Conference proceedings and abstracts were excluded.
- Language: Only studies published in English or French were considered.
- Study design: All types of study designs were included.
- Study population: Studies that included persons diagnosed with possible or definite MS were included in the review without restrictions for disease severity, sex, type of MS or the presence of medical co-morbidities.
- Type of outcome measure: studies that reported on the psychometric properties of one or more of the following utility measures were included: the Quality of Well-Being (QWB) scale [18, 19], the Health Utilities Index Mark 2 (HUI2) [20, 21], the Health Utilities Index Mark 3 (HUI3) [21, 22], the 15D [23, 24], the EQ-5D/EQ-5D-5 Level (EQ-5D-5L) [1, 2], the Assessment of Quality of Life (AQOL) [25, 26] and the SF-6D [3]. The key characteristics of each of these measures are provided in Table 1.
- Psychometric properties: Studies that provided potentially relevant information on the psychometric property of a utility measure, whether this was their objective or not, were included in the review.

2.3 Data Extraction

The following information was extracted from each study: study characteristics (country, study design, and quality assessment of the study), subject characteristics (sample size, age and disease severity), outcome measures and results of psychometric tests.

2.4 Quality Assessment of Studies

The quality of the full-text articles included for review was assessed with a 13-item critical appraisal tool that was developed to assess psychometric properties of clinical measures [27]. Of the 13 items, four of the items were uniquely for articles assessing reliability, four were only for validity studies, and the remaining five items were for either one. The 13 items were scored as ‘yes’, ‘no’ or ‘not applicable’.

The Scottish Intercollegiate Guidelines Network Methodology (2013) was used to provide an overall summary of the level of evidence for each study: (i) two pluses ‘++’ were given when all or most of the quality criteria were fulfilled; (ii) one plus ‘+’ when some of the criteria were fulfilled; and (iii) a minus ‘−’ when few or none of the criteria were fulfilled. Therefore, ‘++’ indicated that the study was of high quality, ‘+’ indicated that it was of moderate quality, and ‘−’ that it was of low quality.

Methodological quality was assessed only for studies whose primary or secondary objectives were to evaluate the psychometric property of a utility measure. If a study’s objective was not to evaluate the psychometric property of a utility measure, its methodological quality was not assessed.

2.5 Psychometric Properties

The following psychometric properties were assessed from the included articles:

- *Content validity*: the extent to which the content of an instrument is an adequate reflection of the construct being measured. It evaluates whether all items included in a measure are relevant for the study population or disease [28].
- *Convergent validity*: considered a subtype of construct validity. It is the extent to which measures of constructs that theoretically should be related to each other are, in fact, observed to be related to each other [29].
- *Discriminative validity (known-groups validity)*: considered a subtype of construct validity. It is the degree to which an instrument can demonstrate different scores for groups known to vary or differ on the variables being measured [29].

Table 1 Summary of generic utility measures

Utility measure	Country of origin	Description of domains	Preferences obtained from	Method of eliciting preferences	No. of health states	Scoring algorithm	Scale range
QWB	USA	Mobility, physical activity, social activity plus 27 symptoms	Public	VAS	945	Regression/additive	0.00 to 1.00
HUI2	Canada	Sensation, mobility, emotion, cognitive, self-care, pain, fertility	Parents	VAS/SG	24,000	MAUF/multiplicative	−0.03 to 1.00
HUI3	Canada	Vision, hearing, speech, ambulation, dexterity, emotion, cognition, pain	Public	VAS/SG	972,000	MAUF/multiplicative	−0.36 to 1.00
15D	Finland	Mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental function, discomfort and symptoms, depression, distress, vitality, sexual activity	Public	VAS	31 billion	MAUF/additive	0.11 to 1.00
EQ-5D/EQ-5D-5L	UK	Mobility, self-care, usual activities, pain/discomfort, anxiety/depression	Public	TTO	243/3,125	Regression/additive	−0.59 to 1.00
AQOL	Australia	Self-care, household tasks, mobility, intimacy, friendships, family role, seeing, hearing, communication, sleep, anxiety and depression, pain	Public	TTO	16.8 million	MAUF/multiplicative	−0.04 to 1.00
SF-6D	UK	Physical functioning, role limitation, social functioning, pain, mental health, vitality	Public	SG	18,000	Regression/additive	0.46 to 1.00

AQOL Assessment of Quality of Life, *EQ-5D-5L* EQ-5D-5 Level, *MAUF* multi-attribute utility function, *HUI2* Health Utilities Index Mark 2, *HUI3* Health Utilities Index Mark 3, *No.* number, *QWB* Quality of Well-Being, *SG* standard gamble, *TTO* time trade off, *VAS* Visual Analogue Scale

- *Responsiveness*: the ability of a measure to detect change over time in the construct being measured [28].
- *Test–retest reliability*: the extent to which a measure provides the same results on repeated trials, assuming that the characteristics being measured do not change [29].
- *Floor/ceiling effect*: the percentage of the sample obtaining the worst and best possible scores. Values >15 % were indicative of a floor or ceiling effect [30].

2.6 Quantitative Analysis of Studies (Meta-Analysis)

The extent to which generic utility measures correlated with other measures of (i) impairment, (ii) activity limitations/participation restrictions, and (iii) health-related quality of life (HRQL) was examined to evaluate convergent validity. Forest plots were drawn to combine the correlation coefficient values. The Schmidt–Hunter method, which is a weighted mean of the correlation coefficient values, was used. This method is based on a random-effects model that weights each study by its sample size. Pooled correlation values of 0.1–0.3 were considered small, 0.4–0.6 were considered medium, and >0.7 were considered large [31]. Heterogeneity, the percentage of total variation across studies that is due to between-study differences rather than chance, was assessed using the I^2 statistic. The I^2 ranges between 0 and 100 %, with higher values indicating greater heterogeneity. A *p* value of <0.05 and an I^2 value >50 % indicated significant heterogeneity. All analysis was carried out using StatsDirect [32].

3 Results

3.1 Number of Articles Sourced

The study selection process is presented in Fig. 1. A total of 337 records were identified through the database searches. Ninety-two records were removed because they were duplicates, leaving 245 abstracts for screening. Of these, 230 articles were excluded because (i) they did not include a generic utility measure, (ii) they included a generic utility measure but did not provide information on its psychometric properties, (iii) study sample was not exclusive to MS, (iv) language was not English or French, and (v) they were conference proceedings or abstracts. This left 15 full-text articles for inclusion in the review.

One of the articles [33] included in this review (that also came up during the electronic database search) was published by the authors (AK and NM). This study reported data on the EQ-5D and the SF-6D (derived from the RAND-36) in 185 people with MS [33]. Although

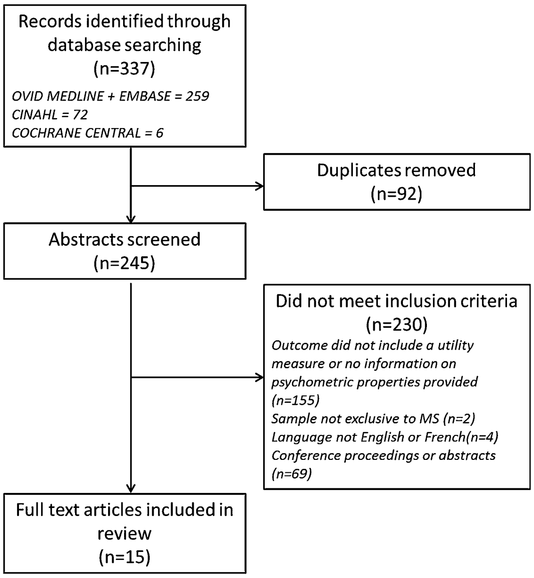

Records identified through database searching (n=337)
OVID MEDLINE + EMBASE = 259
CINAHL = 72
COCHRANE CENTRAL = 6
Duplicates removed (n=92)
Abstracts screened (n=245)
Did not meet inclusion criteria (n=230)
Outcome did not include a utility measure or no information on psychometric properties provided (n=155)
Sample not exclusive to MS (n=2)
Language not English or French(n=4)
Conference proceedings or abstracts (n=69)
Full text articles included in review (n=15)

Fig. 1 Flowchart of study selection process. *MS* multiple sclerosis

available, results on the convergent validity of these measures were not reported (as it was not the aim of that paper); therefore, these important data were incorporated into this review.

3.2 Brief Description of Included Studies for Each Utility Measure

The following generic utility measures were identified in the included articles: the EQ-5D ($n = 9$)/EQ-5D-5L ($n = 1$), followed by the HUI2/HUI3 ($n = 3$), the SF-6D ($n = 2$), the AQOL ($n = 2$), and last the QWB scale ($n = 1$). There were no studies that reported on the psychometric property of the 15D. Table 2 presents key characteristics for each study, and Supplementary Table 1 presents a breakdown of the methodological quality assessment (see the Electronic Supplementary Material).

EQ-5D/EQ-5D-5L: There were nine studies [13, 33–40] that assessed the psychometric properties of the EQ-5D, and one study [41] that assessed the EQ-5D-5L (total = 10 studies). The studies were cross-sectional in design, with sample sizes ranging from 18 to 911 and mean utility scores ranging from 0.49 to 0.80. The studies were of moderate to high quality.

HUI2: Only one study [42] provided information on the psychometric property of the HUI2. The study was cross-sectional in design and consisted of 153 patients with MS who were recruited from two different MS clinics. The study was of moderate methodological quality.

HUI3: There were two studies [13, 43] that provided information on the psychometric properties of the HUI3. Both studies were cross-sectional, with sample sizes of 187 and 302. The mean utility score was presented in only one study, and was 0.57 with a 95 % confidence interval (CI) of 0.52–0.63. Methodological quality was assessed for one of the studies [13] and was graded as high quality. The remaining study [43] was not assessed for methodological quality because its primary objective was not to test psychometric property of the HUI3.

SF-6D: Two studies [13, 33] reported on the psychometric properties of the SF-6D. Both studies were cross-sectional in design and had similar sample sizes (187 and 185). The mean utility value was reported by one of the studies, and was 0.69 standard deviation (SD) 0.13. The studies were of moderate to high quality.

AQOL: The AQOL was evaluated in two studies [44, 45], both of which were conducted by the same author. The first study [44] consisted of a community-based MS group ($n = 101$) in Australia with a mean utility score of 0.46 (SD 0.25) on the AQOL. The second study [45] included a sample of 61 MS patients suffering from chronic pain with mean utility scores ranging from 0.24 to 0.37.

QWB scale: The psychometric property of the QWB scale was reported in only one study [46], which involved 274 patients with MS. The study was cross-sectional in design and did not report the mean utility value for the sample. The methodological quality of the study was not assessed, as its primary objective was not to evaluate the psychometric property of the QWB scale.

3.3 Psychometric Properties of Identified Utility Measures

3.3.1 EQ-5D/EQ-5D-5L

3.3.1.1 Content Validity The content validity of the EQ-5D was evaluated in one study [33] on a sample of 185 people with MS. The objective of this study was to estimate the extent to which the EQ-5D captured domains that were relevant to patients with MS. Certain domains such as walking (mobility) and mood (anxiety/depression) which were identified by patients to be important to their quality of life were included in the EQ-5D. However, other important domains such as fatigue and cognition were not included in the utility measure.

3.3.1.2 Convergent Validity Impairment: Figure 2 is a forest plot for convergent validity of the EQ-5D tested against outcome measures of impairment, such as gait speed and disease severity. The pooled correlation coefficient for convergent validity of the EQ-5D was 0.35 (95 % CI 0.25–0.45). The I^2 statistic for heterogeneity was high at 94.6 % ($p < 0.0001$).

Activity limitations/participation restrictions: Supplementary Fig. 1 presents the correlation coefficient values for convergent validity of the EQ-5D against outcome

Table 2 Description of included studies stratified by type of generic utility measure

Author (year)	Country	Study design	Study setting	Participants	Mean ± SD for utility measure	Psychometric property assessed	Methodological quality
EQ-5D/EQ-5D-5L (n = 10)							
Fogarty et al. (2013) [41]	Ireland	Cross-sectional	Outpatient clinic	$N = 214$, age 47.8 ± 12.7, EDSS 3.6 ± 2.6	0.59 ± 0.33	Known-groups validity; ceiling effect	Moderate quality
Kuspinar and Mayo (2013) [33]	Canada	Cross-sectional	3 outpatient clinics	$N = 185$, age 42.8 ± 10.0, EDSS 2.0 (IQR 1.0–3.5)	0.69 ± 0.18	Content validity; convergent validity (not in paper, but provided by authors)	High quality
Kikuchi et al. (2011) [34]	Japan	Cross-sectional	Inpatient and outpatient settings (8 centers)	$N = 163$, age 42.8 ± 12.3, EDSS 4.0 ± 2.5	Not presented	Convergent validity	Not assessed—primary objective was not to assess psychometric property of utility measure
Twiss et al. (2010) [35]	UK	Longitudinal (but correlations reported for baseline only)	172 centers in multiple countries	$N = 911$, age 36.2 ± 8.4, EDSS 0–4+	Baseline 0.80 ± 0.19, 12 months 0.80 ± 0.21	Convergent validity	Not assessed—primary objective was not to assess psychometric property of utility measure
Ploughman et al. (2010) [36]	Canada	Cross-sectional	Older people with MS	$N = 18$, age 66.5 ± 6.7, EDSS not presented	Not presented	Known-groups validity	Qualitative study—difficult to assess with quality assessment tool
Orme et al. (2007) [37]	UK	Cross-sectional	Community dwelling	$N = 2{,}048$, age 51.4, EDSS 0–9.5 (range)	0.49 ± 0.32	Convergent validity; known-groups validity	Not assessed—primary objective was not to assess psychometric property of utility measure
Fisk et al. (2005) [13]	Canada	Cross-sectional	2 outpatient clinics	$N = 187$, age 51 ± 10, EDSS 6 ± 4	Mean ± SD not presented	Convergent validity; known-groups validity; test–retest reliability; floor/ceiling effects	High quality
Moore et al. (2004) [38]	Canada	Cross-sectional	Outpatient clinic	$N = 114$, age 45 ± 11, EDSS 0–6+	0.61 ± 0.28	Convergent validity	High quality
Nicholl et al. (2001) [39]	UK	Cross-sectional	Rehabilitation center or community dwelling	$N = 96$, age 49.0 ± 8.9, EDSS not presented	Not presented	Convergent validity; known-groups validity	High quality
Rothwell et al. (1997) [40]	UK	Cross-sectional	Rehabilitation center or outpatient clinic	$N = 42$, age 41 (range 28–68), EDSS 5.5 (range 1–8)	Not presented	Convergent validity	Moderate quality

Table 2 continued

Author (year)	Country	Study design	Study setting	Participants	Mean ± SD for utility measure	Psychometric property assessed	Methodological quality
HUI3 (n = 2)							
Jones et al. (2008) [43]	Canada	Cross-sectional	Community-dwelling patients and healthy population	*N* = 302 (MS), age 48.7 (95 % CI 46.6–50.8), EDSS not presented; *N* = 109,741 (healthy), age 44.8 (95 % CI 44.7–44.8)	0.57 (95 % CI 0.52–0.63)	Content validity	Not assessed—primary objective was not to assess psychometric property of utility measure
Fisk et al. (2005) [13]	Canada	Cross-sectional	2 outpatient clinics	*N* = 187, age 51 ± 10, EDSS 6 ± 4	Mean ± SD not presented	Convergent validity; known-groups validity; test–retest reliability; floor/ceiling effects	High quality
HUI2 (n = 1)							
Grima et al. (2000) [42]	Canada	Cross-sectional	2 outpatient clinics	*N* = 153, age 41 ± 15, EDSS 1–6	Mean ± SD not presented	Convergent validity; known-groups validity	Moderate quality
SF-6D (n = 2)							
Fisk et al. (2005) [13]	Canada	Cross-sectional	2 outpatient clinics	*N* = 187, age 51 ± 10, EDSS 6 ± 4	Mean ± SD not presented	Convergent validity; known-groups validity; test–retest reliability; floor/ceiling effects	High quality
Kuspinar and Mayo (2013) [33]	Canada	Cross-sectional	3 outpatient clinics	*N* = 185, age 42.8 ± 10.0, EDSS 2.0 (IQR 1.0–3.5)	0.69 ± 0.13	Content validity; convergent validity (not in paper, but provided by authors)	Moderate quality
AQOL (n = 2)							
Khan et al. (2006) [44]	Australia	Cross-sectional	Community dwelling	*N* = 101, age 49.5 ± 9.2, EDSS 4.9 ± 1.5	0.46 ± 0.25	Known-groups validity	Not assessed—primary objective not to assess psychometric property of utility measure
Khan and Pallant (2007) [45]	Australia	Cross-sectional	Community-dwelling patients with chronic pain	*N* = 61, mean age range 25.6–35.9, EDSS 0–8	Mean utility score range 0.24–0.37	Known-groups validity	Not assessed—primary objective not to assess psychometric property of utility measure
QWB (n = 1)							
Schwartz et al. (1999) [46]	USA	Cross-sectional	13 hospital and clinical sites	*N* = 274, age 46 ± 12, EDSS median 5 (range 0–8.5)	Not presented for total sample	Convergent validity; known-groups validity	Not assessed—Primary objective was not to assess psychometric property of utility measure

AQOL Assessment of Quality of Life, *CI* confidence interval, *EDSS* Expanded Disability Status Scale, *EQ-5D-5L* EQ-5D-5 Level, *HUI2* Health Utilities Index Mark 2, *HUI3* Health Utilities Index Mark 3, *MS* multiple sclerosis, *n* number, *QWB* Quality of Well-Being, *SD* standard deviation

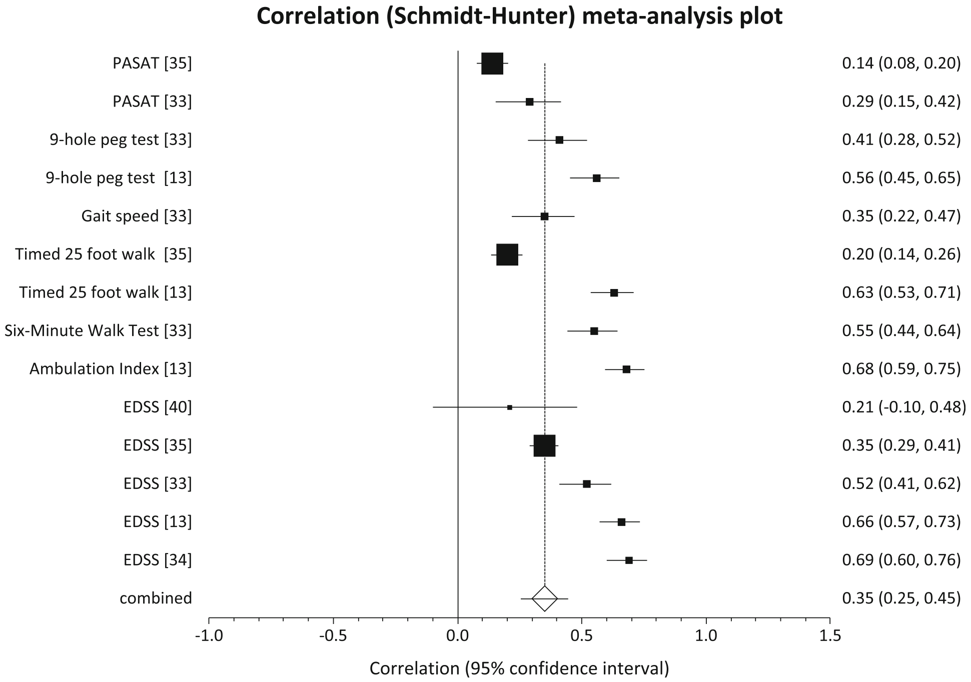


Fig. 2 Forest plot with correlation coefficients (r) of the EQ-5D against outcome measures evaluating impairments of body structure and function. *PASAT* Paced Auditory Serial Addition Test, *EDSS* Expanded Disability Status Scale

measures of activity limitations and participation restrictions (e.g., social function). The pooled correlation was 0.51 (95 % CI 0.45–0.57) and the I^2 statistic for heterogeneity was high at 81.8 % ($p < 0.0001$).

HRQL: Figure 3 presents the combined correlation value for the EQ-5D compared against measures evaluating HRQL, which was 0.56 (95 % CI 0.54–0.59). There was no heterogeneity among the included studies (I^2 statistic = 0 %, $p = 0.53$).

3.3.1.3 Discriminative/Known-Groups Validity Discriminant validity of the EQ-5D was evaluated in three studies [36, 37, 39]. Two of these studies [36, 39] reported that the mobility item lacked discriminative ability because patients who were wheelchair bound did not fit into any response category.

Orme et al. [37] evaluated the extent to which the EQ-5D was able to differentiate between different levels of disease severity. Disease severity was measured using the Expanded Disability Status Scale (EDSS), a classification scheme extending from 0 (normal neurological examination) to 10 (death due to MS). The authors reported that the EQ-5D was able to differentiate between all EDSS levels, except between EDSS levels 3 and 4 (utility score for EDSS 4 was higher than EDSS 3). Fisk et al. [13] found that the decline in utility scores between the mildly (EDSS 0–2.5) and moderately (EDSS 3.0–5.5) impaired MS patients was not statistically significant ($p = 0.30$).

Only one study [41] evaluated the discriminative capacity of the EQ-5D-5L, which showed a linear decline in utility scores from EDSS 0–6, after which point the relationship exhibited greater variability. Furthermore, the discriminative power of the EQ-5D-5L was considerably lower for the domains of self-care and anxiety/depression, compared with the other domains (mobility, pain and usual activities).

3.3.1.4 Test–Retest Reliability The intra-class correlation coefficient for test–retest reliability of the EQ-5D was 0.81 [13].

3.3.1.5 Floor/Ceiling Effect For the EQ-5D, ceiling effects were reported for the mobility item (32 %) and the self-care item (68 %) [13]. No floor effects were found for any of the EQ-5D items. As for the EQ-5D-5L [41], ceiling effects were reported for the self-care item (64 %) and the anxiety/depression item (46 %).

3.3.2 HUI2

3.3.2.1 Content Validity One study [33] evaluated the content validity of the HUI2. The authors identified that the utility measure included domains relevant to patients with MS, such as cognition. However, the authors also identified that the HUI2 was missing certain domains such as fatigue and work.

3.3.2.2 Convergent Validity Impairment: One study [42] calculated the correlation between the EDSS and the HUI2 to be 0.54 ($p < 0.0001$).

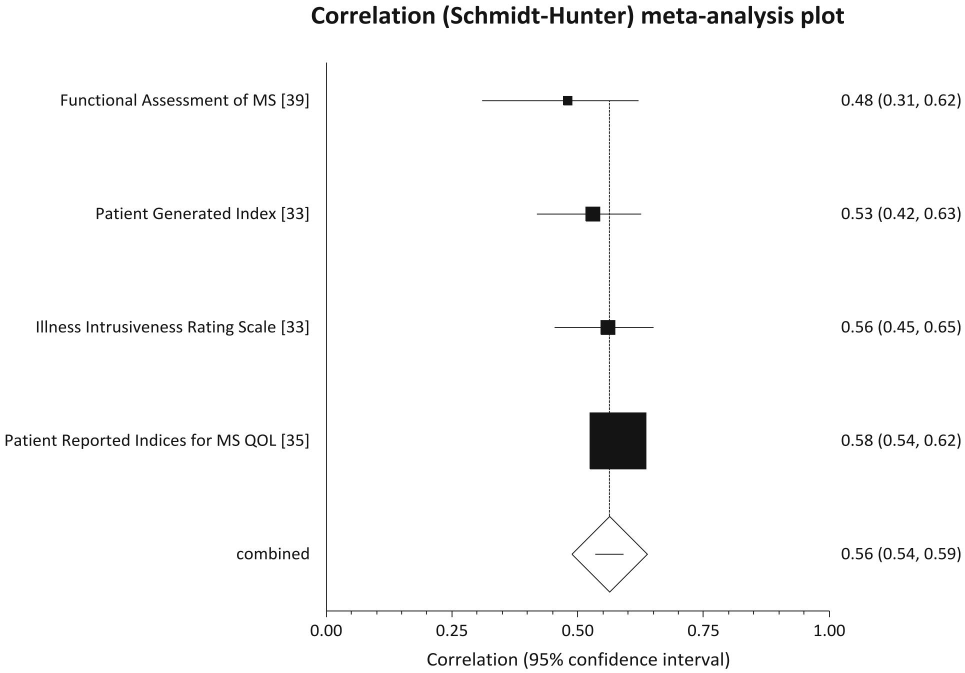


Fig. 3 Forest plot with correlation coefficients (*r*) of the EQ-5D against outcomes of health-related quality of life. MS multiple sclerosis, *Patient Reported Indices for MS QOL* Patient Reported Indices for Multiple Sclerosis Quality of Life subscale

Activity limitations/participation restrictions and HRQL: No studies were available.

3.3.2.3 Discriminative/Known-Groups Validity Mean HUI2 utility scores were 0.83, 0.84, 0.71, 0.71, 0.62 and 0.59 for EDSS levels 1–6, respectively [42].

3.3.2.4 Floor/Ceiling Effect There were no studies that reported on the presence or absence of floor/ceiling effects in the HUI2.

3.3.2.5 Test–Retest Reliability There were no studies that reported about test–retest reliability of the HUI2.

3.3.3 HUI3

3.3.3.1 Content Validity Two studies [33, 43] provided information on the content validity of the HUI3. In the first study, the authors identified that important domains such as fatigue were missing in the HUI3. Furthermore, the HUI3 included domains that were not relevant to many patients with MS, such as self-care, vision and hearing. This may not only affect the measure's ability to detect meaningful change, but may also result in an overestimation of utility scores and false estimates of QALYs. For the second study [43], clinically important differences in scores between patients with MS and the general population were observed for ambulation, pain, dexterity and cognition. However, differences were not observed for hearing and speech, suggesting that these domains or items may not be impacted in MS.

3.3.3.2 Convergent Validity Impairment: When the convergent validity of the HUI3 was tested against outcome measures of impairments, the pooled correlation value was 0.73 (95 % CI 0.68–0.77). The I^2 statistic for heterogeneity was 55 % ($p = 0.082$) (Fig. 4).

Activity limitations/participation restrictions and HRQL: There were no studies that assessed the convergent validity of the HUI3 against measures of activity limitation and participation restrictions, or HRQL.

3.3.3.3 Discriminative/Known-Groups Validity The HUI3 demonstrated known-groups validity by being able to differentiate between mildly, moderately and severely disabled MS patients [13].

3.3.3.4 Test–Retest Reliability The intra-class correlation coefficient for test–retest reliability of the HUI3 was 0.87 [13].

3.3.3.5 Floor/Ceiling Effect There were no ceiling or floor effects for the HUI3 [13]. Only 3 % of subjects obtained a utility score of 1.0 and 10 % of subjects obtained a utility score of <0.

3.3.4 SF-6D

3.3.4.1 Content Validity Only one study [33] reported on the content validity of the SF-6D in MS. The SF-6D was found to include several domains that were important to the quality of life of patients with MS, such as work, fatigue, sports (vigorous physical activities) and social life.

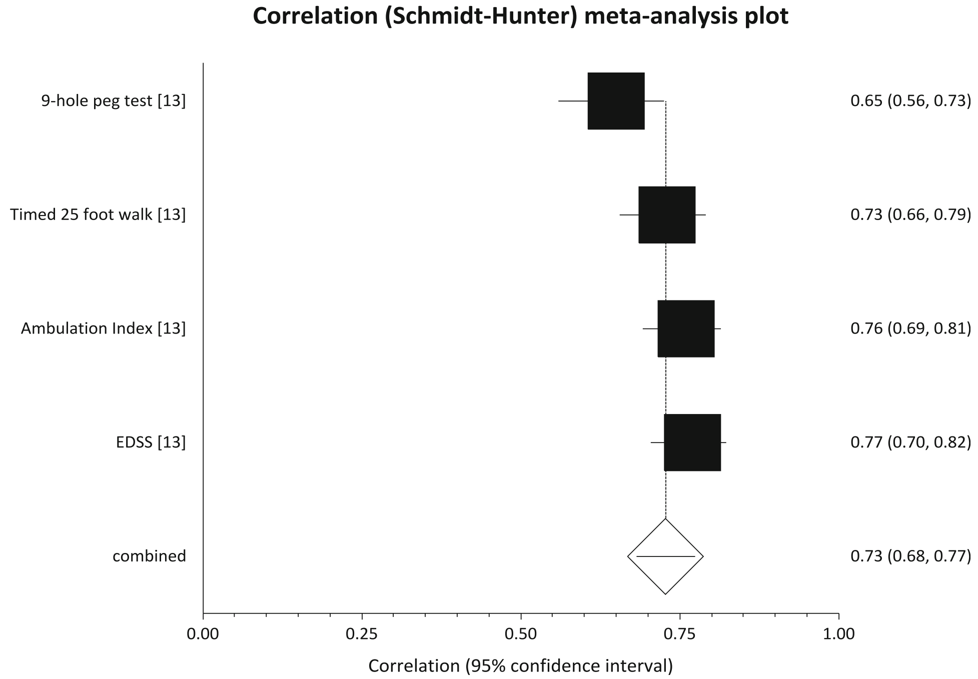


Fig. 4 Forest plot with correlation coefficients (*r*) of the HUI3 against outcomes of impairments of body structure and function. *EDSS* Expanded Disability Status Scale

However, it was missing important domains such as walking and cognition.

3.3.4.2 Convergent Validity Impairment: The pooled correlation value for convergent validity of the SF-6D against outcome measures evaluating impairments of body structure and function (Fig. 5) was 0.39 (95 % CI 0.32–0.46). The I^2 statistic was 66 % ($p = 0.003$).

Activity limitations/participation restrictions: The combined correlation value for the SF-6D against measures evaluating activity limitations and participation restrictions was 0.57 (95 % 0.54–0.59) with an I^2 statistics of 0 % ($p = 0.67$) (Supplementary Fig. 2).

HRQL: When compared against measures evaluating HRQL, the pooled correlation value for convergent validity of the SF-6D was 0.62 (95 % CI 0.50–0.73). The I^2 statistic for heterogeneity was 86 % ($p = 0.008$) (Fig. 6).

3.3.4.3 Discriminative/Known-Groups Validity One study [13] evaluated the discriminative ability of the SF-6D and found that although the index was able to differentiate between mildly and moderately disabled patients, it was unable to differentiate between the more severe patient groups. A flattening of utility scores beyond moderate disability was observed.

3.3.4.4 Test–Retest Reliability The intra-class correlation coefficient for test–retest reliability of the SF-6D was 0.83 [13].

3.3.4.5 Floor/Ceiling Effect For the SF-6D, only 3 and 1 % of subjects reported the lowest and highest possible index scores respectively. However, floor effects were identified for the physical function subscale (41 %) and the role limitation subscale (16 %). Ceiling effects were reported for bodily pain (29 %), social function (39 %), mental health (58 %) and role limitations (84 %) [13].

3.3.5 AQOL

3.3.5.1 Content Validity There were no studies that reported on the content validity of the AQOL in MS.

3.3.5.2 Convergent Validity There were no studies that reported on the convergent validity of the AQOL in MS.

3.3.5.3 Discriminative/Known-Groups Validity The AQOL was able to differentiate between mildly, moderately and severely disabled patients [44], and it was also able to differentiate between patients with different levels of pain intensity [45].

3.3.5.4 Test–Retest Reliability There were no studies that reported on the test–retest reliability of the AQOL.

3.3.5.5 Floor/Ceiling Effect There were no studies that reported on floor or ceiling effects.

3.3.6 QWB Scale

3.3.6.1 Content Validity There were no studies identified that reported on the content validity of the QWB scale in MS.

Fig. 5 Forest plot with correlation coefficients (*r*) of the SF-6D against outcomes of impairments of body structure and function. *PASAT* Paced Auditory Serial Addition Test, *EDSS* Expanded Disability Status Scale

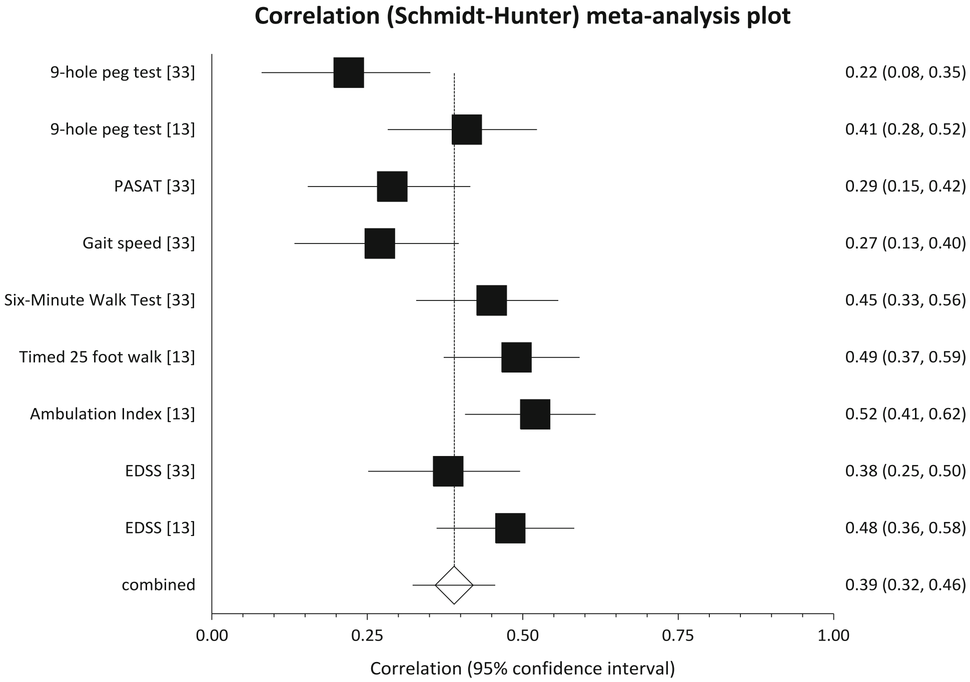
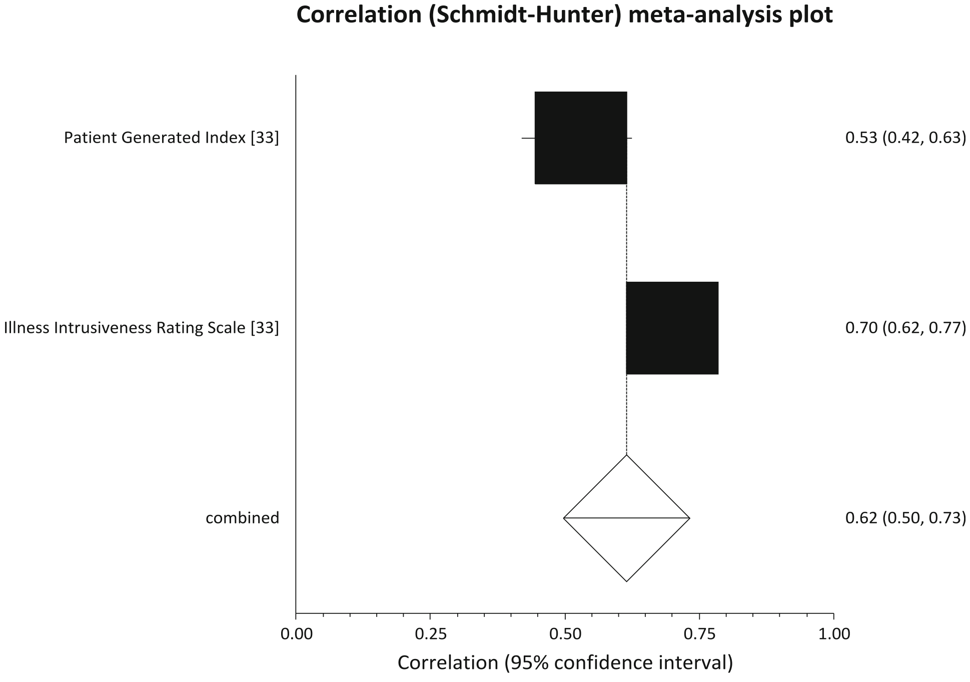


Fig. 6 Forest plot with correlation coefficients (*r*) of the SF-6D against outcomes of health-related quality of life



3.3.6.2 Convergent Validity Impairment: Supplementary Fig. 3 presents the pooled value for convergent validity of the QWB scale against measures evaluating impairment of body structure and function. The combined correlation value was 0.36 (95 % CI 0.24–0.49), and the I^2 statistic was high at 89.1 % ($p < 0.0001$).

Activity limitations/participation restrictions: Supplementary Fig. 4 is a forest plot of the combined correlation coefficient values for convergent validity of the QWB scale when compared against measures of activity limitation and participation restriction. The combined correlation was 0.55 (95 % CI 0.43–0.67), and the I^2 statistic for heterogeneity was high at 87.7 % ($p = 0.004$).

HRQL: There were no studies that evaluated the convergent validity of the QWB scale against measures of quality of life.

3.3.6.3 Discriminative/Known-Groups Validity The QWB scale was able to discriminate between mild and moderate levels of disability, but was not able to differentiate between moderate and severe [46].

3.3.6.4 Test–Retest Reliability There were no studies that reported on the test–retest reliability of the QWB scale.

3.3.6.5 Floor/Ceiling Effect There were no studies that reported on floor or ceiling effects.

4 Discussion

This structured review summarizing the published literature on the reliability and validity of generic utility measures in MS showed that each of the utility measures had their strengths and weaknesses. In terms of content validity, cognition, a domain important to MS, was missing in both the EQ-5D and the SF-6D. Fatigue, another important domain, was missing in the HUI and the EQ-5D. The content validity of the QWB scale and the AQOL were not assessed in any of the included studies. However, if one were to quickly review the items included in these measures, fatigue is missing in the AQOL and the QWB scale, and cognition is missing in the AQOL.

Ceiling and floor effects were present for the EQ-5D and the SF-6D, but not for the HUI3. As for test–retest reliability, the EQ-5D, the SF-6D and the HUI3 all demonstrated excellent reliability. Ceiling/floor effects and test–retest reliability were not assessed for the AQOL or the QWB scale.

In terms of discriminative ability, the SF-6D and the QWB scale were not able to differentiate between moderately and severely disabled MS patients, and the EQ-5D was not able to differentiate between those who were mildly and moderately disabled. Issues were also identified with the mobility item of the EQ-5D, because patients who were wheelchair bound did not fit into any of the response categories. The AQOL and the HUI3 demonstrated good discriminative ability, as both measures were able to differentiate between all levels of disability.

As for convergent validity, the HUI3 was highly correlated ($r = 0.7$) against measurement instruments that evaluated impairments such as disease severity, ambulation and manual dexterity. This is probably not surprising, as the HUI3 was developed with the intention of including only impairment-related domains, and excluding ‘out of skin’ domains such as participation in life roles (i.e., work) [20, 21]. Impairments can impact on participation, but this association is often surprisingly weak in people with disabling health conditions as people learn to create a life even with impairments [47, 48]. In the context of MS, it is relevant to know both the level of impairment and the level to which it restricts participation [47].

The correlations for convergent validity were very similar between the EQ-5D and the SF-6D. Both measures had small to moderate correlations ($r = 0.4$) against instruments evaluating impairments, and slightly stronger correlations against measures of activity limitations/participation restrictions ($r = 0.6$). There is considerable overlap between the EQ-5D and the SF-6D in terms of item or domain coverage. For example, both the EQ-5D and the SF-6D include an item on pain. Self-care in the EQ-5D is covered as bathing and dressing in the SF-6D. Furthermore, the equivalent of the anxiety/depression item in the EQ-5D is feeling tense and downhearted in the SF-6D.

The QWB scale behaved similarly to the EQ-5D and the SF-6D, also demonstrating small to moderate correlations ($r = 0.36$) with measures of impairment and activity limitations/participation restrictions ($r = 0.55$). The QWB scale contains items that are similar to the EQ-5D and the SF-6D (mobility, physical activity, social activity, plus 27 symptoms). Although the QWB scale was the first utility measure to be developed, it is used to a lesser extent than the other utility measures. This may be because it requires substantial training of interviewers and detailed probing of the patient [49]. A more recent self-administered version of the QWB scale has been developed [50]; however, it still takes about 14 min to complete [51]. The EQ-5D and the SF-6D, on the other hand, require only 5 min or less to complete.

To our knowledge this is the first study that reviewed the validity and reliability of generic utility measures in MS. Structured reviews similar to ours have been conducted for other health conditions, such as urinary incontinence [52], spinal cord injury [53], visual disorders [54], schizophrenia [11], diabetes [5] and cardiovascular disease [4]. The results of these studies were mixed, where some reviews found evidence that supported the use of generic utility measures for the health condition under study [4, 5, 52], while others were not able to make such conclusions [11, 53].

There were limitations in the included studies that need to be acknowledged. First, several of the included studies were not specifically designed to test the psychometric properties of utility measures; they provided data that were potentially relevant for this review. Second, the high levels of heterogeneity among the included studies indicate that the pooled correlation coefficients for convergent validity should be interpreted with caution. Third, a full assessment of the psychometric property of the AQOL or the QWB scale was not possible, as we were not able to find information on test–retest reliability and presence of floor or ceiling effects. Fourth, our findings showed that the psychometric property of the 15D in MS has not yet been evaluated; therefore, an analysis of the appropriateness of this utility measure in MS could not be made. Fifth, there were no studies that assessed the responsiveness of these utility measures, making it difficult to draw any conclusions on the ability of these measures to detect clinically important change.

The generic utility measures identified in this review were able to explain only 36 % ($r = 0.6$) of the variance in generic and disease-specific health profiles such as the Patient Generated Index (PGI) and the Patient-Reported Indices for MS (PRIMUS). A large of proportion of the variance (64 %) remained unexplained in these measures, which raises the question of whether generic utility measures are indeed providing an adequate representation of patients' HRQL. Although items that are commonly included in generic measures are also of importance to people with MS, generic utility measures may miss certain domains that are important or specific to the disease. The addition of disease specific 'bolt-ons' or 'dimension extensions' to generic utility measures is one possible method to improve the validity of these measures in MS. A recent review by Lin et al. [55] identified several domains that were specific to different diseases and that could be used as 'bolt-ons' to the EQ-5D. Potential domains that could be included as 'bolt-ons' to generic utility measures are cognition (not found in the EQ-5D or SF-6D) and fatigue (not included in the EQ-5D or HUI2/3). With the bolt-on approach, the wording or phrasing of the bolt-on item and its response options first needs to be developed. Following this, a valuation exercise with the bolt-on item is carried out and a multi-attribute utility function or scoring algorithm calculated. The challenge with the bolt-on approach is that the addition of a new domain may have an impact on the way people value the original dimensions, altering the original regression coefficient values.

Another possible solution to tackle the limitations found with generic utility measures is to develop a disease-specific utility measure for MS. Such a measure would include only domains that are relevant to people with MS and, therefore, provide an accurate assessment of the clinical and cost effectiveness of different treatment options in this population. One of the concerns with disease-specific measures is that they may not be able to capture the impact of co-morbid medical conditions on HRQL. However, in the context of MS, the age of diagnosis is approximately 20–40 years, when co-morbidities are rare. As the context of use for a condition-specific measure is around medication that is usually prescribed around time of diagnosis, most patients will not have co-morbidities. These develop late on with aging, as in any group of people.

As each disease-specific measure will have a different classification system, a concern is whether this will affect comparison of treatments across diseases. However, the issue of comparability is not just limited to the context of disease-specific utility measures but also applies to generic utility measures. As pointed out in this review, there are considerable differences in content coverage (i.e., domains) and methods of valuation (i.e., standard gamble vs. time trade-off vs. VAS) among the generic measures. Furthermore, studies have shown that there are significant discrepancies in utility scores obtained using the EQ-5D, HUI3 and the SF-6D for the same medical condition [9]. For this reason, in the UK, the National Institute of Health and Care Excellence (NICE) advocates for the use of one descriptive system, namely the EQ-5D, for economic evaluation purposes. However, the limitation with this approach is that one measure may not be appropriate for all health conditions. In Canada, the Canadian Agency for Drugs and Technologies in Health (CADTH) does not have a preference for any one utility measure. Provided that the utility measure is reliable and demonstrates validity in the population of interest, it may be used for economic evaluation purposes.

5 Conclusion

The HUI3 demonstrated the strongest psychometric properties when compared with other utility measures. However, the HUI3 only measured impairment and excluded important components of HRQL, such as activity limitations and participation restrictions. The EQ-5D, the SF-6D and the QWB scale, on the other hand, did include items on participation in life roles. However, these measures demonstrated a lack of content validity in MS by missing certain domains that were important to the disease, as well as difficulty in differentiating between different levels of disability. The addition of MS-specific 'bolt-ons' to generic utility measures and the development of an MS specific utility measure are possible areas of exploration for future research.

Conflicts of interest The authors have no conflicts of interest to declare.

Authors' contribution AK was primarily responsible for writing the manuscript in close cooperation with NM. Both authors read, edited, and approved the final manuscript. AK is the overall guarantor for the content.

References

1. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095–108.
2. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005;43:203–20.

3. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21:271–92.
4. Dyer MT, Goldsmith KA, Sharples LS, Buxton MJ. A review of health utilities using the EQ-5D in studies of cardiovascular disease. Health Qual Life Outcomes. 2010;8:13.
5. Janssen MF, Lubetkin EI, Sekhobo JP, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with type 2 diabetes mellitus. Diabet Med. 2011;28:395–413.
6. Richardson J, Iezzi A, Khan MA, Maxwell A. Validity and reliability of the Assessment of Quality of Life (AQoL)-8D multi-attribute utility instrument. Patient Patient Cent Outcomes Res 2013;1–12.
7. Tran BX, Ohinmaa A, Nguyen LT. Quality of life profile and psychometric properties of the EQ-5D-5L in HIV/AIDS patients. Health Qual Life Outcomes. 2012;10:132.
8. Garau M, Shah KK, Mason AR, Wang Q, Towse A, Drummond MF. Using QALYs in cancer. Pharmacoeconomics. 2011;29:673–85.
9. Espallargues M, Czoski-Murray CJ, Bansback NJ, Carlton J, Lewis GM, Hughes LA, Brand CS, Brazier JE. The impact of age-related macular degeneration on health status utility values. Invest Ophthalmol Vis Sci. 2005;46:4016–23.
10. Barton GR, Bankart J, Davis AC, Summerfield QA. Comparing utility scores before and after hearing-aid provision: results according to the EQ-5D, HUI3 and SF-6D. Appl Health Econ Health Policy. 2004;3:103–5.
11. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. Value Health. 2011;14:907–20.
12. Myers JA, McPherson KM, Taylor WJ, Weatherall M, McNaughton HK. Duration of condition is unrelated to health-state valuation on the EuroQoL. Clin Rehabil. 2003;17:209–15.
13. Fisk JD, Brown MG, Sketris IS, Metz LM, Murray TJ, Stadnyk KJ. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. J Neurol Neurosurg Psychiatry. 2005;76:58–63.
14. Feeny D. Preference-based measures: utility and quality-adjusted life years. In: Fayers PM, Hays RD, editors. Assessing quality of life in clinical trials. 2nd ed. Oxford University Press; 2005. p. 405–29.
15. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2004;13:873–84.
16. Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG. Multiple sclerosis. N Engl J Med. 2000;343:938–52.
17. Terwee CB, Mokkink LB. Measurement in medicine: a practical guide. London: Cambridge University Press; 2011.
18. Kaplan RM, Anderson JP, Wu AW, Mathews WC, Kozin F, Orenstein D. The Quality of Well-Being Scale. Applications in AIDS, cystic fibrosis, and arthritis. Med Care. 1989;27:S27–43.
19. Kaplan RM, Ganiats TG, Sieber WJ, Anderson JP. The Quality of Well-Being Scale: critical similarities and differences with SF-36. Int J Qual Health Care. 1998;10:509–20.
20. Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. J Clin Oncol. 1992;10:923–8.
21. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. Pharmacoeconomics. 1995;7:490–502.
22. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, Denton M, Boyle M. Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. Med Care. 2002;40:113–28.
23. Sintonen H, Pekurinen M. A fifteen-dimensional measure of health-related quality of life (15D) and its applications. In: Walker SR, Rosser R, editors. Quality of life assessment: key issues in the 1990s. Springer, Netherlands; 1993. p. 185–95.
24. Sintonen H. The 15D instrument of health-related quality of life: properties and applications. Ann Med. 2001;33:328–36.
25. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. Qual Life Res. 1999;8:209–24.
26. Richardson J, Atherton Day N, Peacock S, Iezzi A. Measurement of the quality of life for economic evaluation and the Assessment of Quality of Life (AQoL) Mark 2 instrument. Aust Econ Rev. 2004;37:62–88.
27. Brink Y, Louw QA. Clinical instruments: reliability and validity critical appraisal. J Eval Clin Pract. 2012;18:1126–32.
28. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63:737–45.
29. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. New York: Oxford university press; 2008.
30. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4:293–307.
31. Polgar S, Thomas SA. Introduction to research in the health sciences. Elsevier Health Sciences; 2013.
32. StatsDirect L. StatsDirect statistical software. StatsDirect, UK; 2005.
33. Kuspinar A, Mayo NE. Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis? Health Qual Life Outcomes. 2013;11:71.
34. Kikuchi H, Mifune N, Niino M, Ohbu S, Kira J, Kohriyama T, Ota K, Tanaka M, Ochi H, Nakane S, Maezawa M, Kikuchi S. Impact and characteristics of quality of life in Japanese patients with multiple sclerosis. Qual Life Res. 2011;20:119–31.
35. Twiss J, Doward LC, McKenna SP, Eckert B. Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS). Health Qual Life Outcomes. 2010;8:117.
36. Ploughman M, Austin M, Stefanelli M, Godwin M. Applying cognitive debriefing to pre-test patient-reported outcomes in older people with multiple sclerosis. Qual Life Res. 2010;19:483–7.
37. Orme M, Kerrigan J, Tyas D, Russell N, Nixon R. The effect of disease, functional status, and relapses on the utility of people with multiple sclerosis in the UK. Value Health. 2007;10:54–60.
38. Moore F, Wolfson C, Alexandrov L, Lapierre Y. Do general and multiple sclerosis-specific quality of life instruments differ? Can J Neurol Sci. 2004;31:64–71.
39. Nicholl CR, Lincoln NB, Francis VM, Stephan TF. Assessing quality of life in people with multiple sclerosis. Disabil Rehabil. 2001;23:597–603.
40. Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ. 1997;314:1580–3.
41. Fogarty E, Walsh C, Adams R, McGuigan C, Barry M, Tubridy N. Relating health-related Quality of Life to disability progression in multiple sclerosis, using the 5-level EQ-5D. Mult Scler. 2013;19:1190–6.
42. Grima DT, Torrance GW, Francis G, Rice G, Rosner AJ, Lafortune L. Cost and health related quality of life consequences of multiple sclerosis. Mult Scler. 2000;6:91–8.
43. Jones CA, Pohar SL, Warren S, Turpin KV, Warren KG. The burden of multiple sclerosis: a community health survey. Health Qual Life Outcomes. 2008;6:1–7.

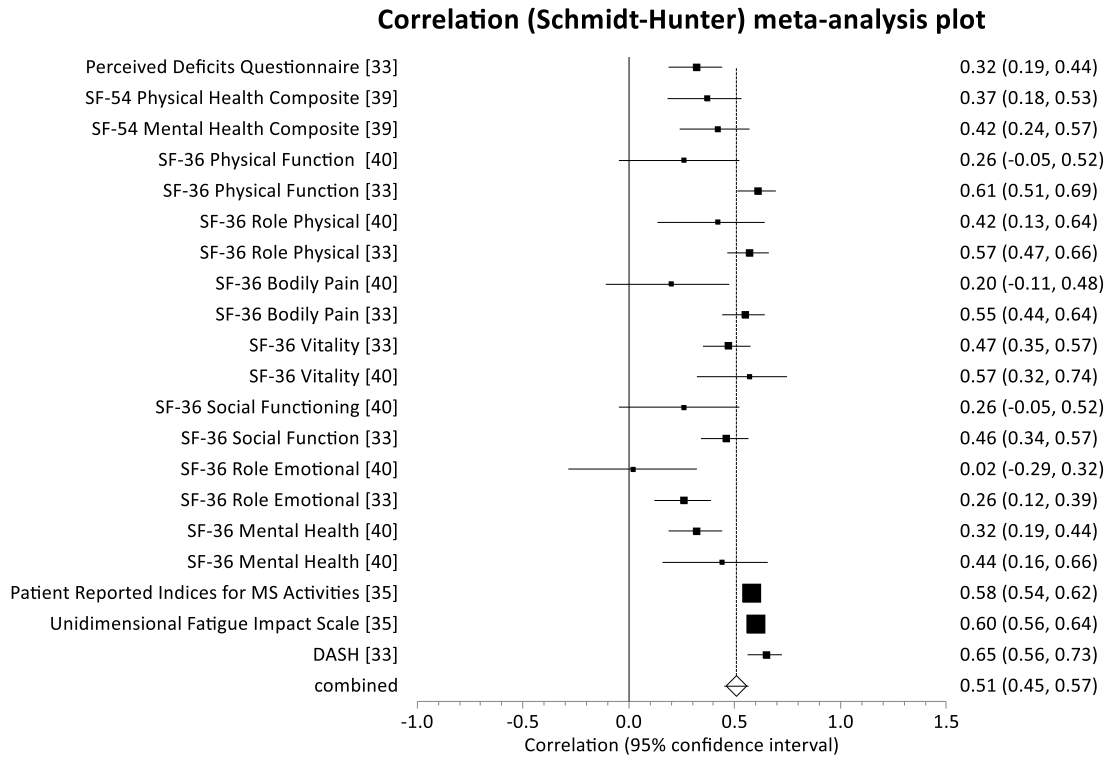
44. Khan F, McPhail T, Brand C, Turner-Stokes L, Kilpatrick T. Multiple sclerosis: disability profile and quality of life in an Australian community cohort. Int J Rehabil Res. 2006;29:87–96.
45. Khan F, Pallant J. Chronic pain in multiple sclerosis: prevalence, characteristics, and impact on quality of life in an Australian community cohort. J Pain. 2007;8:614–23.
46. Schwartz CE, Vollmer T, Lee H. Reliability and validity of two self-report measures of impairment and disability for MS. Neurology. 1999;52:63.
47. Yorkston KM, Kuehn CM, Johnson KL, Ehde DM, Jensen MP, Amtmann D. Measuring participation in people living with multiple sclerosis: a comparison of self-reported frequency, importance and self-efficacy. Disabil Rehabil. 2008;30:88–97.
48. Fougeyrollas LN. Long-term consequences of spinal cord injury on social participation: the occurrence of handicap situations. Disabil Rehabil. 2000;22:170–80.
49. Read JL, Quinn RJ, Hoefer MA. Measuring overall health: an evaluation of three important approaches. J Chronic Dis. 1987;40:7S–21S.
50. Kaplan RM, Sieber WJ, Ganiats TG. The Quality of Well-Being Scale: comparison of the interviewer-administered version with a self-administered questionnaire. Psychol Health. 1997;12:783–91.
51. Andresen EM, Rothenberg BM, Kaplan RM. Performance of a self-administered mailed version of the Quality of Well-Being (QWB-SA) questionnaire among older adults. Med Care. 1998;36:1349–60.
52. Davis S, Wailoo A. A review of the psychometric performance of the EQ-5D in people with urinary incontinence. Health Qual Life Outcomes. 2013;11:20.
53. Whitehurst DG, Noonan VK, Dvorak MF, Bryan S. A review of preference-based health-related quality of life questionnaires in spinal cord injury research. Spinal Cord. 2012;50:646–54.
54. Tosh J, Brazier J, Evans P, Longworth L. A review of generic preference-based measures of health-related quality of life in visual disorders. Value Health. 2012;15:118–27.
55. Lin FJ, Longworth L, Pickard AS. Evaluation of content on EQ-5D as compared to disease-specific utility measures. Qual Life Res. 2013;22:853–74.

Supplementary Material

Search Strategy for OVID MEDLINE

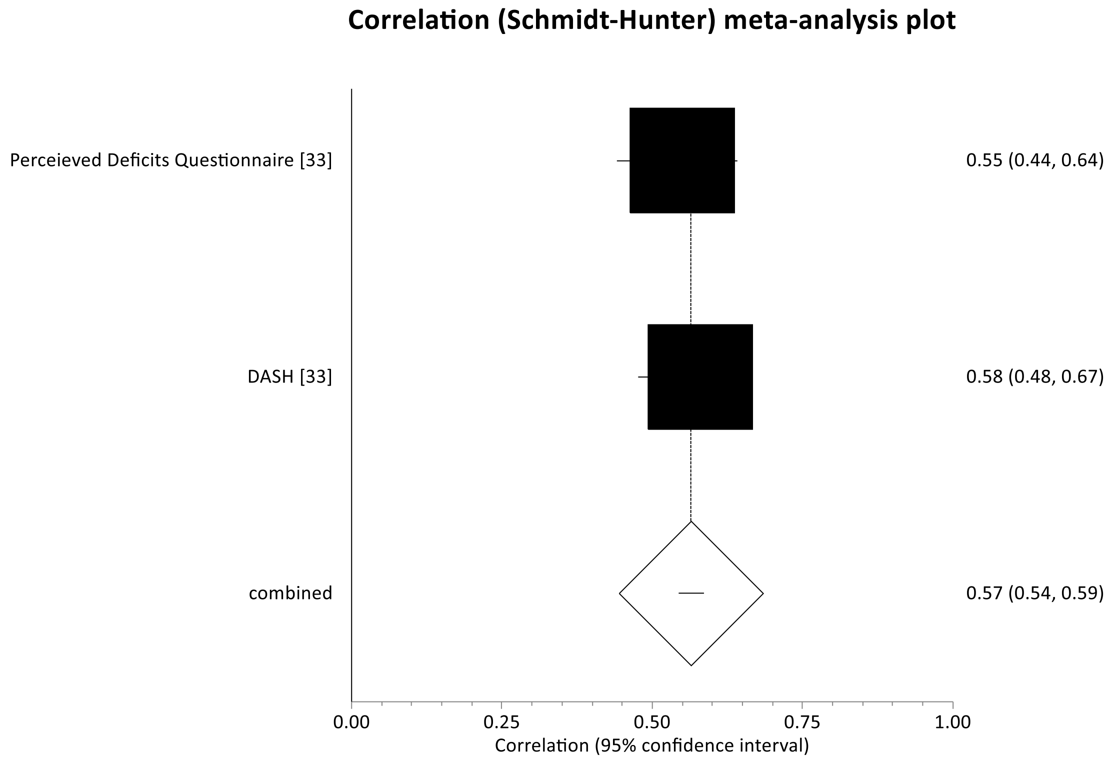
1. Multiple sclerosis.mp.
2. Health utilities index.mp.
3. HUI2.mp.
4. HUI3.mp.
5. EQ-5D.mp.
6. EuroQol.mp.
7. 15D.mp.
8. SF-6D.mp.
9. Assessment of Quality of Life.mp.
10. AQOL.mp.
11. Quality of Well Being.mp.
12. QWB.mp.
13. 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12
14. 1 and 13

Supplementary Figure 1 Forest plot with correlation coefficients (*r*) of the EQ-5D against outcomes of activity limitations and participation restrictions.



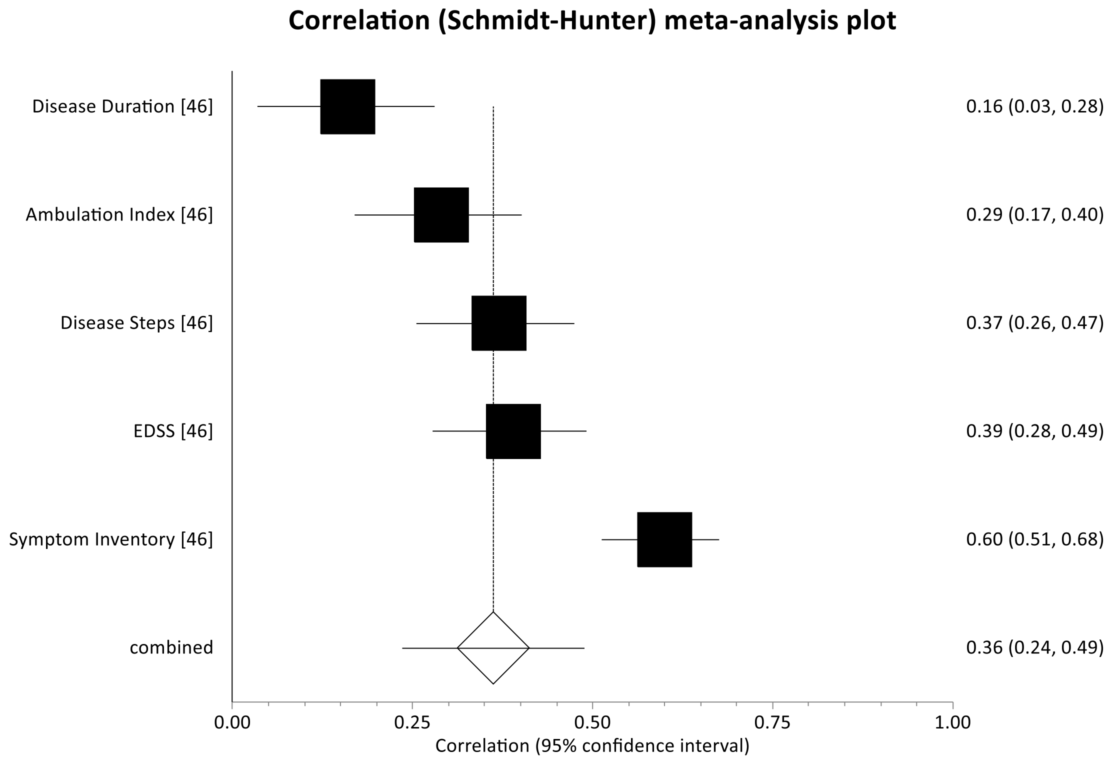
DASH Disabilities of the Arm Shoulder and Hand.

Supplementary Figure 2 Forest plot with correlation coefficients (*r*) of the SF6D against outcomes of activity limitations/participation restrictions.



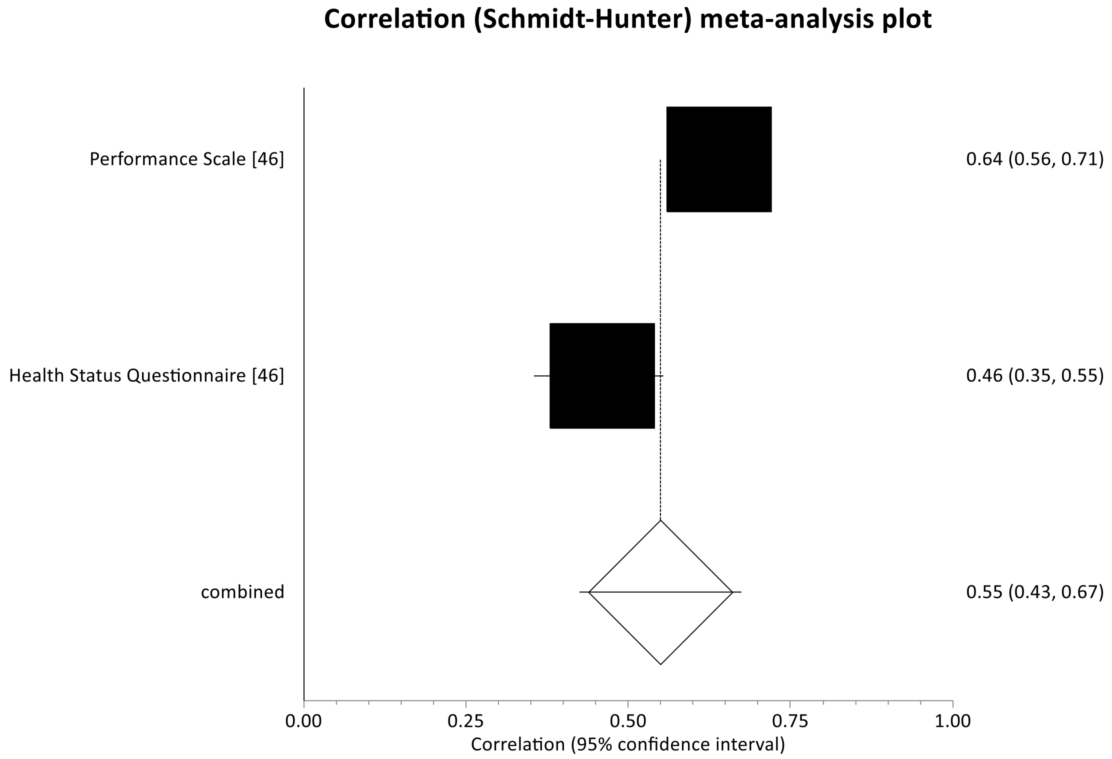
DASH Disabilities of the Arm Shoulder and Hand.

Supplementary Figure 3 Forest plot with correlation coefficients (*r*) of the Quality of Well Being scale against outcome measures evaluating impairments of body structure and function.



EDSS Expanded Disability Status Scale.

Supplementary Figure 4 Forest plot with correlation coefficients (*r*) of the Quality of Well Being scale against outcomes of activity limitations and participation restrictions.



Supplementary Table 1 Methodological quality assessment of included studies

	Item												
Author, Year	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Fogarty (2013)[41]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	N	Y
Kuspinar (2013)[33]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	Y	Y
Fisk (2005)[13]	Y	N/A	Y	N/A	Y	Y	N	Y	Y	Y	Y	Y	Y
Moore (2004)[38]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
Nicholl (2001)[39]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
Rothwell (1997)[40]	Y	Y	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	N	Y
Grima (2000)[42]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	Y	Y

Y Yes, *N* No, *N/A* Not applicable, *1* If human subjects were used, did the authors give a detailed description of the sample of subjects used to perform the (index) test?, *2* Did the authors clarity the qualification, or competence of the rater(s) who performed the (index) test?, *3* Was the reference standard explained?, *4* If interrater reliability was tested, were raters blinded to the findings of other raters?, *5* If intrarater reliability was tested, were raters blinded to their own prior findings of the test under evaluation? *6* Was the order of examination varied? *7* If human participants were used, was the time period between the reference standard and the index test short enough to be reasonably sure that the target condition did not change between the two tests?, *8* Was the stability (or theoretical stability) of the variable being measured taken into account when determining the suitability of the time interval between repeated measures?, *9* Was the reference standard independent to the index test?, *10* Was the execution of the (index) test described in sufficient detail to permit replication of the test?, *11* Was the execution of the reference standard described in sufficient detail to permit its replication?, *12* Were withdrawals from the study explained?, *13* Were the statistical methods appropriate for the purpose of the study?

CHAPTER 8: Integration of Manuscripts 3 and 4

Research questions of Manuscripts 3 and 4

Manuscript 3:

A review of the psychometric properties of generic utility measures in multiple sclerosis.

Manuscript 4:

Using existing data to identify candidate items for a health state classification system in multiple sclerosis.

Integration of Manuscripts 3 and 4

The previous manuscript evaluated the psychometric properties of existing generic preference-based measures in people with MS. It demonstrated that there were weaknesses with each of the generic measures, in terms of their lack of content coverage, their weak to moderate correlations with other HRQL measures, and their inability to discriminate between different levels of disability. The previous manuscript reinforced the need for a MS specific preference-based measure which can be used to evaluate the clinical and cost-effectiveness of different interventions for MS.

The structure of a preference-based measure is its classification system which has two components: the items and the response options, which are valued in combination with other items to produce a utility value. The next manuscript will describe the methodology used to identify the items and the response options for a MS specific preference-based measure. The discriminative capacity of the response options will be tested by cross walking onto a visual analogue scale (VAS) of health rating.

CHAPTER 9 (MANUSCRIPT 4)

Using existing data to identify candidate items for a health state classification system in multiple sclerosis

Ayse Kuspinar[1], Lois Finch[2], Simon Pickard[3], Nancy E. Mayo[1,2]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada

[2]Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

[3]Center for Pharmacoepidemiology and Pharmacoeconomic Research and Department of PharmacySystems, Outcomes and Policy, University of Illinois at Chicago, Chicago, IL, USA.

Published in *Quality of Life Research* (2014) 23:1445-1457.

Communication addressed to:

Ayse Kuspinar, M.Sc., Ph.D. Candidate
School of Physical & Occupational Therapy
Faculty of Medicine, McGill University
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
Tel: 514-934-1934 ext 31564
E-mail: ayse.kuspinar@mail.mcgill.ca

DOI 10.1007/s11136-013-0604-5

Using existing data to identify candidate items for a health state classification system in multiple sclerosis

Ayse Kuspinar · Lois Finch · Simon Pickard ·
Nancy E. Mayo

Accepted: 10 December 2013 / Published online: 15 December 2013

Abstract

Purpose In multiple sclerosis (MS), the use of preference-based measures is limited to generic measures such as Health Utilities Index Mark 2 and 3, the EQ-5D and the SF-6D. However, the challenge of using such generic preference-based measures in people with MS is that they may not capture all domains of health relevant to the disease. Therefore, the main aim of this paper is to describe the development of a health state classification system for MS patients. The specific objectives are: (1) to identify items best reflecting the domains of quality of life important to people with MS and (2) to provide evidence for the discriminative capacity of the response options by cross-walking onto a visual analog scale of health rating.

Methods The data come from an epidemiologically sampled population of people with MS diagnosed post-1994. The dataset consisted of 206 items relating to impairments, activity limitations, participation restrictions, health perception and quality of life. Important domains were identified from the responses to the Patient Generated Index, an individualized measure of quality of life. The extent to which the items formed a uni-dimensional, linear construct was estimated using Rasch analysis, and the best item was selected using the threshold map.

Results The sample was young (mean age 43) and predominantly female ($n = 140/189$; 74 %). The P-PBMSI classification system consisted of five items, with three response levels per item, producing a total of 243 possible health states. Regression coefficient values consistently decreased between response levels and the linear test for trend were statistically significant for all items. The linear test for trend indicated that for each item the response options provided the same discriminative ability within the magnitude of their capacity. A scoring algorithm was estimated using a simple additive formula. The classification system demonstrated convergent validity against other measures of similar constructs and known-groups validity between different clinical subgroups.

Conclusion This study produced a health state classifier system based on items impacted upon by MS, and demonstrated the potential to discriminate the health impact of the disease.

Keywords Health-related quality of life · Utility · Preference-based measures · Multiple sclerosis

Introduction

Several new therapies, behavioral [1–7], medical [8–12] and surgical [13–18] have been developed for multiple sclerosis (MS). Preference-based measures of health-related quality of life (HRQL) allow us to assess the positive and negative effects of these interventions from the patients' perspective. Preference-based measures are developed using multi-attribute utility theory [19–21], and consist usually of one item per dimension. Ideally, these

A. Kuspinar (✉) · N. E. Mayo
Faculty of Medicine, School of Physical and Occupational Therapy, McGill University, 3654 Prom Sir William Osler, Montreal, QC H3G 1Y5, Canada
e-mail: ayse.kuspinar@mail.mcgill.ca

L. Finch · N. E. Mayo
Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

S. Pickard
Department of Pharmacy Systems, Outcomes and Policy, Center for Pharmacoepidemiology and Pharmacoeconomic Research, University of Illinois at Chicago, Chicago, IL, USA

dimensions are independent from each other [19–22], but in some health conditions, such as mental illness [23] or complex diseases such as MS, restricting the content to independent dimensions may not adequately reflect the HRQL of the population targeted. For example, Mavranezouli et al. [23] developed a preference-based measure for mental health (CORE-6D), where 5 out of the 6 items in the measure were highly correlated with each other (loneliness, anxiety, humiliation, risk/harm to self and general functioning). Preference-based measures attach weights to the various dimensions of health, allowing trade-offs to be made between them [24, 25]. They provide a single value for overall HRQL [24, 26–29] that can range from 0 (for death or worst possible health state) to 1 (for perfect health or best possible health state) [20, 24, 27, 30, 31]. This single value can be linked to life expectancy to calculate quality-adjusted life years (QALYs) [22, 26, 32, 33]. QALYs provide a single comprehensive measure of health improvement that captures the effect of an intervention on both mortality (quantity of life) and morbidity (quality of life) [22, 30, 34]. The QALY can be used to compare and make decisions about the clinical and cost-effectiveness of different interventions [22].

In MS, the use of preference-based measures is limited to generic measures such as Health Utilities Index Mark 2 and 3 (HUI II and III), the EQ-5D and the SF-6D [26]. However, the challenge of using such generic preference-based measures in people with MS is that they may not capture all domains of health relevant to the disease [35] either as benefits or harms. For example, fatigue, the most common symptom occurring in MS is included neither in the EQ-5D nor in the HUI. Walking, another common problem among patients with MS is not included in the SF-6D, and cognition is not included in the EQ-5D. Also, generic preference-based measures may be criticized for having low construct validity in MS. Studies have shown generic preference-based measures to have low correlations with disease-specific health profiles [36, 37], to have limited discriminative ability between different levels of disease severity and are prone to ceiling and floor effects [38].

Preference-based measures are typically generic in nature. They are designed to include a common set of dimensions that most people will value highly [39] and therefore may not include those dimensions that are specific to a disease. Disease-specific measures are designed to fill in the gaps in generic measures by tapping specific domains.

Disease-specific preference-based measures have been developed for stroke [40], cancer [41, 42], coronary artery disease [43], pulmonary hypertension [44], asthma [45], rhinitis [46], urinary incontinence [47], erectile dysfunction [48] and mental health [23]. They are designed to provide additional information not captured by generic measures. Disease-specific preference-based measures are more tailored to patients' needs and better able to detect changes in patients' health status over time [22, 27, 49]. They are also better able to capture subtle changes in clinical status that are not captured with generic measures [22, 26, 47, 50]. Furthermore, in line with Food and Drug Administration guidelines [51], preferences for health states can be obtained directly from patients rather than the general public [40, 52]. The classification system that will be developed will be an MS-specific instrument with weights obtained using patient values.

The structure of a preference-based measure is its classification system which has two components: the items and the response options, which are valued either alone or in combination with other items to produce a utility value. This paper will describe the methodology used to identify items and the response options for the classification system. No attempt at valuing the classification system will be made at this time.

Therefore, the main aim of this paper is to describe the development of a health state classification system for MS patients, for which preference weights will be obtained later on, to develop into a preference-based measure. As it is not a preference-based measure at its current state, the health state classification system will be referred to as the *prototype* Preference-Based MS Index (P-PBMSI). The specific objectives of this developmental work were: (1) to identify items best reflecting the domains of quality of life important to people with MS and (2) to provide evidence for the discriminative capacity of the response options by cross-walking onto a visual analog scale (VAS) of health rating.

Methods

Data source

This study performed secondary analysis on an existing dataset that assessed the life-impact of people diagnosed with MS during the era of imaging and disease modifying therapies (the New MS) [53]. The available study population consisted of both men and women who had been registered from 1994 to 2008 at the three participating MS clinics in Greater Montreal, Quebec, Canada. The study was approved by all regional ethics committees. Inclusion criteria for the study were diagnosis of MS or clinically isolated syndrome (which is the initial neurological manifestation of MS) after 1994.

From a pool of 5,000 patients, a center-stratified random sample of 550 patients was drawn, of which 394 were contacted. Stratification strategy was by hospital site or clinic. There were 3 MS clinics involved. From those who were contacted, the first 189 persons who responded were enrolled. Duration of the disease, type of MS and patients score on Expanded Disability Status Scale (EDSS) were determined from patients' medical charts. All subjects were asked to complete a comprehensive questionnaire package. The questionnaire package included over 206 items relating to function (impairments, activity limitations and participation restrictions), health perception and quality of life. The World Health Organization's international classification of functioning, disability and health (ICF) was used as a framework to identify the items that needed to be included in the questionnaire package. If an item could be found in an existing patient-reported outcome measure, then it was included in the questionnaire package, if not, then it was created from scratch. This process was conducted with a multi-disciplinary team of neurologists, epidemiologists, psychologists, psychiatrists, physical therapists and occupational therapists. The multi-disciplinary approach and the use of the ICF as a framework insured that all areas of function and HRQL important to MS were captured in the questionnaire package.

The following measures that were in the existing dataset were used in this secondary analysis to assess construct validity of the classification system: the RAND-36, the EQ-5D, the Patient Generated Index (PGI), the Perceived Deficits Questionnaire (PDQ), the Six-Minute Walk Test (6MWT) and the EDSS. The RAND-36 is one of the most widely used generic health profiles with good internal consistency, convergent and discriminate validity with other health measures [54, 55]. The EQ-5D is a generic utility or preference-based measure that has shown moderate correlations with measures of MS disability and ambulation [38, 56, 57]. The PGI is an individualized measure of quality of life where patients are asked to identify up to five of the most important areas of their lives affected by their condition. The reliability, validity and responsiveness of the PGI have been assessed on patients with low back pain, menorrhagia, suspected peptic ulcer and varicose veins [58]. PGI scores for the four conditions showed significant small to moderate correlations with the eight subscales of the SF-36. The PDQ is a patient-reported outcome of cognitive impairment developed specifically for use in MS. The PDQ has demonstrated convergent validity, internal consistency and test–retest reliability in MS [59, 60]. The 6MWT is a simple performance-based test that measures functional exercise capacity. The reliability of the 6MWT has been assessed in persons with MS. The intra-class correlation coefficient is 0.96 for test–retest reliability and 0.93 for inter-rater reliability [61]. The EDSS is a widely used scale to measure level of disability in patients with MS. It is a classification scheme extending from 0 (normal neurological examination) to 10 (death due to MS) [62].

Data analysis and procedure

The authors decided beforehand that the classification system would be based on a limited number of key items to reduce administration time, response burden and the number of possible 'unrealistic' health states. Preference-based measures generate n^i unique health states ($n =$ number of response options and $i =$ number of items). Producing a large number of health states is not a hindrance in itself, as a value can be assigned to any possible health state using mathematical modeling. However, some 'theoretically' possible health states will not be 'realistically' possible [22]. Hence, reducing the number of items (or response options) is likely to minimize the number of 'unrealistic' health states and response burden.

Figure 1 gives an overview of the steps leading from domains to items to a potential scoring algorithm. For this study, the important domains were identified from the responses to the PGI [63]. Items on function came from a variety of patient-reported outcome measures, and health rating came from a VAS that was anchored from 0 (worst imaginable health state) to 100 (best imaginable health state). The VAS was equivalent to the one that accompanies the EQ-5D.

Each patient's response on the PGI was mapped to the ICF domains independently by four raters (Box A of Fig. 1) and the results have been reported previously [35]. The top ten domains that patients identified to be most important to their quality of life were included for further investigation in the study. The ICF provided a coding framework and standardized description of health-related problems at the level of body structure/function (e.g., fatigue, cognition), activity (e.g., dressing, feeding) and participation (e.g., school, work). These levels are also known as impairments, activity limitations and participation restrictions, respectively. Any discrepancies between raters were resolved by discussion. For example, fatigue was a key domain and 23 items relating to fatigue were available in the database. All other questionnaire items were mapped onto each domain. Items were selected from a range of patient-reported outcomes that were available in the database and that measured the domain of interest. The extent to which the items formed a uni-dimensional, linear (latent) construct was estimated using Rasch analysis. Factor analysis was not used to identify and select items for the health state classification system for two important reasons. First, factor analysis is faulted for mistaking ordinal observations for linear measures [64, 65]. All of our items were ordinal with three response levels each. Second, factor analysis is unable to identify the location of each variable along a linear continuum [64–66]. As item selection was primarily based on the location of each response level (i.e., thresholds) along the latent scale, Rasch analysis

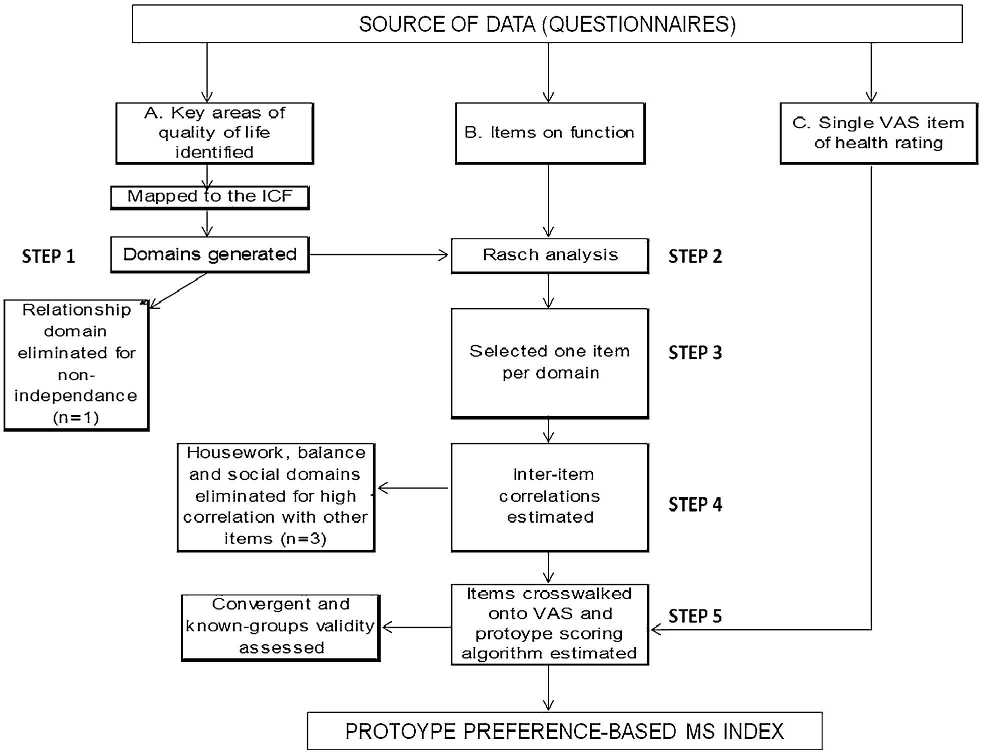

SOURCE OF DATA (QUESTIONNAIRES)
A. Key areas of quality of life identified
B. Items on function
C. Single VAS item of health rating
Mapped to the ICF
STEP 1
Domains generated
Rasch analysis
STEP 2
Relationship domain eliminated for non-independance (n=1)
Selected one item per domain
STEP 3
Housework, balance and social domains eliminated for high correlation with other items (n=3)
Inter-item correlations estimated
STEP 4
Convergent and known-groups validity assessed
Items crosswalked onto VAS and protoype scoring algorithm estimated
STEP 5
PROTOYPE PREFERENCE-BASED MS INDEX

Fig. 1 Flowchart of the steps involved in developing the P-PBMSI

served the purpose of this study better than factor analysis and thus was the preferred method for constructing the classification system.

Rasch analysis is conventionally used to develop new measures or refine existing ones. The Rasch measurement models developed in this study were used to select the items that best reflected the full continuum of the latent construct. This procedure followed the method by Brazier [22] and Young [67].

The procedure for the Rasch analysis was the same for all domains (Step 2). All analyses were performed using the Rasch Unidimensional Measurement Model (RUMM2020) software [68]. Good model fit was indicated by a non-significant Chi square statistic with Bonferroni adjustment and item/person fit residuals that were close to zero with a standard deviation of 1. The residual correlation matrix was observed for possible response dependency, and uni-dimensionality of the models was verified using principal component analysis of the residuals followed by independent t-tests if there was a question of multi-dimensionality. Individual item fit was assessed using the Chi square probability value and the fit residual values. Misfit was indicated by a Chi square probability value of <0.05 (using a Bonferroni adjustment) and fit residual values ≥ 2.5 [69, 70]. Reliability of the scale was assessed using the Person Separation Index (Rasch-based equivalent to Cronbach's Alpha), where a value of 0.7 or greater was considered acceptable [69, 71].

One item per domain was selected using the procedure described by Brazier [22] and Young [67] (Step 3). This procedure involves selecting an item based on the spread of threshold values across the latent scale that are determined through the threshold map. With k ordinal response options, there are $k - 1$ thresholds which indicate the number of progressions or steps (from lower to higher) there are in an item. An item with three response levels would have 2 thresholds. As the EQ-5D uses a 3-level response, this was chosen in order to facilitate rating and minimize cognitive burden on patients. Rasch analysis converts ordinal responses to linear through a logit transformation. On the logit scale a construct ranges from negative logit values (capturing people with severe problems for that item) to zero or neutral (capturing people with moderate problems for that item) to positive logit values (capturing people with no problems for that item). Thus, the ideal item would be centered on logit zero (neutral) and have a negative and a positive threshold. This would ensure that response levels captured people with a range of disability (severe, moderate and mild).

Item fit was also considered. A small Chi square with a fit residual close to zero, the better that item represented the underlying construct [67]. The point–biserial correlation (the correlation between the item score and the domain score) was also taken into consideration, where values greater than 0.3 were acceptable [72].

After the best item per domain was selected, polychoric correlation coefficients were calculated, and best-domain items that were highly correlated with each other were eliminated to increase structural independence (Step 4).

Having selected the best candidate items for each domain the next step was to value whether the response options for each item adequately discriminated the construct of interest (Step 5). Each item was mapped onto the VAS. The regression weights obtained were not preference weights, but provided estimates of how the response options were spread across the health construct. Level 1 of each item was denoted as the best level and level 3 was the worst level. The gradient across levels was estimated from the regression coefficient values and from the linear test for trend. Regression coefficient values were expected to decrease between each response level when treated as categorical variables (i.e., larger negative values), and we expected a statistically significant ($p < 0.05$) trend of decreasing health by increasing severity from the linear test for trend. In order to provide evidence that the classification system related to other measures of MS disability and quality of life, we created a regression-based scoring system for the P-PBMSI by regressing the items onto the VAS scores provided by each respondent. The regression coefficients were used in a simple additive formula: 100 + [the decrement (i.e., negative weight)] associated with the level of each item.

The frequency and proportion of individuals at different levels of disability (none, mild, mild to moderate, moderate to severe and severe) based on the P-PBMSI health state classifier system were reported. Individuals who had all item levels equal to 1 were classified as having no disability (11111). Individuals with one item level equal to 2 was classified as having mild disability (e.g., 21111), individuals with two or more item levels equal to 2 (but no 3 s) were classified as having mild to moderate disability (e.g., 22111). Moderate to severe disability was indicated by having one or two items with a response level equal to 3 (e.g., 33111), and severe disability was indicated by having three or more items with response levels equal to 3 (e.g., 33311).

The construct validity of the P-PBMSI health state classifier system was evaluated using convergent validity. We hypothesized, using Cohen's criteria [73], that there would be strong correlations ($r > 0.5$) between the (a) P-PBMSI walking item and the Physical Function subscale of the RAND-36, (b) P-PBMSI fatigue item and the vitality subscale of the RAND-36, (c) P-PBMSI cognition item and the perceived deficits questionnaire (cognition questionnaire), (d) P-PBMSI mood item and the mental health subscale of the RAND-36, (e) P-PBMSI work item and the work question from the illness intrusiveness rating scale and (f) P-PBMSI total score and the PGI (for quality of life). Spearman's rank correlation was used for ordinal variables, and Pearson's correlation for continuous variables.

Furthermore, the known-groups method was used to test the discriminative ability of the P-PBMSI index and the EQ-5D index against different clinical subgroups as measured using the EDSS, the 6MWT and the general health perception item of the RAND-36.

Statistical analysis was carried out using the Statistical Analysis Systems (SAS) Version 9.2.

Results

The dataset included 189 persons with MS. The sample was young (mean age 43) and predominantly female (Table 1). Both men and women had mild disability with a median EDSS score of 2. The average number of years since diagnosis was 6 years, and 59 % of the sample was on disease modifying therapies.

The top ten domains that patients identified to be most affected by their MS were: school/work, fatigue, sports, social life, relationships, walking, cognition, balance, housework and mood (Step 1), all of which were identified in the ICF core sets for MS. Relationship was excluded from the preference-based measure for non-independence, because the literature shows that it is a downstream effect of other domains such as mood and fatigue [74–76].

All Rasch measurement models met the criteria for good fit, with non-significant Chi square probability values with Bonferroni adjustment, high reliability (Persons Separation Index > 0.7) and mean item and person fit residuals close to zero with standard deviation of one (Step 2). The domains walking and sports were combined in one Rasch model and analyzed together, as they were highly correlated and were part of the broader construct of physical function. The Rasch model fit statistics are presented in Table 2.

One item was selected from each Rasch model (Step 3). All selected items had threshold values, as expected,

Table 1 Demographic and clinical characteristics of sample ($n = 189$)

Characteristics	Mean (SD) or *N* (%)
Age (year)	43.0 (10.2)
Women/men	140/49 (74/26)
Definite MS/CIS[a]	170/15 (92/8)
Year since diagnosis	6.2 (3.6)
EDSS, median (IQR)	2.0 (1.0–3.5)
On DMT/not on DMT/no information	112/21/56 (59/11/30)
Patient generated index[a]	0.50 (0.25)
EQ-5D	0.69 (0.18)

SD standard deviation, *no.* number, *CIS* clinically isolated syndrome, *EDSS* expanded disability status scale, *IQR* inter-quartile range, *DMT* disease modifying therapies

[a] Missing data on four subjects

Table 2 Rasch model goodness of fit statistics for each domain

Domain (*N* items/*N* thresholds)	Chi Sq goodness of fit (*N* degrees of freedom)	*p* value (Chi Sq)	Person Separation Index	Item fit mean (SD)	Person fit mean (SD)	Threshold range
Walking[a] (6/12)	22.4 (18)	0.21	0.94	−0.53 (0.52)	−0.27 (0.39)	−6.61 to 7.47
Work (4/6)	16.1 (8)	0.04	0.78	−0.49 (0.61)	−0.23 (0.64)	−3.60 to 1.35
Fatigue (6/18)	15.4 (18)	0.63	0.93	0.06 (0.89)	−0.37 (1.11)	−3.35 to 3.66
Social life (4/10)	16.1 (12)	0.19	0.82	−0.33 (1.93)	−0.42 (0.82)	−2.76 to 2.90
Balance (10/19)	25.1 (20)	0.20	0.94	−0.27 (0.59)	−0.39 (0.54)	−8.70 to 4.16
Mood (8/20)	32.7 (24)	0.11	0.84	−0.31 (1.26)	−0.38 (0.98)	−3.68 to 4.62
Cognition (19/72)	96.3 (76)	0.06	0.94	−0.05 (1.08)	−0.29 (1.29)	−4.47 to 3.86

Rasch analysis for housework not carried out as there was only one item in the patient-reported outcome measure that represented this domain

N number, *Chi Sq* Chi square, *SD* standard deviation

[a] Rasch model includes items on sports

ranging from negative to positive logits with the middle response level centered on logit zero. Figures 2a, b present, as examples, the Rasch measurement models for walking and fatigue. One item was selected from each model based on the threshold map. Figure 3 presents the spread of threshold values across the Rasch scale for all of the selected items. Table 3 presents further information on the threshold range and fits statistics for the selected items. All items fit the Rasch measurement model indicating that the item was representative of the domain. All of the point–biserial correlation coefficient values ranged between 0.63 to 0.89, indicating that the item correlated with the overall scale and was representative of the construct.

Three items, namely balance, housework and social life, were eliminated after inter-item correlations were calculated (Step 4). Balance was highly correlated with walking ($r = 0.8$), housework was highly correlated with work ($r = 0.8$) and social life was highly correlated with mood and fatigue ($r = 0.7$).

The P-PBMSI classification system consisted of five items, with three response levels per item, producing a total of 243 possible health states. Table 4 presents the regression coefficient values for each corresponding response level of the P-PBMSI to the VAS. Regression coefficient values consistently decreased between response levels and the linear test for trend was statistically significant for all items. The linear test for trend indicated that for each item the response options provided the same discriminative ability within the magnitude of their capacity. Using an additive formula (Step 5), the simple linear regression coefficient values of each item and its corresponding response levels were summed together. A value of 0 was given to an item if the response level was 1, because the reduction for that item was zero. To provide an illustration of the discriminative ability of our classification system, consider the following scenarios: a patient who had no problems with any of the items (health state 11111) would have a P-PBMSI score of 100. A patient with health state 23112 would have a P-PBMSI score of 45.1 $[(100 + (-16.0 - 24.7 - 0 - 0 - 14.2))]$.

Table 5 presents the frequency (and percentage) of subjects with none, mild, mild to moderate, moderate to severe and severe levels of disability, based on the P-PBMSI classification system. The P-PBMSI identified 17 subjects (9 %) as having no disability, 23 subjects (12 %) as having mild disability, 73 subjects (39 %) as having mild to moderate disability, 56 subjects (30 %) as having moderate to severe disability and 20 subjects (10 % of the sample) as having severe disability. The overall frequency and proportion of individuals were normally distributed for the P-PBMSI, indicating that the P-PBMSI classification system may have potential to discriminate between different levels of disability.

The P-PBMSI items and total score demonstrated convergent validity (Table 6). As hypothesized moderate to strong correlations (>0.5) were observed between the: P-PBMSI walking item and Physical Function subscale of the RAND-36, P-PBMSI fatigue item and vitality subscale of the RAND-36, P-PBMSI cognition item and PDQ, P-PBMSI mood item and mental health subscale of the RAND-36, P-PBMSI work item and hours of paid work per week, and the P-PBMSI total score and the PGI.

For known-groups validity (Table 7), both the P-PBMSI and the EQ-5D indices were able to discriminate between different clinical subgroups, functional walking capacity and general health perception. However, the P-PBMSI provided a wider range of values than the EQ-5D (36 vs. 26 on the EDSS, 43 vs. 27 on the 6MWT and 66 vs. 52 on general health perception).

Discussion

This study used an existing dataset on the life-impact of MS to estimate a scoring algorithm for a prototype preference-based index for MS, targeting the health effects and

Fig. 2 **a** Threshold map for the Rasch measurement model on walking. **b** Threshold map for the Rasch measurement model on fatigue

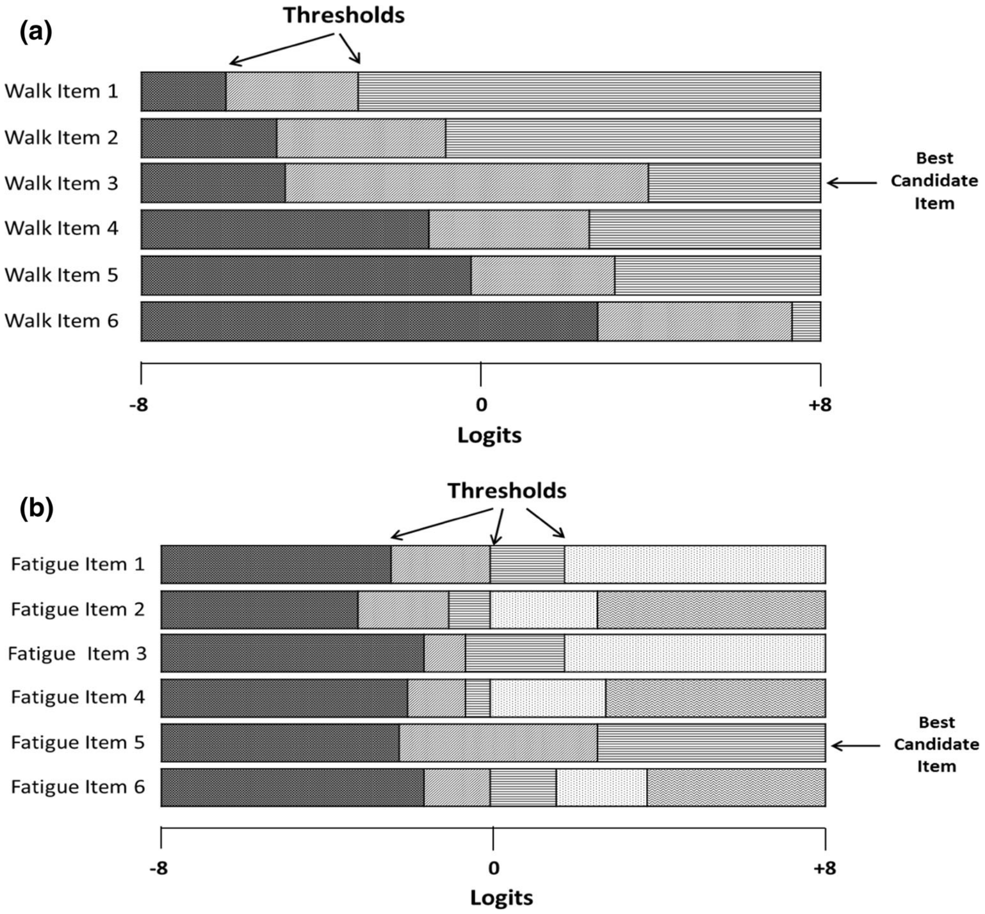


Fig. 3 Threshold map for all of the selected items demonstrating that, for each item, the threshold value ranged from negative to positive logits, with the middle response level centered on 0. The exact threshold value for each item can be found in Table 3

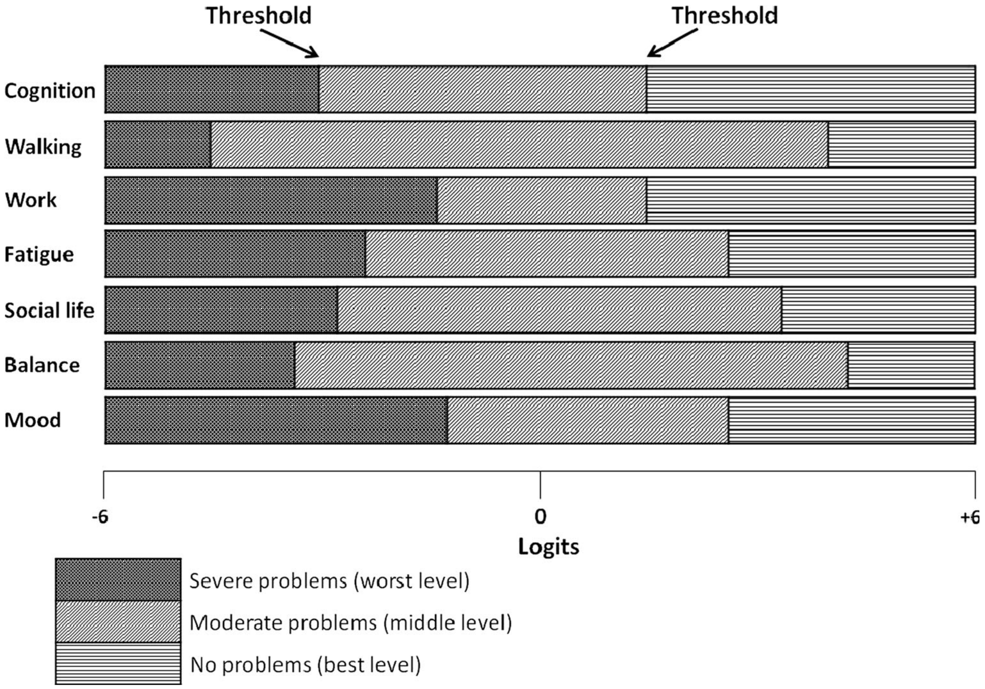


Table 3 Individual fit for the selected items

Item	Location	SE	Fit residual	Chi Sq (*N* df)	*p* value	Threshold range	Point-biserial correlation
Cognition	−0.84	0.13	−1.79	6.89 (4 df)	0.14	−2.97 to 1.30	0.63
Walking[a]	−0.37	0.18	−0.75	2.96 (3 df)	0.12	−4.76 to 4.02	0.85
Work	0.08	0.13	−1.28	3.16 (2 df)	0.21	−1.20 to 1.35	0.89
Fatigue	0.13	0.13	0.38	3.60 (3 df)	0.31	−2.16 to 2.45	0.79
Social life	0.16	0.16	−0.32	1.50 (3 df)	0.68	−2.76 to 3.28	0.77
Balance	0.37	0.21	0.51	3.24 (2 df)	0.20	−3.42 to 4.16	0.66
Mood	0.66	0.11	−2.32	7.66 (3 df)	0.05	−1.18 to 2.49	0.77

Chi Sq Chi square, *SE* standard error, *N df* number of degrees of freedom (equal to the number of class intervals −1)

[a] Walking and sport item combined in a subtest in the Rasch Model

Table 4 Regression coefficient values for simple linear regression analyses and linear test for trend

Items	Simple linear regression Regression coefficient (SE)	Linear trend test Regression coefficient (*p* value)
Walking 1[a]	Referent (0)	−16.6 ($p < 0.0,001$)
Walking 2	−16.0 (2.1)	
Walking 3	−35.7 (5.1)	
Fatigue 1[a]	Referent (0)	−12.0 ($p < 0.0001$)
Fatigue 2	−11.2 (2.6)	
Fatigue 3	−24.7 (3.4)	
Cognition 1[a]	Referent (0)	−7.1 ($p < 0.0001$)
Cognition 2	−9.1 (2.6)	
Cognition 3	−13.4 (3.2)	
Mood 1[a]	Referent (0)	−6.8 ($p = 0.0002$)
Mood 2	−6.2 (2.8)	
Mood 3	−13.9 (3.6)	
Work 1[a]	Referent (0)	−13.2 ($p < 0.0001$)
Work 2	−14.2 (2.2)	
Work 3	−26.0 (2.8)	

[a] The first response level of each item is the intercept: Walking 1 = 83.8, Fatigue 1 = 82.3, Cognition 1 = 78.2, Mood 1 = 79.2 Work 1 = 82.3

health decisions required by people in the early stages of MS. We particularly chose our sample to represent people diagnosed with MS post-1994, when results of neuroimaging were the diagnostic criterion, and when disease modifying drugs became available. This group has been labeled as having the New MS [53]. Thus, the domains selected for prototype scoring and future valuation represent the priorities for this target group.

The P-PBMSI classification system consisted of five items: walking, fatigue, cognition, mood and work. No one item covered the full range from worst possible health state (0) to best possible health state (100), indicating that no one item is sufficient to represent the full spectrum of MS impact [77]. Interestingly, the item with the widest range of impact was walking in response to decrements of −16 and −36. No single item provided a negative health impact more than −36. The five items in the classification system were different from those found in generic preference-based measures namely, the HUI II, HUI III, EQ-5D and SF-6D. Fatigue which affects 75–90 % of patients with MS [78–81] is not included in the EQ-5D or the HUI measures. Fatigue was originally included in the EQ-5D but was later dropped because when combined with the other variables, it did not reach statistical significance. This is probably because preferences were obtained from members of the general public who had never experienced MS fatigue. Our results indicated that this important item had the largest impact (largest regression coefficient value) on health after walking. Cognition, another important impairment in MS, is not included in the EQ-5D or the SF-6D. Cognitive impairment is recognized as an important consequence of MS [82] affecting up to 70 % of patients [83]. Absence of a cognition item in generic utility measures questions the content validity of these measures in people with MS. Content validity of a PRO can be judged only by the individuals or populations being assessed (i.e., the patients themselves). The new classification system has content

Table 5 Frequency and percentage of individuals ($n = 189$) with none, mild, mild to moderate, moderate to severe and severe levels of disability that the P-PBMSI classification system identified

Level of disability	Description	P-PBMSI *N* (%)
None	All items 1 (11111)	17 (9 %)
Mild	One item with 2 (21111, 12111, etc.)	23 (12 %)
Mild to moderate	Two or more items with 2, but no 3 (22111, 22211, etc.)	73 (39 %)
Moderate to severe	One or two items with 3 (33111, 33211, etc.)	56 (30 %)
Severe	Three or more items with 3 (33321, 33332, etc.)	20 (10 %)

Table 6 Convergent validity (correlations) between the P-PBMSI (individual items and total score) and different measures of similar construct

Measure	Walking[a]	Fatigue[a]	Cognition[a]	Mood[a]	Work[a]	Total score
PFI RAND-36	0.78	0.37	0.22	0.19	0.69	0.71
VIT RAND-36	0.40	0.65	0.47	0.41	0.50	0.69
PDQ[a]	0.27	0.51	0.69	0.48	0.41	0.64
MHI RAND-36	0.12	0.43	0.46	0.63	0.28	0.49
Work h/week	0.32	0.16	0.17	0.17	0.52	0.42
PGI	0.49	0.39	0.29	0.25	0.47	0.59

PFI RAND-36 physical function index subscale of RAND-36, *VIT RAND-36* vitality subscale of RAND-36, *MHI RAND-36* mental health index subscale of RAND-36, *PDQ* perceived deficits questionnaire, *Work h/week* hours of paid work per week, *PGI* patient generated index

[a] All variables were ordered so that higher scores were better

Table 7 Known-groups validity of the P-PBMSI total score (calculated using a mapping function against the VAS) and the EQ-5D index against external measures of disease severity

Measure	P-PBMSI Mean (SD)	EQ-5D Mean (SD)
EDSS		
0–2.5 (minimal disability)	69.3 (22.9)*	74.7 (12.9)*
3–5.5 (moderate disability)	51.1 (22.2)	63.3 (20.0)
6 + (severe disability)	33.7 (18.0)	48.7 (21.9)
6MWT		
600 + m	79.5 (19.3)*	78.2 (10.3)*
300–599 m	60.6 (23.0)	70.5 (15.7)
0–299 m	36.7 (23.8)	51.2 (23.0)
General health perception		
Excellent	84.6 (16.7)*	81.3 (6.6)*
Very good	73.7 (19.1)	76.4 (11.6)
Good	53.7 (24.8)	68.0 (17.4)
Fair	42.0 (19.4)	54.1 (19.7)
Poor	18.7 (19.3)	29.8 (27.1)

EDSS expanded disability status scale, *6MWT* six-minute walk test, *PBMSI* preference-based multiple sclerosis index, *m* meters, *SD* standard deviation

* $p < 0.0001$ with one way analysis of variance

validity in MS because it has been developed based on domains that were identified by patients to be most important to their quality of life. The absence of important domains may add doubt to the interpretability of scores produced by these measures in this population and may result in a false estimate of QALYs. Work, a participation item, is not found in the HUI II or HUI III. This is probably because the HUI measures were developed with the intention of evaluating ‘within-the-skin’ experiences that excluded items relating to participation in life roles.

The development of disease-specific measures is an emerging area of interest in the literature, as there are several potential benefits to using these measures in both clinical and cost-effectiveness research. Disease-specific preference-based measures are designed to include domains that are specific to a disease, therefore, are likely to be more sensitive to disease-specific changes which may be positive or negative. Furthermore, they not only provide descriptive information on the various domains of health, but also provide a value for each one, thus allowing trade-offs to be made between them. This advantage may be particularly important when interventions may have undesirable side effects and these effects are not part of the generic classifications (e.g., fatigue, nausea, cognitive changes, weight gain). Disease-specific utility measures serve the potential to overcome one of the challenges associated with disease-specific health profiles—that domains cannot be combined into a single index, which makes it difficult to conclude whether an intervention resulted in a net improvement or decline in HRQL. Furthermore, disease-specific utility measures can be used to make decisions on the cost-effectiveness of different treatments in MS.

A major strength of preference-based measures (generic or disease-specific) is their ability to take the health index score and link it to life expectancy to calculate QALYs [22, 26, 32, 33]. QALYs provide a single measure of health improvement that captures the effect of an intervention on both quantity of life and quality of life [22, 30, 34]. QALYs can provide information and help make decisions on the clinical and cost-effectiveness of different interventions in MS [22].

In this study, health state valuations were obtained directly from patients themselves, whereas many generic or disease-specific preference-based measures have been developed by asking the general public to value hypothetical health states. The main argument for the use of general population values is that it is society who pays for the services, and thus they should be the ones involved in health care decision making [22, 84, 85]. The challenge with this is that the general public has no experience of the health states that they are asked to value. Patients, on the other hand, know their health state better than anyone else and are the ones receiving the health care service or program [22, 84, 85]. An argument against the use of patient preferences is that it may make it difficult to compare the cost per QALY of a treatment for MS with, for example, the cost per QALY of a treatment for cancer. However, an underlying assumption of QALYs is that a QALY gained or lost is blind to health conditions and individual characteristics such as age, sex, disease severity, social roles, place of residence and other personal characteristics [86].

There is the reality that patients, who experience a particular health state, may not devalue it as much as a person without any experience with disability or illness. As a result, health states that are modifiable by interventions may not show up as desirable by the patients yet any change for better in this health state would be rated as highly desirable by the general public. There is concern that this utility value may play against patients having access to interventions that modify these health states. For example, society highly values walking, but patients who have walking disability and use a wheelchair may find this mode of mobility easier than technologically assisted walking [87–89]. For patients, the QALY assigned to change from not being able to walk at all to walking with some difficulty may not be as high as with general public values. This brings forth the question of whether this is bias or if it is the truth.

An alternative approach to the one we used is to derive a utility measure from an existing MS-specific health profile. We chose not to use this approach because we wanted to develop a measure that was specifically focused on the needs of our population, MS patients diagnosed in the era of magnetic resonance imaging (MRI) and disease modifying therapies. These are the people who are faced with treatment decisions; hence, we wanted to include content relevant to this target population. Although deriving a utility from an existing health profile would be useful for secondary data analysis when a utility measure was not used in primary data collection, the entire disease-specific measure would have to be administered. In MS, for example, all 54 items of the MSQOL-54 would need to be completed, over time, and the algorithm applied. In many clinical and research situations, having a shorter measure that produces the same utility is likely to be more feasible recognizing that there is a need to also collect data on performance measures to quantify MS impact. Thus, we chose the more parsimonious approach of directly asking patients to derive content and only querying this content in the utility measure.

There were limitations in this study that need to be noted. First, the outcome was global health rating using the VAS. Patients were asked to provide one number on the VAS that would best describe their current health; however, we do not know the appraisal process involved in selecting that value. The more accurate measurement would be to ask patients to value specific health states using one of the many standardized techniques (VAS or others). Second, the range for the prototype scoring algorithm was 100 (for health state 11111) to −13.0 (for health state 33333). As we did not have a value for dead, we could not anchor our results on a scale from dead to perfect health.

In conclusion, this study identified items that best reflected the domains of quality of life important to people with MS and mapped these items onto a VAS of health rating. The next step will involve conducting cognitive interviews with patients (following FDA guidelines) to ensure that phrasing of items and their response options are appropriately comprehended by patients [90, 91]. Last, a valuation study will be conducted for specific health states to obtain a final scoring algorithm for the preference-based MS index.

Acknowledgments This work was supported in part by the Canadian Institutes of Health Research and a scholarship from the Fonds de la Recherche en Santé du Quebec.

References

1. Miller, D. M., Moore, S. M., Fox, R. J., Atreja, A., Fu, A. Z., Lee, J. C., et al. (2011). Web-based self-management for patients with multiple sclerosis: A practical, randomized trial. *Telemedicine Journal & E-Health, 17*, 5–13.
2. Barlow, J., Turner, A., Edwards, R., & Gilchrist, M. (2009). A randomised controlled trial of lay-led self-management for people with multiple sclerosis. *Patient Education and Counseling, 77*, 81–89.
3. Bombardier, C. H., Cunniffe, M., Wadhwani, R., Gibbons, L. E., Blake, K. D., & Kraft, G. H. (2008). The efficacy of telephone counseling for health promotion in people with multiple sclerosis: A randomized controlled trial. *Archives of Physical Medicine and Rehabilitation, 89*, 1849–1856.
4. McAuley, E., Motl, R. W., Morris, K. S., Hu, L., Doerksen, S. E., Elavsky, S., et al. (2007). Enhancing physical activity adherence and well-being in multiple sclerosis: A randomised controlled trial. *Multiple Sclerosis, 13*, 652–659.
5. Grossman, P., Kappos, L., Gensicke, H., D'Souza, M., Mohr, D. C., Penner, I. K., et al. (2010). MS quality of life, depression, and fatigue improve after mindfulness training: A randomized trial. *Neurology, 75*, 1141–1149.
6. Forman, A. C., & Lincoln, N. B. (2010). Evaluation of an adjustment group for people with multiple sclerosis: A pilot randomized controlled trial. *Clinical Rehabilitation, 24*, 211–221.
7. Cosio, D., Jin, L., Siddique, J., & Mohr, D. C. (2011). The effect of telephone-administered cognitive-behavioral therapy on quality of life among patients with multiple sclerosis. *Annals of Behavioral Medicine, 41*, 227–234.
8. Kavia, R. B., De, R. D., Constantinescu, C. S., Stott, C. G., & Fowler, C. J. (2010). Randomized controlled trial of Sativex to treat detrusor overactivity in multiple sclerosis. *Multiple Sclerosis, 16*, 1349–1359.
9. Moller, F., Poettgen, J., Broemel, F., Neuhaus, A., Daumer, M., & Heesen, C. (2011). HAGIL (Hamburg Vigil Study): A randomized placebo-controlled double-blind study with modafinil for treatment of fatigue in patients with multiple sclerosis. *Multiple Sclerosis, 17*, 1002–1009.
10. Freeman, J. A., Thompson, A. J., Fitzpatrick, R., Hutchinson, M., Miltenburger, C., Beckmann, K., et al. (2001). European Study Group on Interferon-beta: Interferon-beta1b in the treatment of

secondary progressive MS: Impact on quality of life. *Neurology, 57*, 1870–1875.

11. Rudick, R. A., Miller, D., Hass, S., Hutchinson, M., Calabresi, P. A., Confavreux, C., et al. (2007). AFFIRM and SENTINEL investigators: Health-related quality of life in multiple sclerosis: Effects of natalizumab. *Annals of Neurology, 62*, 335–346.
12. Fox, R. J., Miller, D. H., Phillips, J. T., Hutchinson, M., Havrdova, E., Kita, M., et al. (2012). Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *New England Journal of Medicine, 367*, 1087–1097.
13. Freedman, M. S., Bar-Or, A., Atkins, H. L., Karussis, D., Frassoni, F., Lazarus, H., et al. (2010). The therapeutic potential of mesenchymal stem cell transplantation as a treatment for multiple sclerosis: Consensus report of the International MSCT Study Group. *Multiple Sclerosis, 16*, 503–510.
14. Zamboni, P., Menegatti, E., Galeotti, R., Malagoni, A. M., Tacconi, G., Dall'Ara, S., et al. (2009). The value of cerebral Doppler venous haemodynamics in the assessment of multiple sclerosis. *Journal of the Neurological Sciences, 282*, 21–27.
15. Al-Omari, M. H., & Rousan, L. A. (2010). Internal jugular vein morphology and hemodynamics in patients with multiple sclerosis. *International Angiology, 29*, 115–120.
16. Baracchini, C., Perini, P., Calabrese, M., Causin, F., Rinaldi, F., & Gallo, P. (2011). No evidence of chronic cerebrospinal venous insufficiency at multiple sclerosis onset. *Annals of Neurology, 69*, 90–99.
17. Centonze, D., Floris, R., Stefanini, M., Rossi, S., Fabiano, S., Castelli, M., et al. (2011). Proposed chronic cerebrospinal venous insufficiency criteria do not predict multiple sclerosis risk or severity. *Annals of Neurology, 70*, 51–58.
18. Zivadinov, R., Marr, K., Cutter, G., Ramanathan, M., Benedict, R. H., Kennedy, C., et al. (2011). Prevalence, sensitivity, and specificity of chronic cerebrospinal venous insufficiency in MS. *Neurology, 77*, 138–144.
19. Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., et al. (2002). Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care, 40*, 113–128.
20. Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal. *J Health Economics, 5*, 1–30.
21. Torrance, G. W., Boyle, M. H., & Horwood, S. P. (1982). Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research, 30*, 1043–1069.
22. Brazier, J., Ratcliffe, J., Salomon, J. A., & Tsuchiya, A. (2007). *Measuring and valuing health benefits for economic evaluation*. New York: Oxford University Press Inc.
23. Mavranezouli, I., Brazier, J. E., Young, T. A., & Barkham, M. (2011). Using Rasch analysis to form plausible health states amenable to valuation: The development of CORE-6D from a measure of common mental health problems (CORE-OM). *Quality of Life Research, 20*, 321–333.
24. Kind, P. (2005). Values and valuation in the measurement of HRQoL. In P. Fayers & D. Hays (Eds.), *Assessing quality of life in clinical trials* (pp. 391–404). New York: Oxford University Press Inc.
25. Feeny, D., Torrance, G. W., & Furlong, W. (1996). Health utilities index. In B. Spilker (Ed.), *Quality of life and pharmaeconomics in clinicals trials* (pp. 239–252). Philadelphia: Lippincott-Raven Publishers.
26. Feeny, D. (2005). Preference-based measures: Utility and quality-adjusted life years. In P. Fayers & D. Hays (Eds.), *Assessing quality of life in clinical trials* (pp. 405–429). New York: Oxford University Press Inc.
27. Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life. *Annals of Internal Medicine, 118*, 622–629.
28. Revicki, D. A., & Kaplan, R. M. (1993). Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Quality of Life Research, 2*, 477–487.
29. Berzon, R., Mauskopf, J. A., & Simeon, G. P. (1996). Choosing a health profile (descriptive) and/or a patient-preference (utility) measure for a clinical trial. In B. Spilker (Ed.), *Quality of life and pharmaeconomics in clinical trials* (pp. 375–379). Philadelphia: Lippincott-Raven Publishers.
30. Feeny, D. H., & Torrance, G. W. (1989). Incorporating utility-based quality-of-life assessment measures in clinical trials. Two examples. *Medical Care, 27*, S190–S204.
31. Torrance, G. W. (1987). Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases, 40*, 593–603.
32. Kind, P., Lafata, J. E., Matuszewski, K., & Raisch, D. (2009). The use of QALYs in clinical and patient decision-making: Issues and prospects. *Value Health, 12*(Suppl 1), S27–S30.
33. Hawthorne, G., & Richardson, J. (2001). Measuring the value of program outcomes: A review of multiattribute utility measures. *Expert Review of Pharmacoeconomics & Outcomes Research, 1*, 215–228.
34. Guyatt, G. H., Veldhuyzen Van Zanten, S. J., Feeny, D. H., & Patrick, D. L. (1989). Measuring quality of life in clinical trials: A taxonomy and review. *Canadian Medical Association Journal, 140*, 1441–1448.
35. Kuspinar, A., & Mayo, N. E. (2013). Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis? *Health and Quality of Life Outcomes, 11*, 71.
36. Hemmett, L., Holmes, J., Barnes, M., & Russell, N. (2004). What drives quality of life in multiple sclerosis? *QJM, 97*, 671–676.
37. Nicholl, C. R., Lincoln, N. B., Francis, V. M., & Stephan, T. F. (2001). Assessing quality of life in people with multiple sclerosis. *Disability and Rehabilitation, 23*, 597–603.
38. Fisk, J. D., Brown, M. G., Sketris, I. S., Metz, L. M., Murray, T. J., & Stadnyk, K. J. (2005). A comparison of health utility measures for the evaluation of multiple sclerosis treatments. *Journal of Neurology, Neurosurgery and Psychiatry, 76*, 58–63.
39. Williams, A. (2005). The EuroQol instrument. In P. Kind, R. Brooks, & R. Rabin (Eds.), *EQ-5D concepts and methods: A developmental history* (pp. 1–17). Dordrecht: Springers.
40. Poissant, L., Mayo, N. E., Wood-Dauphinee, S., & Clarke, A. E. (2003). The development and preliminary validation of a Preference-Based Stroke Index (PBSI). *Health and Quality of Life Outcomes, 1*, 43.
41. Pickard, A. S., Shaw, J. W., Lin, H. W., Trask, P. C., Aaronson, N., Lee, T. A., et al. (2009). A patient-based utility measure of health for clinical trials of cancer therapy based on the European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire. *Value Health, 12*, 977–988.
42. Rowen, D., Brazier, J., Young, T., Gaugris, S., Craig, B. M., King, M. T., et al. (2011). Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *Value Health, 14*, 721–731.
43. Melsop, K. A., Boothroyd, D. B., & Hlatky, M. A. (2003). Quality of life and time trade-off utility measures in patients with coronary artery disease. *American Heart Journal, 145*, 36–41.
44. McKenna, S. P., Ratcliffe, J., Meads, D. M., & Brazier, J. E. (2008). Development and validation of a preference based measure derived from the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses. *Health and Quality of Life Outcomes, 6*, 65.
45. Revicki, D. A., Leidy, N. K., Brennan-Diemer, F., Sorensen, S., & Togias, A. (1998). Integrating patient preferences into health outcomes assessment: The multiattribute Asthma Symptom Utility Index. *Chest, 114*, 998–1007.
46. Revicki, D. A., Leidy, N. K., Brennan-Diemer, F., Thompson, C., & Togias, A. (1998). Development and preliminary validation of

the multiattribute Rhinitis Symptom Utility Index. *Quality of Life Research, 7*, 693–702.

47. Brazier, J., Czoski-Murray, C., Roberts, J., Brown, M., Symonds, T., & Kelleher, C. (2008). Estimation of a preference-based index from a condition-specific measure: The King's Health Questionnaire. *Medical Decision Making, 28*, 113–126.
48. Torrance, G. W., Keresteci, M. A., Casey, R. W., Rosner, A. J., Ryan, N., & Breton, M. C. (2004). Development and initial validation of a new preference-based disease-specific health-related quality of life instrument for erectile function. *Quality of Life Research, 13*, 349–359.
49. Guyatt, G. H., Bombardier, C., & Tugwell, P. X. (1986). Measuring disease-specific quality of life in clinical trials. *Canadian Medical Association Journal, 134*, 889–895.
50. Brazier, J., & Fitzpatrick, R. (2002). Measures of health-related quality of life in an imperfect world: A comment on Dowie. *Health Economics, 11*, 17–19.
51. Patrick, D. L., Burke, L. B., Powers, J. H., Scott, J. A., Rock, E. P., Dawisha, S., et al. (2007). Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health, 10*(Suppl 2), S125–S137.
52. Lin, F. J., Longworth, L., Pickard, A. S. (2013). Evaluation of content on EQ-5D as compared to disease-specific utility measures. *Quality Life Research, 22*(4), 853–874.
53. Mayo, N. (2008). Setting the agenda for multiple sclerosis rehabilitation research. *Multiple Sclerosis, 14*, 1154–1156.
54. Freeman, J. A., Hobart, J. C., Langdon, D. W., & Thompson, A. J. (2000). Clinical appropriateness: A key factor in outcome measure selection: The 36 item short form health survey in multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry, 68*, 150–156.
55. Nortvedt, M. W., Riise, T., Myhr, K. M., & Nyland, H. I. (2000). Performance of the SF-36, SF-12, and RAND-36 summary scales in a multiple sclerosis population. *Medical Care, 38*, 1022–1028.
56. Kikuchi, H., Mifune, N., Niino, M., Ohbu, S., Kira, J., Kohriyama, T., et al. (2011). Impact and characteristics of quality of life in Japanese patients with multiple sclerosis. *Quality of Life Research, 20*, 119–131.
57. Twiss, J., Doward, L. C., McKenna, S. P., & Eckert, B. (2010). Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS). *Health and Quality of Life Outcomes, 8*, 117.
58. Ruta, D. A., Garratt, A. M., & Russell, I. T. (1999). Patient centred assessment of quality of life for patients with four common conditions. *Quality Health Care, 8*, 22–29.
59. Sullivan, M. J., Edgley, K., & Dehoux, E. (1990). A survey of multiple sclerosis: I. Perceived cognitive problems and compensatory strategy use. *Canadian Journal of Rehabilitation, 4*, 99–105.
60. Marrie, R. A., Miller, D. M., Chelune, G. J., & Cohen, J. A. (2003). Validity and reliability of the MSQLI in cognitively impaired patients with multiple sclerosis. *Multiple Sclerosis, 9*, 621–626.
61. Paltamaa, J., West, H., Sarasoja, T., Wikstrom, J., & Malkia, E. (2005). Reliability of physical functioning measures in ambulatory subjects with MS. *Physiotherapy Research International, 10*, 93–109.
62. Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology, 33*, 1444–1452.
63. Ruta, D. A., Garratt, A. M., Leng, M., Russell, I. T., & MacDonald, L. M. (1994). A new approach to the measurement of quality of life. The Patient-Generated Index. *Medical Care, 32*, 1109–1126.
64. Schumacker, R. E. (1996). Editor's note. *Structural Equation Modeling: A Multidisciplinary Journal, 3*, 1–2.
65. Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 3*, 3–24.
66. Chang, C. H. (1996). Finding two dimensions in MMPI-2 depression. *Structural Equation Modeling: A Multidisciplinary Journal, 3*, 41–49.
67. Young, T., Yang, Y., Brazier, J. E., Tsuchiya, A., & Coyne, K. (2009). The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. *Quality of Life Research, 18*, 253–265.
68. Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2004). *Rasch unidimensional measurement models (RUMM) 2020*. Perth, Western Australia: Rumm Laboratory Pty Ltd.
69. Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*, 1–18.
70. Gibbons, C. J., Mills, R. J., Thornton, E. W., Ealing, J., Mitchell, J. D., Shaw, P. J., et al. (2011). Rasch analysis of the hospital anxiety and depression scale (HADS) for use in motor neurone disease. *Health and Quality of Life Outcomes, 9*, 82.
71. Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human science*. New Jersey: Lawrence Erlbaum Associates Inc.
72. Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove: Waveland Press Inc.
73. Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.
74. Broome, H. (2012) *The association between cognition, social functioning, physical impairment, and relationship factors in individuals with multiple sclerosis* (pp. 1–195). The University of Hull.
75. Mead, D. E. (2002). Marital distress, co-occurring depression, and marital therapy: A review. *Journal of Marital and Family Therapy, 28*, 299–314.
76. Lee, E. K. O., & Oh, H. (2012). Marital satisfaction among adults with disabilities in South Korea. *Journal of Disability Studies Policy, 23*, 215–224.
77. Mayo, N. E., Hum, S., & Kuspinar, A. (2013). Methods and measures: What's new for MS? *Multiple Sclerosis, 19*, 709–713.
78. Krupp, L. B., & Pollina, D. A. (1996). Mechanisms and management of fatigue in progressive neurological disorders. *Current Opinion in Neurology, 9*, 456–460.
79. Fisk, J. D., Pontefract, A., Ritvo, P. G., Archibald, C. J., & Murray, T. J. (1994). The impact of fatigue on patients with multiple sclerosis. *Canadian Journal of Neurological Sciences, 21*, 9–14.
80. Freal, J. E., Kraft, G. H., & Coryell, J. K. (1984). Symptomatic fatigue in multiple sclerosis. *Archives of Physical Medicine and Rehabilitation, 65*, 135–138.
81. Murray, T. J. (1985). Amantadine therapy for fatigue in multiple sclerosis. *Canadian Journal of Neurological Sciences, 12*, 251–254.
82. Deloire, M. S., Bonnet, M. C., Salort, E., Arimone, Y., Boudineau, M., Petry, K. G., et al. (2006). How to detect cognitive dysfunction at early stages of multiple sclerosis? *Multiple Sclerosis, 12*, 445–452.
83. Peyser, J. M., Edwards, K. R., Poser, C. M., & Filskov, S. B. (1980). Cognitive function in patients with multiple sclerosis. *Archives of Neurology, 37*, 577–579.
84. Brazier, J., Akehurst, R., Brennan, A., Dolan, P., Claxton, K., McCabe, C., et al. (2005). Should patients have a greater role in valuing health states? *Applied Health Economics and Health Policy, 4*, 201–208.
85. Ubel, P. A., Loewenstein, G., & Jepson, C. (2003). Whose quality of life? A commentary exploring discrepancies between health

state evaluations of patients and the general public. *Quality of Life Research, 12*, 599–607.

86. Whitehead, S. J., & Ali, S. (2010). Health outcomes in economic evaluation: The QALY and utilities. *British Medical Bulletin, 96*, 5–21.
87. Stallard, J., & Major, R. E. (1995). The influence of orthosis stiffness on paraplegic ambulation and its implications for functional electrical stimulation (FES) walking systems. *Prosthetics and Orthotics International, 19*, 108–114.
88. Brissot, R., Gallien, P., Le Bot, M. P., Beaubras, A., Laisne, D., Beillot, J., et al. (2000). Clinical experience with functional electrical stimulation-assisted gait with Parastep in spinal cord-injured patients. *Spine (Phila Pa 1976), 25*, 501–508.
89. Thoumie, P., Perrouin-Verbe, B., Le, C. G., Bedoiseau, M., Busnel, M., Cormerais, A., et al. (1995). Restoration of functional gait in paraplegic patients with the RGO-II hybrid orthosis. A multicentre controlled study. I. Clinical evaluation. *Paraplegia, 33*, 647–653.
90. Rothman, M., Burke, L., Erickson, P., Leidy, N. K., Patrick, D. L., & Petrie, C. D. (2009). Use of existing patient-reported outcome (PRO) instruments and their modification: The ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO task force report. *Value Health, 12*, 1075–1083.
91. Food, U. S. (2009). Drug administration: Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register, 74*, 65132–65133.

CHAPTER 10: Integration of Manuscripts 4 and 5

Research questions of Manuscripts 4 and 5

Manuscript 4:

Using existing data to identify candidate items for a health state classification system in multiple sclerosis.

Manuscript 5:

The development of a bilingual MS-specific health classification system: the Preference-Based Multiple Sclerosis Index (PBMSI).

Integration of Manuscripts 4 and 5

Manuscript 4 involved the development of a prototype classification system for the PBMSI. In the study, we identified items best reflecting the domains of quality of life important to people with MS, and provided evidence for the discriminative capacity of the response options by cross walking onto a visual analogue scale (VAS) of health rating. Five items were selected for inclusion in the PBMSI.

These 5 items came from various existing questionnaires. As a result, each one had a different recall period, set of instructions and response options. The next study describes the qualitative review process undertaken to revise these items, in English and French, based on two key sources: expert opinion and patients. At the expert level, items were revised to make them more uniform with regard to their instructions and response options. At the patient level, cognitive interviews were conducted to assess readability and comprehension in English and French.

CHAPTER 11 (MANUSCRIPT 5)

The development of a bilingual MS-specific health classification system: the Preference-Based Multiple Sclerosis Index (PBMSI)

Ayse Kuspinar[1], Vanessa Bouchard[1], and Nancy E. Mayo[1,2]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada

[2]Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

Submitted to *International Journal of MS Care*

Communication addressed to:

Ayse Kuspinar, M.Sc., Ph.D. Candidate
School of Physical & Occupational Therapy
Faculty of Medicine, McGillUniversity
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
Tel: 514-934-1934 ext 31564
E-mail: ayse.kuspinar@mail.mcgill.ca

ABSTRACT

Objective: The US Food and Drug Administration's (FDA) guidelines for the development of patient reported outcomes requires patient input in the development of self-reported assessments. Conducting cognitive interviews with patients are important when developing questionnaires in order to increase the accuracy of reporting and minimize measurement error. The objective of this study was to perform qualitative review, in English and French, of items in the Preference-Based MS Index (PBMSI) using expert and patient feedback.

Methods: Cognitive interviews were conducted with MS patients in both English and French. The verbal-probing method was used to conduct the interviews. For each PBMSI item, the interviewer probed for specific information on what types of difficulty the participant had with the item and the basis for their response for each item. Furthermore, the respondent was asked to provide information on the clarity of the item, the meaning of the item, the appropriateness of the response options and the recall time period. To minimize respondent burden, each participant was interviewed on 2 to 3 items only. All interviews were recorded with a digital voice recorder and transcribed onto a computer.

Results: Cognitive interviews were performed on 22 patients with MS. Each interview took about 30 minutes to complete. The average age of the sample was 52 years (range 29 to 88) and 82% were women. The average number of years since diagnosis was 12, and the highest level of education completed was university or college for 86% of the sample. During the cognitive interview process, modifications were made to each item, in terms of recall period, instructions and phrasing.

Conclusion: The process of qualitative review was an important and necessary step to produce the best items for use in the PBMSI. Patient feedback allowed us to clarify items, simplify language and make the items more uniform in terms of their instructions and response options. In the future, this will not only help minimize unnecessary cognitive burden on patients when filling out the questionnaire, but will also increase the accuracy of reporting and reduce measurement error.

INTRODUCTION

The US Food and Drug Administration's (FDA) guidelines for the development of patient reported outcomes requires patient input in the development of self-reported assessments.[1] Conducting cognitive interviews with patients are important when developing questionnaires in order to help reduce respondent burden and minimize measurement error.

Preference-based measures are patient-reported outcomes of health-related quality of life (HRQL) that are commonly used for economic evaluation in health care.[2;3] Preference-based measures can often generate hundreds and thousands of health states. The most commonly used preference-based measure is the EuroQol-5D (EQ-5D), which was developed by a team of researchers in Europe.[4;5] The EQ-5D consists of 5 items: mobility, self-care, usual activities, pain and anxiety/depression. Each item has 3 response options, providing a total of 243 (3^5) unique health states. The EQ-5D is self-administered and takes 1-2 minutes to complete.[2]

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system (CNS) that can produce a range of symptoms, such as muscle weakness, fatigue and cognitive impairment.[6] In MS, the use of preference-based measures is limited to generic measures like the EQ-5D. However, the challenge with using such generic preference-based measures in MS is that these measures may not capture all domains of health relevant to the disease. Our previous work has shown that there are limitations with the use of these measures in MS with regards to content and construct validity.[7;8] Therefore, a MS specific preference-based measure may be more appropriate for use in economic evaluation of treatments involving MS.

In a previous study,[9] we identified 5 items that were most important to the quality of life of people with MS: walking, fatigue, mood, cognition and work. These 5 items came from various existing questionnaires.[9] As a result, each one of these items had different recall periods, instructions, and response options. This study describes the qualitative review process undertaken to revise these items, in English and French, based on two key sources: expert opinion and patients. This qualitative process will not only ensure that the items are comprehended and interpreted as intended by patients, but will also make the items more uniform with regard to their instructions and response options.

Therefore, the global aim of this study was to contribute to the development of the Preference-Based MS Index (PBMSI). The specific objective of this foundational work was to qualitatively review the 5 items selected for inclusion in the PBMSI, using expert and patient feedback.

METHODS

Domain generation and item selection

The methods for domain generation and item selection for the PBMSI have been reported previously.[7;9] Briefly, the *domains* for the PBMSI were created based on semi-structured interviews with 185 patients with MS. Patients were asked to identify the most important aspects of their lives that were affected by MS.

These same patients were also asked to complete a comprehensive questionnaire package consisting of over 200 items, which came from existing patient-reported outcomes or were created from scratch by a multi-disciplinary team of clinicians and researchers. The items for the PBMSI came from this questionnaire package. Modern methods of measurement (i.e. Rasch analysis) were used to select one item per domain.[9]

Revision and rewriting of items

The items selected for inclusion in the PBMSI had different phrasing styles and recall periods. Due to these variations, the items needed to be rewritten for uniformity and coherence. As presented in Figure 1, item revision was conducted in 2 phases. The first phase involved item revision and rewriting by experts simultaneously in English and French. In the second phase, the items were cognitively debriefed with 22 MS patients, 14 in English and 8 in French. During the cognitive interview process, each item went through several iterations before being accepted as the final version to include in the PBMSI.

Phase 1: Focus group with experts

A focus group was conducted with experts in the field of MS to rewrite items that would convey the same information in English and French. Experts were recruited from the 4 major hospitals in Montreal, Canada (Royal Victoria Hospital, Montreal General Hospital, Montreal Neurological Institute and Notre Dame Hospital).

The focus group was conducted in a round table format, where participants were paired up with the person sitting next to them. Each pair was given a copy of the PBMSI items at the start of the session, and was asked to discuss each item in terms of the following four points: (i) Is the wording clear and appropriate for the item? If no, how would you change it? (ii) Are the response options clear and appropriate for the item? If no, how would you change them? (ii) How difficult would it be for patients to answer the question? And (iv) Do you have any suggestions to improve the item? While the items were being rewritten in English, wording in French was suggested by two French speaking researchers and problematic wording addressed in both language. The end product was a set of items with parallel English and French wording.

Phase 2: Cognitive interviews with patients

Recruitment of patients

Participants were recruited through advertising on the MS Society of Canada's website, during the 2012 Quebec Summit on Multiple Sclerosis, and through flyers placed at the outpatient MS Clinic of the Montreal Neurological Hospital. Interested participants contacted the study coordinator (AK) by email or telephone, and the study coordinator sent the consent form to be signed and returned. Patients were eligible to participate in the cognitive interview if: (1) they were diagnosed with MS, (2) were at least 18 years of age, and (3) were able to speak and read English or French.

Cognitive interviewing process

Interviews were conducted by 2 physiotherapists, who were also doctoral candidates. Each interview took about 30 minutes to complete. One physiotherapist conducted the English interviews while the other conducted the French interviews. All interviews were carried out by telephone.

Prior to the phone interview, participants were sent a questionnaire package with basic socio-demographic questions, the PBMSI questions and a visual analogue scale (VAS) of their health state today. They were permitted to look at the package beforehand and were asked to have it on hand for the interview. During the interview, respondents were first asked to provide their answers to the socio-demographic questions and the PBMSI items. Following this, the cognitive interview process for each item began.

The verbal probing method was employed, as it is known to help facilitate the interview process and place less burden on the respondent. As shown in Table 1, for each PBMSI item, the interviewer probed for specific information on what types of difficulty the participant had with the item and the basis for their response for each item. Furthermore, the respondent was asked to provide information on the clarity of the item, the meaning of the item, the appropriateness of the response options and the recall time period.

To minimize respondent burden, each participant was interviewed on 2 to 3 items only. All interviews were recorded with a digital voice recorder and transcribed onto a computer.

Once all of the items were endorsed or finalized in English, cognitive interviews were also performed on the French items. The same format and type of questions that were used during the English interviews were also used for the French ones. The French speaking interviewer asked the respondent about the meaning of specific words in the item, the overall meaning of the item, and why they had chosen a specific response option. For some items, the respondent was also asked to consider alternative wording for those items. On the basis of the cognitive interviews, some revisions were made to the original translations.

Analysis of cognitive interview data

After each interview, the interviewer reviewed the comments to determine issues with recall period, comprehension, clarity and response options. If an item was found to be problematic during the interview, it was revised based on the respondents' suggestions and then tested on the next respondent. When at least 3 respondents in a row stated that they had no problems with an item, the item was accepted as the final version.

RESULTS

The focus group consisted of a total of 24 clinicians and researchers. The group included a neurologist, a clinical psychologist, a neuro-psychologist, an epidemiologist, eleven physiotherapists, three occupational therapists, one nurse, and five graduate students. All participants had experience working with MS patients or other neurological conditions such as stroke.

During the focus group, it was decided that similar to the EQ-5D, the recall period would be based on the patient's 'health state today'. Therefore, statements such as 'past 4 weeks' were removed from the items. Item #5 on '*ability to work'* was revised to '*roles and responsibilities (work, family or household)*' to include patients who did not work, but carried out work-related activities such as household chores. Response options were also simplified and unnecessary wording was removed to reduce cognitive burden on patients.

Once the items were reviewed and finalized among experts, they were then taken to patients for cognitive interviewing. Table 2 presents the demographic and clinical characteristics of the patients who participated in the cognitive interview. There were 14 participants who underwent cognitive interviewing in English, and 8 who underwent cognitive interviewing in French. The English cognitive interview participants, compared to the French cognitive interview participants, were slightly older and consisted of a greater proportion of men. However, the mean number of years since diagnosis was the same for both groups (11 years).

Table 3 presents the step-by-step changes that were made to each item in English during the cognitive interview process. The items underwent several iterations: walking had 5 iterations, fatigue had 7 iterations, mood had 4 iterations, cognition had 4 iterations and roles and responsibilities had 3 iterations. The changes are explained in detail below.

Walking: The item on walking was revised to include people with high levels of physical function (i.e. individuals who could walk briskly for recreation or sports). Furthermore, certain words such as 'community' were removed because patients found them to be too vague or ambiguous.

Fatigue: In the original version, fatigue was described as 'exhausted'. However, patients found this to be a 'heavy word'. In fact, one patient stated that if fatigue were on a scale from zero to ten, where zero was fatigue, exhaustion would be a ten. Therefore, as per patients' suggestions, the word 'exhaustion' was removed from the item. When patients were asked, how would they describe MS related fatigue? They expressed that the need to rest should be incorporated into the item. Therefore, the response options were revised to '*I never felt so tired I had to rest…I felt so tired I had to rest one or more times throughout the day…I felt so tired I had to rest most of the day.*'

Mood: A small yet important modification was made to the mood item as a result of feedback from patients. The original response levels for this item were '*I do not feel sad… I feel somewhat sad…I feel very sad*'. Patients reported that the word 'depressed' should be incorporated into the response options, as it was not clear that the question was referring to depression. Therefore, the response levels were revised to '*I do not feel sad or depressed… I feel somewhat sad or depressed…I feel very sad or depressed*'.

Cognition: The aspect of cognition assessed in the PBMSI was on decision making (e.g. planning your day, planning meals etc.). However, when patients were interviewed on this item they reported to have no problems with decision making. Instead, patients stated that, rather than decision making, concentration was an area of cognition that was a major concern for them. As a result of this feedback, the cognition item was changed to 'concentration' and was phrased to '*did you have trouble concentrating in the past week (on things like conversations, books, movies or daily routines)?*'

Roles & Responsibilities: Very minor changes were made to the response levels of this item. Generally, patients stated that roles and responsibilities as described by '*ability to do the things you needed to do at work, at home, and to take care of yourself and your family'* was clear and easy to comprehend.

Recall period: As MS has an unpredictable course and symptoms can change from day to day, patients reported that '*today*' was not an accurate representation of their symptoms. Patients stated that '*over the past week'* was an appropriate time frame, as it was more representative of their experience and easy to recall. Patients stated that the recall period '*over the past month*' was difficult to remember.

Table 4 presents the step-by-step changes that were made to each item in French during the cognitive interview process. The walking item underwent 3 iterations, fatigue underwent 4 iterations, mood underwent 2 iterations, cognition underwent 1 iteration, and roles and responsibilities underwent 2 iterations. Examples of changes include 'la plupart du temps' being revised to 'le plus souvent', and 'que j'ai eu' being revised to 'au point où j'ai eu'.

Table 5 presents a summary of the items: (i) in their original version, (ii) after being rewritten by experts in the focus group, and (iii) at the end of the cognitive interviews. Table 6 presents the same items in French.

The PBMSI Questionnaire

A copy of the PBMSI questionnaire (in English and French) can be found at the end of the paper.

DISCUSSION

This study described the simultaneous development of a bilingual MS-specific health classification system, the PBMSI. Experts in the field of MS were brought together in a focus group to rewrite items simultaneously in English and French. The purpose of the focus group was to clarify confusing items, to simplify language, and to ensure that there was consistency in the style of the questions and response options. Forward-backward translation of the items was not necessary, as items were developed simultaneously in both languages at the expert level. Later, cognitive interviews were conducted with 14 English speaking and 8 French speaking patients. Based on patient feedback, revisions were made to each of the items in terms of content, instructions and phrasing.

Two well-known methods of developing questionnaires in multiple languages are (i) sequential and (ii) simultaneous.[10] In the sequential approach items are developed in only one language (the source language) with subsequent translation into the target languages using a forward and backward translation process. In the simultaneous approach, native speakers from each language develop items simultaneously. The PBMSI items were developed at the expert level (i.e. focus group) using the latter approach. The advantage of the simultaneous method, compared to the sequential one, is that any problematic wording and discrepancies between the language versions are resolved during the item generation process.[10] Following item writing at the expert level, we conducted cognitive interviews with patients to ensure that there was semantic and conceptual equivalence between languages. We assessed whether patients understood the questions the same way in both English and French.

Our study's sample size was similar to other studies that involved cognitive interviews to develop questionnaires. Our sample size of 22 patients was sufficient and within the recommended range in the literature. Willis[11] recommended that samples of 5 to 15 individuals were sufficient when revising questionnaire items. Also, Sheatsley[12] suggested that it usually takes no more than 12 to 25 interviews to reveal major flaws in a questionnaire.

Furthermore, the method we used to conduct the cognitive interviews, verbal probing, is a well-established and accepted methodology.[13] The use of probing helps guide the respondent and shapes the interchange in a way that is controlled mainly by the interviewer. The advantage of this methodology is that it helps avoid irrelevant or unnecessary discussion during the interview, and helps the interviewer to concentrate on areas that appear to be important sources of error.[11;13] The alternative method, which is the think-aloud method, also has its own advantages. For example, minimal interviewer training is required as the interviewer is required to mainly listen to the respondent talk. Furthermore, because minimal guidance is provided, the respondent or patient may provide information that is unanticipated by the interviewer. However, the disadvantage of the think-aloud method is that all respondents may not be outgoing and elaborate very much on a question. Also, this method places a significant amount of burden on the respondent, and may result in the individual wandering off-track and delving into unrelated topics.[11;13]

We were sensitive to avoid wording that could be subjected to response shift. Response shift is defined as a change in one's evaluation of a target construct (i.e. fatigue) as a result of a change in the respondent's internal standards of measurement, values and conceptualization of the target construct[14]. "Difficulty" is a word that has been flagged as a potential source for response shift, as patients may recalibrate how they interpret what difficulty means to them over time[15]. In the PBMSI, the only item that would be close to being subjected to response shift would be concentration, which used the word "trouble". However, in the context of this item, the word "trouble" was used as a noun, and not as an adverb to describe difficulty.

The choice of recall period can depend on the disease or the condition's characteristics.[1] In this study, based on feedback from patients, a 7-day recall period was used for the PBMSI items. As MS has an unpredictable disease course and symptoms can vary from day to day, a recall period using the 'past week' was found to be most appropriate. Asking patients to answer a question based on their health state 'today' would not be an accurate representation of their experiences. As one patient pointed out, symptoms such as fatigue, can vary not only form one day to the next, but can also vary within a single day (i.e. morning to afternoon). Also, to avoid having patients average their responses over the past week, we asked patients to select a response based on the state that they were *most often* in the past week. For example, the response options for the question ‘Describe your fatigue in the past week’ were ‘*Most often*…(i) I never felt so tired I had to rest, (ii) I felt so tired I had to rest one or more times throughout the day, (iii) I felt so tired I had to rest most of the day’. A time frame of ‘in the past month’ was disapproved by patients, as it was a long period of time to remember, and was likely to be influenced by their state at the time of recall.

A strength of this study was that the items went through several processes of review to ensure that they were clear and easy for patients to understand. Furthermore, our sample of MS patients were not only of a sufficient number, but were also representative of various age groups and disease characteristics (i.e. number of years since diagnosis ranged from 1 to 38 years).

A limitation of this study was that all of the items were changed from their original format, and as a result the items may function differently. However, these changes were carried out to make the items more uniform and easy for interpretation by patients. We believe that the methods of qualitative review conducted in this study did not worsen the items, but rather improved them in terms of phrasing and clarity.

The process of qualitative review was an important and necessary step to produce the best items for use in the PBMSI. Item writing by experts and cognitive interviews with patients allowed us to clarify items, simplify language and make the items more uniform in terms of their instructions and response options. This method in the future will not only help minimize unnecessary cognitive burden on patients when filling out the questionnaire, but will also increase the accuracy of reporting. The next step in the development of the PBMSI will be to elicit patient preferences for each of the items using standard valuation methods and to calculate a scoring algorithm for the index.

REFERENCE LIST

(1) Food and Drug Administration. Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register* 2009;74:65132-65133.

(2) Feeny D. Preference-based measures: utility and quality-adjusted life years. *Assessing quality of life in clinical trials* 2005;405-429.

(3) Feeny D, Furlong W, Torrance GW et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care* 2002;40:113-128.

(4) Kind P, Brooks R, Rabin R. *EQ-5D concepts and methods: a developmental history*. Springer, 2006.

(5) Kind P. The EuroQoL instrument: an index of health-related quality of life. *Quality of life and pharmacoeconomics in clinical trials* 1996;2:191-201.

(6) Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG. Multiple Sclerosis. *N Engl J Med* 2000;343:938-952.

(7) Kuspinar A, Mayo NE. Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis? *Health Qual Life Outcomes* 2013;11:71.

(8) Kuspinar A, Mayo NE. A review of the psychometric properties of generic utility measures in multiple sclerosis. *Pharmacoeconomics* 2014;32:759-773.

(9) Kuspinar A, Finch L, Pickard S, Mayo NE. Using existing data to identify candidate items for a health state classification system in multiple sclerosis. *Qual Life Res* 2014;23:1445-1457.

(10) Marquis P, Keininger D, Acquadro C, de la Loge C. Translating and evaluating questionnaires: cultural issues for international research. *Assessing quality of life in clinical trials* 2005;77-93.

(11) Willis GB. *Cognitive interviewing: A tool for improving questionnaire design*. Sage Thousand Oaks, CA, 2005.

(12) Sheatsley PB. Questionnaire construction and item writing. 1983. Handbook of Survey Research, PH Rossi, JD Wright and AB Anderson, eds. New York: Academic Press.

(13) Collins D. Pretesting survey instruments: an overview of cognitive methods. *Qual Life Res* 2003;12:229-238.

(14) Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med* 1999;48:1531-1548.

(15) Barclay-Goddard R, Lix LM, Tate R, Weinberg L, Mayo NE. Health-related quality of life after stroke: does response shift occur in self-perceived physical function? *Arch Phys Med Rehabil* 2011;92:1762-1769.

Figure 1 A summary of the simultaneous development of the PBMSI items in English and French

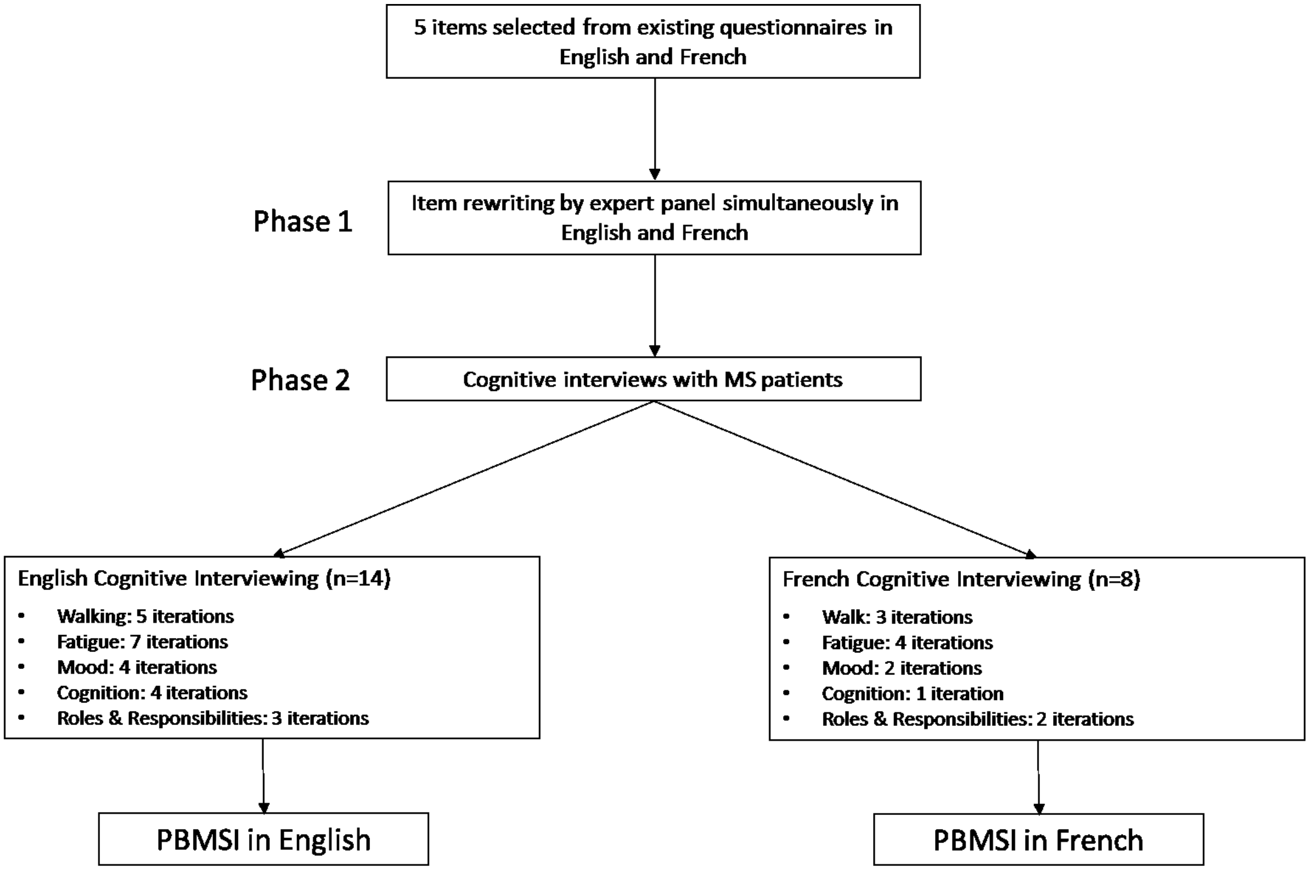

5 items selected from existing questionnaires in English and French
Phase 1
Item rewriting by expert panel simultaneously in English and French
Phase 2
Cognitive interviews with MS patients
English Cognitive Interviewing (n=14)
Walking: 5 iterations
Fatigue: 7 iterations
Mood: 4 iterations
Cognition: 4 iterations
Roles & Responsibilities: 3 iterations
French Cognitive Interviewing (n=8)
Walk: 3 iterations
Fatigue: 4 iterations
Mood: 2 iterations
Cognition: 1 iteration
Roles & Responsibilities: 2 iterations
PBMSI in English
PBMSI in French

Table 1 Cognitive interview questions

Recall period
What do you think about the recall time period?
Which time frame is most representative of your health “today, past week, or past month?”
Why did you choose that time frame?
Items
What does this question mean to you?
In your own words, what do you think this question is asking?
Was this question easy to understand?
Are there any words in this question that are not clear?
How would you change the wording to make it clearer?
Response choices
What do you think about the response options?
Are there any words in the response choices that are not clear?
How would you make the response choices clearer?
Overall impression of the questionnaire
Do you have any comments on the questionnaire as whole?
Is there anything that you would change in the questionnaire?

Table 2 Demographic and clinical characteristics of cognitive interview patients

Characteristics	Cognitive interviewing in English (n=14)	Cognitive interviewing in French (n=8)
	Mean (SD) or N (%)	
Age (y)	53.7 (9.7)	48.4 (17.5)
Women / Men	11 / 3 (79 / 21)	7/ 1 (88 / 12)
Year since diagnosis	11.9 (10.3)	11.3 (6.4)
University/College/High School	11 / 2 / 1 (79 / 14 / 7)	5 / 1 / 2 (63 / 13 / 25)
EQVAS (0 to 100)	63.6 (15.2)	76.9 (10.7)

EQVAS, EuroQoL Visual Analogue Scale of health state today.

Table 3 Development of the items in *English* during cognitive interviewing process

Item	**Version (problem identified with the item)**	**Patient Number**													
1	***WALKING***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	V1 (problem identified with word ‘today’)		x												
	V2 (‘today’ changed to ‘past week’. Problem with ‘community’ and ‘difficulty walking outside’							x							
	V3 (‘community’ changed to ‘walk for recreation or sports’. ‘Difficulty walking outside changed to ‘neighborhood, shopping mall or public building’.)								√	√					
	V4 (Word ‘neighborhood’ not clear.)											x			
	V5 (Second response level changed to ‘walk to accomplish the tasks I needed to do during the day (to and from transportation, public building or within work environment)												√	√	√
2	***FATIGUE***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	V1 (problem identified with the word ‘today’)	x	x												
	V2 (‘today changed to ‘past week’)			√											
	V3 (problem with the word ‘different times throughout the day’.)					x									
	V4 (‘different times throughout the day’ changed to one or more times throughout the day’.)						x								
	V5 (Problem with word ‘exhausted’)							√	x						
	V6 (‘exhausted changed to ‘tired’ but ‘tired’ alone does not describe MS fatigue)									x					
	V7 (‘tired’ changed to ‘I was so tired I needed to rest’)										√	√	√	√	√

3	***MOOD***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	V1 (problem identified with the word ‘today’)	x													
	V2 (‘today changed to ‘past week’			√		√									
	V2 (the word ‘sad’ alone does not describe mood)						x								
	V4 (response options changed to ‘sad or depressed’)											√	√	√	
4	***COGNITION***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	V1 (stem and response options are clear)	√	√												
	V2 (‘today’ changed to ‘past week’ for uniformity with other items)				√	√									
	V3 (decision making is not a problem, but concentration is)							x	x	x					
	V4 (item revised to assess concentration)										√	√	√		
5	***ROLES & RESPNSIBILITIES***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	V1 (problem identified with the word ‘today’)				x										
	V2 (problem with first response option ‘all of the things I needed to do’)						x								
	V3 (first response option modified to ‘all or most of the things I needed to do’)										√	√	√		

V, version; x, problem identified with item; √, no problems identified with item.

Table 4 Development of items in *French* throughout the cognitive interview process

Item	**Version (problem identified with the item)**	**Patient Number**							
1	***WALKING***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	V1 (translation for first response option not clear)	x							
	V2 (first response option changed to 'j'ai pu faire de la marche comme activité ou sport')		√	x					
	V3 (first response option clarified to 'J'ai pu faire de la marche *rapide* comme loisir ou sport')				√	√		√	√
2	***FATIGUE***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	V1 (problem with 'toute la journée')	x							
	V2 ('toute la journée' changed to 'une grande partie de la journée')		√						
	V3 (problem with 'que j'ai eu')					x			
	V4 ('que j'ai eu' changed to 'au point où j'ai eu')						√	√	√
3	***MOOD***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	V1 (problem with' 'la plupart du temps')	x							
	V2 ('la plupart du temps' changed to 'le plus souvent')		√			√	√	√	√
4	***COGNITION***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	V1 (no problems with item)		√		√	√	√		
5	***ROLES & RESPONSIBILITIES***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	V1 (problem with 'devais')			x					
	V2 ('devais' changed to 'fallait')				√		√	√	√

V, version; x, problem identified with item; √, no problems identified with item.

Table 5 PBMSI items in English after focus group with experts (Phase 1) and after cognitive interviews with patients (Phase 2).

Item	Initial Version	Focus group with experts (*Phase 1*)	Cognitive interview with patients (*Phase 2*)
1	***WALKING***		
	How would you best describe your ability to walk with or without a walking aid? □ I am able to walk in the community, as I need to □ I am able to walk inside the house, but I have difficulty walking alone outside □ I am able to walk only a few steps or I use a wheelchair	**Describe your ability to walk today.** □ I am able to walk in the community without using a walking aid □ I am able to walk inside the house, but I have difficulty walking outside without a walking aid □ I am able to walk only a few steps or I use a wheelchair	**Describe your ability to walk in the past week.** **Most often:** □ I could walk briskly for recreation or sports □ I could walk to accomplish the tasks I needed to do during the day (to and from transportation, public building or within work environment) □ I could walk only a few steps or I always used a wheelchair
2	***FATIGUE***		
	During the past week, I felt exhausted. □ None of the time or rarely or a little of the time (0 to 2 days) □ Occasionally or a moderate amount of the time (3 to 4 days) □ Most or all of the time (5 to 7 days)	**Describe your fatigue today.** □ I rarely feel exhausted □ I feel exhausted at different times throughout the day □ I feel exhausted all day long	**Describe your fatigue in the past week.** **Most often:** □ I never felt so tired I had to rest □ I felt so tired I had to rest one or more times throughout the day □ I felt so tired I had to rest most of the day
3	***MOOD***		
	How often did the following statement apply to you during the past 4 weeks? **I felt sad:** □ Not at all or a little bit □ Somewhat □ Quite a bit or very much	**Describe your mood today.** □ I do not feel sad □ I feel somewhat sad □ I feel very sad	**Describe your mood in the past week.** **Most often:** □ I did not feel sad or depressed □ I felt somewhat sad or depressed □ I felt very sad or depressed

4 *COGNITION*

During the past 4 weeks, how often did you have trouble making decisions?

- □ Never or rarely
- □ Sometimes or often
- □ Almost always

Describe your ability to make everyday decisions (like planning your day, planning meals etc.).

- □ I never or rarely have trouble
- □ I have trouble some of the time
- □ I have trouble most of the time

Did you have trouble concentrating in the past week (on things like conversations, books, movies or daily routines)?

Most often:

- □ I never or rarely have trouble
- □ I had trouble some of the time
- □ I had trouble most of the time

5 *ROLES & RESPONSIBILITIES*

How would you best describe your ability to accomplish work or any other activities?

- □ I can work or perform activities as I used to
- □ I do not always perform my work or activities as I used
- □ I can no longer work or perform activities as I used to

Describe your ability to do the things you need to do at work, at home, and to take care of yourself and your family today.

- □ I can do the things I need to do
- □ I can do some of the things I need to do
- □ I can no longer do the things I need to do

Describe your ability to do the things you needed to do at work, at home, and to take care of yourself and your family in the past week.

Most often:

- □ I could do all or most of the things I needed to do
- □ I could do some of the things I needed to do
- □ I could not do the things I needed to do

Table 6 PBMSI items in French after focus group with experts (Phase 1) and after cognitive interviews with patients (Phase 2).

Item	**Initial Version**	**Focus group with experts (*Phase 2*)**	**Cognitive interview with patients (*Phase 3*)**
1	***WALKING***		
	Comment décririez-vous votre capacité à marcher avec ou sans aide.	**Décrivez votre capacité à marcher aujourd'hui**	**Décrivez votre capacité à marcher au cours de la dernière semaine.** **Le plus souvent:**
	❑ Je peux marcher à l'extérieur autant que je veux. ❑ Je peux marcher chez moi, mais j'ai de la difficulté à marcher seul à l'extérieur. ❑ Je peux faire seulement quelques pas, ou j'utilise un fauteuil roulant.	❑ Je peux marcher dans la communauté sans utiliser d'aide à la marche ❑ Je peux marcher dans la maison, mais j'ai de la difficulté à marcher à l'extérieur sans aide à la marche ❑ Je peux marcher seulement quelques pas ou j'utilise un fauteuil roulant	❑ J'ai pu faire de la marche rapide comme loisir ou sport ❑ J'ai pu marcher pour accomplir les tâches que j'avais à faire dans la journée (pour vous rendre à un transport, un endroit public ou à votre travail) ❑ J'ai pu marcher seulement quelques pas ou j'utilise un fauteuil roulant
2	***FATIGUE***		
	Durant la dernière semaine, je me sentais épuisé.	**Décrivez votre fatigue aujourd'hui**	**Décrivez votre fatigue au cours de la dernière semaine.** **Le plus souvent:**
	❑Rarement ou jamais quelques fois (1 à 2 jours) ❑Souvent ou plusieurs fois (3 à 4 jours) ❑Majorité du temps ou tout le temps (5 à 7 jours)	❑ Je me sens rarement épuisé ❑ Je me sens épuisé à différents moments pendant la journée ❑ Je me sens épuisé toute la journée	❑ Je ne me suis jamais senti fatigué au point où j'ai eu à me reposer. ❑ Je me suis senti fatigué au point où j'ai eu à me reposer une ou quelques fois pendant la journée ❑ Je me suis senti fatigué au point où j'ai eu à me reposer une grande partie de la journée

3	*MOOD*		
	Dites nous à quelle fréquence cet énoncé a été appliqué à votre situation au cours des quatre dernières semaines. Je me sentais triste. □Pas du tout ou un peu □Moyennement □Souvent ou beaucoup	**Décrivez votre humeur aujourd'hui.** □ Je ne me sens pas triste □ Je me sens un peu triste □ Je me sens très triste	**Décrivez votre humeur au cours de la dernière semaine.** **Le plus souvent:** □ Je ne me suis pas senti triste ou déprimé □ Je me suis senti un peu triste ou déprimé □ Je me suis senti très triste ou déprimé
4	*COGNITION*		
	Durant les 4 dernières semaines avez-vous souvent eu de la difficulté à prendre des décisions? □ Jamais ou rarement □ Quelques fois ou souvent □ Pratiquement toujours	**Décrivez votre capacité à prendre des décisions de tous les jours (comme planifier votre journée, planifier les repas etc.)** □ Je n'ai jamais ou rarement de difficulté □ J'ai quelques fois de la difficulté □ J'ai presque toujours de la difficulté	**Avez-vous eu des problèmes à vous concentrer au cours de la dernière semaine (en suivant une conversation, lisant un livre, regardant un film ou en complétant votre routine quotidienne)?** **Le plus souvent:** □ Je n'ai jamais ou rarement eu de difficulté □ J'ai quelques fois eu de la difficulté □ J'ai presque toujours eu de la difficulté
5	*ROLES & RESPONSIBILITIES*		
	Comment décririez-vous votre capacité à accomplir votre travail ou toutes autres activités. □ Je peux faire mon travail/mes activités comme avant. □ Je ne peux pas toujours faire mon travail/mes activités comme avant. □ Je ne peux plus faire mon travail/mes activités comme avant.	**Décrivez votre capacité à accomplir les choses que vous devez faire au travail, à la maison, et pour prendre soin de vous et de votre famille aujourd'hui.** □ Je peux faire les choses que je dois faire □ Je peux faire quelques-unes des choses que je dois faire □ Je ne peux plus faire les choses que je dois faire	**Décrivez votre capacité à accomplir les choses que vous devez faire au travail, à la maison, et pour prendre soin de vous et de votre famille au cours de la dernière semaine.** **Le plus souvent:** □ J'ai pu faire toutes ou la plupart des choses qu'il fallait que je fasse □ J'ai pu faire quelques-unes des choses qu'il fallait que je fasse □ Je n'ai pas pu faire les choses qu'il fallait que je fasse

PREFERENCE-BASED MULTIPLE SCLEROSIS INDEX

For each of the items listed below, choose the option you were most often in, over the past week.

1) Walking

Describe your ability to walk in the past week.

Most often:

□ I could walk briskly for recreation or sports

□ I could walk to accomplish the tasks I needed to do during the day (to and from transportation, public building or within work environment)

□ I could walk only a few steps or I always used a wheelchair

2) Fatigue

Describe your fatigue in the past week.

Most often:

□ I never felt so tired that I had to rest

□ I felt so tired that I had to rest one or more times throughout the day

□ I felt so tired that I had to rest most of the day

3) Mood

Describe your mood in the past week.

Most often:

□ I did not feel sad or depressed

□ I felt somewhat sad or depressed

□ I felt very sad or depressed

PREFERENCE-BASED MULTIPLE SCLEROSIS INDEX

For each of the items listed below, choose the option you were most often in, over the past week.

4) **Concentration**

Did you have trouble concentrating in the past week (on things like conversations, books, movies or daily routines)?

Most often:

□ I never or rarely had trouble

□ I had trouble some of the time

□ I had trouble most of the time

5) **Roles & responsibilities**

Describe your ability to do the things you needed to do at work, at home, and to take care of yourself and your family in the past week.

Most often:

□ I could do all or most of the things I needed to do

□ I could do some of the things I needed to do

□ I could not do the things I needed to do

PREFERENCE-BASED MULTIPLE SCLEROSIS INDEX

Pour chacun des items suivants, choisissez l'option qui correspond à l'état dans lequel vous avez été le plus souvent au cours de la dernière semaine.

1) Marche

Décrivez votre capacité à marcher au cours de la dernière semaine.

Le plus souvent:

□ J'ai pu faire de la marche rapide comme loisir ou sport

□ J'ai pu marcher pour accomplir les tâches que j'avais à faire dans la journée (pour vous rendre à un transport, un endroit public ou à votre travail)

□ J'ai pu marcher seulement quelques pas ou j'utilise un fauteuil roulant

2) Fatigue

Décrivez votre fatigue au cours de la dernière semaine.

Le plus souvent:

□ Je ne me suis jamais senti fatigué au point où j'ai eu à me reposer.

□ Je me suis senti fatigué au point où j'ai eu à me reposer une ou quelques fois pendant la journée

□ Je me suis senti fatigué au point où j'ai eu à me reposer une grande partie de la journée

3) Humeur

Décrivez votre humeur au cours de la dernière semaine.

Le plus souvent :

□ Je ne me suis pas senti triste ou déprimé

□ Je me suis senti un peu triste ou déprimé

□ Je me suis senti très triste ou déprimé

PREFERENCE-BASED MULTIPLE SCLEROSIS INDEX

Pour chacun des items suivants, choisissez l'option qui correspond à l'état dans lequel vous avez été le plus souvent au cours de la dernière semaine.

4) Concentration

Avez-vous eu des problèmes à vous concentrer au cours de la dernière semaine (en suivant une conversation, lisant un livre, regardant un film ou en complétant votre routine quotidienne)?

Le plus souvent :

☐ Je n'ai jamais ou rarement eu de difficulté

☐ J'ai quelques fois eu de la difficulté

☐ J'ai presque toujours eu de la difficulté

5) Rôles & responsabilités

Décrivez votre capacité à accomplir les choses que vous devez faire au travail, à la maison, et pour prendre soin de vous et de votre famille au cours de la dernière semaine.

Le plus souvent :

☐ J'ai pu faire toutes ou la plupart des choses qu'il fallait que je fasse

☐ J'ai pu faire quelques-unes des choses qu'il fallait que je fasse

☐ Je n'ai pas pu faire les choses qu'il fallait que je fasse

CHAPTER 12: Integration of Manuscripts 5 and 6

Research questions of Manuscripts 5 and 6

Manuscript 5:

The development of a bilingual MS-specific health classification system: the Preference-Based Multiple Sclerosis Index (PBMSI).

Manuscript 6:

Developing a valuation function for a multiple sclerosis specific classification system: comparison of standard gamble and rating scale.

Integration of Manuscripts 5 and 6

In Manuscript 5, the PBMSI items were revised using patient and expert feedback. The qualitative review process allowed us to clarify any ambiguous phrasing and make the items more uniform in terms of their instructions and response options. This process will help minimize unnecessary cognitive burden on patients when answering the questionnaire, increase the accuracy of reporting and reduce measurement error.

Preference-based measures produce n^i unique health states (n = number of response levels and i = number of items). Because these measures can generate hundreds and thousands of health states, it is simply not practical to directly value all of the health states described by the classification system. As a result, the typical procedure used when developing a preference-based measure is to value a *subset* of health states, and then combine them in a mathematical model to predict a score for all possible health states described by the classification system.

Two well-known methods of valuing health states are the Standard Gamble (SG) and the Rating Scale (RS). To date, no agreement has been reached in terms of which method should be used in the valuation of health states. There are strong conceptual differences between the two methods which could affect patients' capacity to understand and respond appropriately to the task demanded. Therefore, a head-to-head comparison was thought to be of use in the context of MS and in the context of developing a preference based measure. In the next manuscript, we elicited patient preferences for the different items in the PBMSI using the SG and RS, and compared the two methods on absolute and utility values, level of difficulty, and discriminative ability.

CHAPTER 13 (MANUSCRIPT 6)

Developing a valuation function for a multiple sclerosis specific classification system: comparison of standard gamble and rating scale

Ayse Kuspinar[1], Simon Pickard[2], Nancy E. Mayo[1,3]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada

[2]Center for Pharmacoepidemiology and Pharmacoeconomic Research and Department of PharmacySystems, Outcomes and Policy, University of Illinois at Chicago, Chicago, IL, USA.

[3]Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

In preparation for submission to *Quality of Life Research*

Communication addressed to:

Ayse Kuspinar, M.Sc., Ph.D. Candidate
School of Physical & Occupational Therapy
Faculty of Medicine, McGill University
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
Tel: 514-934-1934 ext 31564
E-mail: ayse.kuspinar@mail.mcgill.ca

ABSTRACT

Objective: When making clinical decisions about which treatment is better or worse for a given patient, the patient's perspective on the benefits and risks is relevant. In this study we elicited patient preferences for different items in a Multiple Sclerosis (MS) specific health classification system using the Standard Gamble (SG) and the Rating Scale (RS). The purpose of this study is to contribute preliminary evidence towards the similarities and differences in the SG and the RS to reflect patient preferences for the different items in an MS specific health state measure, where contrasts were on absolute and utility values, level of difficulty, and discriminative ability.

Methods: Two different samples were recruited for the study. The first (development) sample provided cross-sectional data to generate the preference weights for the valuation of health-states which were then used to develop ($_D$) the MAUF$_D$. For the development sample, the distribution of SG and RS were compared across levels of perceived difficulty in completing the valuation. The parameters from the MAUF$_D$ were applied to a second sample (the validation sample) to produce the MAUF$_V$ and the distribution compared across key measures known to reflect the impact of MS.

Results: Health states that were assessed using the RS were rated lower than when assessed with the SG. The lowest mean health state value with the RS was 0.39 and the highest was 0.65. The mean SG values were much greater, with the lowest being 0.80 and the highest being 0.91. Correlations between the two methods were very low ranging from -0.29 to 0.15. Two different MAUF were calculated, one based on SG values and the other on RS values. Bland-Altman plots to assess agreement revealed that the difference in scores produced by each MAUF was clinically meaningful and a paired t-test analysis demonstrated that this difference was statistically significant.

Conclusion: The SG compared to the RS, produced higher utility and was more difficult for patients with MS to understand. Although the SG is a classical technique of measuring preferences, similar to other studies, we did not find the SG practical in this patient population. Furthermore, in the broader policy arena of allocating resources across multiple health conditions, the standard approach of using generic preference-based measures with general population weights would be difficult to disapprove. However, in the context of use here, which would be to evaluate the effect

of interventions that are expected to impact widely on the health of individuals with MS, the PBMSI with patient preferences shows promise.

INTRODUCTION

Multiple Sclerosis (MS) is an inflammatory, demyelinating disease of the central nervous system (CNS)[1], with wide ranging effects on function, health, and quality of life. From a clinical perspective, the most widely used measure is the Expanded Disability Status Scale (EDSS), which is a single-item disability classification scale used by MS neurologists to quantify disability. It is known to have a number of psychometric limitations[2-5]. From the patient's perspective, a number of MS specific and generic health indices have been used with the most common being the generic SF-36[6]. The only measures which yield one value for quantifying the overall health impact of MS are ones that have preference weights applied to the different dimensions measured. The psychometric properties of these generic preference-based measures have recently been reviewed and limitations identified[7]. A feature of all these measures is that the preference weights are obtained by asking members of the general population to consider the health-impact of each item, whether or not they have experienced the effect. Currently, there is no MS-specific preference-based measure and the choice of who would weight the items has not been resolved.

It has been argued that general population weights are the most appropriate particularly for informing policies about resource allocation in a global context where all health conditions are competing for the same resources. However, for making clinical decisions about which treatment is better or worse for a given patient, perhaps the patient's perspective on the benefits and risks is relevant. Patient' preferences for health states have been shown to differ systematically from those obtained from the general population[8], with patients valuing sub-optimal health states higher. When the health states are hypothetical with both the patient sample and the general population sample naive to the health state, little differences are observed[9].

In addition to who is being asked to value the health-state, there are also differences in how the valuation is done. Two of the most well-known methods of valuing health states are the standard gamble (SG) and the rating scale (RS). The RS scale typically asks individuals to place a given state on a vertical ruler-like scale (i.e. feeling thermometer). The distance between the placements of health states corresponds to the subject's understanding about the relative differences between the health states. With the SG, respondents are presented with a given health state, and are asked to consider whether they would prefer to remain in that health state for the rest of their life or take a chance with a new (imaginary) treatment. They are told that the new treatment has the ability to return them to perfect health immediately but also has the ability to cause immediate death. The probability of returning to full health on taking the new treatment is gradually decreased (and the chance of death increased) until the patient decides to remain in the given health state. The greater the risk of death the subject is willing to consider, the lower the utility value of the health state of interest. Both the RS and the SG provide a score from 0 (dead) to 1 (perfect health).

To date, no agreement has been reached in terms of which method should be used in the valuation of health states. The SG is a classical method of measuring preferences and is based on the axioms of expected utility theory of Von Neumann and Morgenstern.[10] The SG is strongly preferred by health economists, as it is the only valuation method that includes theoretical foundations of economics and an element of risk (one of the axioms of utility theory).[11;12] However, the SG has been criticized for placing high cognitive burden on respondents[13-15] and being prone to risk aversion bias.[16;17] The RS is based on psychometric or measurement theory[18] and has gained popularity over the years because of its simplicity and ease of use.[13;16] However, the RS has been critiqued for not including an element of choice or decision making under uncertainty, and not being rooted in economic theory.[16]

There are such strong conceptual differences between the two methods that could affect patients' capacity to understand and respond appropriately to the task demanded, a head-to-head comparison was thought to be of use in the context of MS and in the context of developing a preference based measure. The purpose of this study is to contribute preliminary evidence towards the similarities and differences in the SG and the RS to reflect patient preferences for the different items in an MS specific health state measure, where contrasts were on absolute and utility values, level of difficulty, and discriminative ability.

METHODS

A MS specific classification system, titled the Preference-Based Multiple Sclerosis Index (PBMSI), was recently developed using input from patients and clinical experts.[19] Domains for the classification system were developed based on semi-structured interviews from 185 patients with MS, and one item per domain was selected using Rasch analysis.[19;20] Then, each selected item was qualitatively reviewed by a group of clinical experts and patients with MS.

Figure 1 presents the methodological steps for this study. Two different samples were recruited. The first (development) sample provided cross-sectional data to generate the preference weights for the valuation of health-states which were then used to develop (D) the MAUF_D. For the development sample, the distribution of SG and RS were compared across levels of perceived difficulty in completing the valuation. The next step was to produce the MAUF_D based on valuations obtained from both the SG and RS. The second sample provided additional cross-sectional data to validate (V) the MAUF, termed MAUF_V. The parameters from the MAUF_D were applied to the validation sample to produce the MAUF_V and the distribution compared across key measures known to reflect the impact of MS.

Selection of Subjects

The development sample for the valuation of health states was recruited through advertising in three venues: MS Society of Canada website; the 2012 Quebec Summit on Multiple Sclerosis; and outpatient MS clinic of the Montreal Neurological Hospital. To participate, individuals had to be diagnosed with MS and be older than 18 years of age. The study was approved by the hospital's ethics committee and written informed consent was obtained from participants prior to doing the online survey.

The validation sample was subjects with MS who were participating in a clinical trial of exercise. The protocol for this study has been published.[21] Briefly, participants were recruited from 3 MS clinics in the Montreal area and were aged 19-65, diagnosed after 1994, ambulatory, and able to speak and read English or French. Participants were excluded if they had an additional illness that restricted their function, had suffered at least one relapse during the past 30 days, or were unable to respond to simple questions on orientation and memory. This sample was ideal for the assessment as they were stable at time of recruitment and had the language and cognitive capacity to understand the questions. The ethics committees of each participating hospital approved the study.

Measures

The main measure for this study was the PBMSI, administered to both the development and validation samples. Two methods of valuing the health states from the PBMSI were the SG and RS used to derive $MAUF_D$. Measures of global disability, walking capacity and general health perception were used to validate $MAUF_V$.

PBMSI: The PBMSI is a brief self-administered questionnaire consisting of five items: walking, fatigue, mood, concentration, and roles and responsibilities. Each item has three response options, and the recall time frame is 'over the past week'. The classification system produces 243 (3^5) health states.

Selection of health states for valuation: Each patient valued 12 health states: 5-single attribute level states, 5 corner states, all worst and all intermediate states. These states are as follows:

- Single-attribute level states: a given item was described at less than full function (response level 2) while all other items were set at their best level (response level 1).
- Corner states: a given item was described at its worst level (response level 3) while all other items were set at their best level (response level 1).
- All worst was described as the worst level on all items (response level 3), and all intermediate was described as less than full function on all items (response level 2). Patients also assigned a value for the state 'dead' on the RS. A value for the state 'dead' was not required for the SG, as it was anchored from dead to perfect health.

Preferences for the above health states were obtained from patients with MS using an online survey. In the survey, patients were asked to fill out the PBMSI and answer certain socio-demographic and clinical questions. Then they were asked to value selected health states using the SG and RS.

Standard Gamble: Patients were asked to rate the single-attribute and corner states using the standard gamble (SG). In the SG, patients were presented with a less than perfect health state (i.e. a corner state or single-attribute state), and asked to imagine themselves in that health state for the rest of their life. Then they were asked to imagine that they were given a treatment. If the treatment was successful, they would be restored to full health. But if the treatment were to fail, they have a probability of dying immediately. Essentially respondents are asked to indicate the highest risk of death (in percentage) they would accept with the treatment. However, the questionnaire that elicited these probabilities, referred to death as “failure”. This is a common procedure in the literature.[22-26] The response options were given in a drop down menu, as follows: ‘0% chance of ‘failure’ (100% chance of ‘success’)…5% chance of ‘failure’ (95% chance of ‘success’)…etc.’ Patients were asked to select only one response option from the list provided. The probability of ‘success’ that they were willing to accept with the treatment was their SG value (i.e. 100% ‘success’ is equal to a SG value of 1.0, 95% ‘success’ is equal to a SG value of 0.95 etc.)

The format also allowed for the assessment of states worse than dead if respondents indicated that they would take the treatment even if it had 0% chance of ‘success’ (100% chance of ‘failure’).

Rating Scale: Patients were asked to rate each of the single-attribute and corner states on a RS from 0 to 100, where zero was the worst imaginable health state and 100 was the best imaginable health state. Patients were also asked to provide on the RS a value for the state ‘dead’. If state dead was identified as the worst state and was placed at the 0 end of the scale, then preferences were simply equal to the scale value given to each health state. If death was not identified as the worst state but was placed on some intermediate point on the scale (d), then preferences were measured as: $(x-d)/(1-d)$, where x was the rating given to a health state and d was the rating given to death.

Difficulty: At the end of the survey patients were asked to rate how difficult it was to answer the PBMSI items, the RS, and the SG. Responses were recorded on a four-point Likert scale (very easy, fairly easy, fairly difficult, and very difficult).

Global disability: Global disability was measured using Patient-Determined Disease Steps (PDDS), self-reported outcome of disability in MS.[27] It has nine ordinal levels ranging between 0 (normal) and 8 (Bedridden) and PDDS scores can be converted into classifications of mild, moderate, or severe disability.[28] The PDDS is a surrogate measure of the Expanded Disability Status Scale (EDSS) and has shown to be strongly correlated with the EDSS.[29] A score of 0 on the PDDS is normal and is equal to an EDSS score of 0. A score of 3 characterizes gait disability without the need for an assistive device and corresponds to an EDSS score of 4.0 to 4.5. PDDS scores of 4, 5, and 6 represent need for assistive devices and is equivalent to EDSS scores of 6 to

6.5. For both the PDDS and EDSS, scores of 7 correspond to being wheelchair bound, and scores of 8 correspond to being confined to bed.[28]

Functional exercise capacity: The 6MWT is a simple performance-based test that measures functional exercise capacity. The reliability of the 6MWT has been assessed in persons with MS. The intra-class correlation coefficient is 0.96 for test-retest reliability and 0.93 for inter-rater reliability.[30]

General Health Perception: The RAND-36 is one of the most widely used generic health profiles and the first question measures general health perception, which is formulated as, "In general, would you say your health is…," with five nominal response options ranging from excellent to poor.[31] General health perception is easy to measure and can provide information on the person's well-being and overall HRQL. Furthermore, it has been shown to be a predictive factor in the progression of disease.[32] Patients with MS who evaluate their health as "poor" or "fair" have twice the chance of experiencing a worsening in disability 1 year later, versus patients who evaluate their health as "good", "very good", or "excellent".[32] General health perception, is a patient-reported outcome (PRO) and is important in providing additional information on disease activity in patients with MS not captured by direct measurement or observation, and is sensitive to the presence of symptoms (e.g. weakness, sensation, bladder, bowel, and fatigue), their severity and type.[33]

EQ-5D: The EQ-5D[34] is a generic preference-based measure of HRQL that consists of two parts. The first part includes 5 separate domains; mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each domain has 3 levels: no problems, some problems, extreme problems. The second part consists of a Visual Analogue Scale (EQVAS) to measure self-perceived health on a vertical scale from 0 to 100, where 0 is the worst imaginable health state, and 100 is the best imaginable health state.[34] The MAUF_{D} was compared against the EQ-5D, as it is a commonly used preference-based measure in MS and is recommended by the National Institute for Health and Care Excellence (NICE) for economic evaluation.

Statistical Methods

For the development sample, the distribution of SG and RS values was obtained for each health state and plotted by quartile; Pearson correlation coefficients were also calculated.

Concordance between the reported levels of difficulty for the SG and RS was presented and agreement assessed using un-weighted and weighted Kappa. Generalized estimating equations (GEE) were used to assess the impact that reported level of difficulty had on SG and RS values, considering the correlation arising from multiple valuations per person.

Two MAUF (i.e. scoring algorithms) were developed (MAUF$_D$): one based on SG values and the other based on RS values. The methodology used to develop the MAUF$_D$ closely followed the procedures described in the manual for the development of the HUI3.[35]

The person-mean approach was used to develop the valuation functions.[35] In other words, the functions were estimated from the mean responses of the sample for the single-attribute health states and corner states.

A utility scale runs from 0.0 (dead) to 1.0 (all best/perfect health). Disutility equals one minus utility (disutility = 1 – utility). Thus, the disutility scale ranges from 0.0 for all best/perfect health to 1.0 for dead.

If the sum of the disutility corner states is equal to 1.0, then the valuation function is additive. However, if the sum of the corner states is not equal to one, then the valuation function is multiplicative. The multiplicative function, as specified by MAUT was:

$$u' = \left(1/c\right)\left[\prod_{j=1}^{n}(1 + c * c_j * u'_j) - 1\right]$$

(Eq. 1)

where, $u'$is the required disutility of any PBMSI health state on the perfect health = 0.0, dead=1.0 scale; j is the number of PBMSI items which was 5; c_j is the person-mean disutility for the corner state; u'_j is single-attribute level disutility score; and $\prod_{j=1}^{n}$ is the product of all $(1 + c * c_j * u'_j)$. The scaling parameter c was calculated by iteratively solving the following equation:

$$1 + c - [\prod_{j=1}^{5}(1 + c * c_j)] = 0$$

(Eq. 2)

where $\prod_{j=1}^{5}$ is the product of all (1 + c x c_j) from c_1 to c_5; and c_j is the person-mean disutility for the corner state.

The scaling parameter c depends on the sum of the corner disutility states:

If	$\sum_{j=1}^{5} c_j > 1$	then -1 < c < 0;	(Eq. 3a)
if	$\sum_{j=1}^{5} c_j = 1$	then c = 0, and the valuation function is additive;	(Eq. 3b)
and if	$\sum_{j=1}^{5} c_j < 1$	then c > 0.	(Eq. 3c)

If the valuation function is additive, c = 0 is the only root of equation 2. If the valuation function is not additive, equation 2 will have 2 roots: (i) a trivial solution (c = 0) and (ii) a non-trivial solution ($c \neq 0$). We will be searching for the non-trivial solution, and the sum of the corner states will tell us where to search for it (i.e. if sum of corner states is greater than 1, then -1 < c < 0; if sum of corner states is less than 1, then c > 0).

Excel Solver was used to iteratively solve for the scaling parameter c. All other analyses were conducted using SAS9.3.

Required sample size for the MAUF$_D$

We estimated the sample size for this valuation to yield a 95% confidence interval (95%CI) around the mean value for the SG and RS of ± 0.05 points. Clinically meaningful difference on the SG (as well as the RS) is approximately 0.10 points[13]; half the difference was chosen as it would not be meaningful and, therefore, this CI would indicate precision in the estimates of value.

Calculation of the 95% CI requires an estimate of the population standard deviation (SD). To our knowledge, there are no studies have reported the SD for the SG in people with MS. Therefore, sample size calculations were based on the values obtained for the RS in the MS Life-Impact Study[19;20;36] conducted in a similar population. The SD of the RS value for ‘best imaginable health’ was 0.08. Based on this information the number of people required per health state was equal to 10 (calculated using the following formula: *1.96*(0.08/√n) = 0.05*). As there were 5 corner states, the required sample size for this study was 50 people.

Agreement between the SG MAUF and RS MAUF for both samples was depicted using scatter plots. For perfect agreement, all data points are expected to be on the diagonal line, the line of equality. For both the development and the validation samples, the Bland-Altman method was used to analyze agreement between the SG MAUF and the RS MAUF. This method contrasts the mean difference between two MAUF (y axis) against the average of the two MAUF, which represents the latent trait of "utility". The graph shows 95% limits of agreement around the mean difference (1.96 SD). Perfect agreement between the SG MAUF and the RS MAUF would be indicated by a mean difference equal to 0 and no pattern across the latent trait. The distribution of the differences in values between the MAUF SG and MAUF RS were plotted using a histogram. A paired t-test was used to contrast these two values.

The distribution of items on the PBMSI obtained from the clinical trial validation sample was identified. The known-groups method was used to test the discriminative ability of the SG and RS MAUF_V against different measures of disability, namely the PDDS, the 6MWT and the general health perception item of the RAND-36. The MAUF_V was also compared against the generic preference-based measure EQ-5D. The linear test for trend was employed to test if gradients across levels of disability were statistically significant.

RESULTS

Sample

Table 1 presents the demographic and clinical characteristics of the two samples, development and validation. These samples were chosen using quite different sampling frames, and hence were expected to differ somewhat. However, the two samples were similar on age (mean ~ 47 years) and proportion women (75%-79%). The clinical trial (validation) sample was comprised of people recruited into an exercise intervention trial and showed lower disability in walking (level 1), lower fatigue, better mood, but more challenges with regular roles and responsibilities. Also shown is the number of people in the most common health states. For example, 8% of the validation sample had the health state 11111, reflecting the best level on all 5 dimensions. Furthermore, approximately 13% of the samples had the health state 22111, reflecting some problems with walking and fatigue, but no problems with mood, concentration, and roles and responsibilities. No statistical comparison between samples was done because it was known from the outset that these two samples did not arise from the same population.

Table 2 presents for the development sample the mean SG and RS values for level 2 and level 3 of each item in the PBMSI as well as two multi-attribute health states, all at level 2 and all at level 3. All health states were rated lower using the RS than the SG. The mean RS values ranged from 0.20 to 0.65, whereas the mean SG values ranged from 0.60 to 0.91. Also presented are the correlation coefficients between the SG and RS; weak correlations were observed ranging from -0.29 to 0.15.

Figure 2 presents a distribution of the RS values by percentile for each of the corner states (i.e. level 3, worst, for each item). Higher scores on the RS indicate better health. The RS values were fairly evenly distributed. The median value, which is represented by the end of the light blue bar, was 0.5 for severe walking impairment, severe fatigue and depression. The median value for impaired concentration and restricted roles and responsibilities were 0.6 and 0.4, respectively.

Figure 3 presents the percentile distribution of the SG values for the corner states. The SG values were on a scale from 0 to 1, where higher scores indicate better health. The SG values were considerably higher than RS values for all of the items, with the median values being 0.9 or 0.95. Twenty-five percent of the sample rated having severe walking impairments, severe fatigue, and severe impaired concentration equivalent to perfect health (1.0).

Table 3 presents the percent agreement between the levels of difficulty reported by patients for the SG (rows) and RS (columns). Across all levels of difficulty, 38% (23/61) found both methods to be of equal difficulty (diagonal cells); 50% (30/61) rated the SG at a higher level of difficulty than the RS (cells below the diagonal). Only 5 people rated the RS harder than the SG (cells above the diagonal), but the 6 people rating SG as "very easy" scored all health states with virtually the same value, 0.95 (data not shown). Chance corrected agreement was poor using un-weighted Kappa (κ 0.09; 95% CI: 0.08 to 0.25) and weighted Kappa (κ 0.13; 95% CI: -0.08 to 0.34).

To answer the question as to whether level of difficulty had an impact on health state values, we regressed method of valuation (SG, RS) onto the 12 health state values using GEE, which considered the correlation (non-independence) of the valuation, including the interaction between method and health state. The model was health state value = method (RS/SG) + item (1-12) + method*item. As the interaction term was non-significant, it was dropped. For the RS, the effect of difficulty across all items when compared to the SG was equal to -0.25. When the model was adjusted for difficulty, the difference was accentuated to -0.32. The difference between RS and SG did not depend on item (non-significant interaction).

Table 4 presents the parameters used to develop the MAUF$_\text{D}$ based on the SG and RS values obtained in the development sample. The first column presents the mean RS and SG utility values for each response level, where level 1 was the best, level 2 was intermediate, and level 3 was the worst. The first level of each item was 1.0 (perfect health). As expected, there was a drop in utility values from level 1 to level 2 to level 3. For each item, response level 3 was the corner state utility value. The second column of Table 4 presents the disutility values (1-utility) for each of the item response levels. The third column presents the mean utility values rescaled so that the third response level of each item was 0.0, and the first response level was 1.0. The fourth column is the rescaled mean disutility score, which is equal to 1 - the rescaled mean utility score (presented in third column). These are the parameters used to develop the valuation function (MAUF$_\text{D}$).

Table 5 presents the MAUF$_\text{D}$ developed using the SG values presented in Table 4. The sum of the corner states was equal to 0.85, which is less than 1.0; therefore the MAUF$_\text{D}$ was multiplicative and yielded two solutions for equation 2. Based on equation 3c, the non-trivial solution was greater than 0. Using the iterative solution (Eq. 2) an exact value for the non-trivial solution *c* was calculated, and found to be equal to 0.4821.

The SG MAUF$_\text{D}$ for the PBMSI in dis-utilities was:

$$
\begin{aligned}
\textit{PBMSI Disutility}_{\text{(perfect health = 0, dead = 1)}} = (1/0.4821) \text{ x } & \\
([1 + \{0.4821\} \text{ x } 0.18 \text{ x } u'_1] \text{ x } & \\
([1 + \{0.4821\} \text{ x } 0.19 \text{ x } u'_2] \text{ x } & \\
([1 + \{0.4821\} \text{ x } 0.16 \text{ x } u'_3] \text{ x } & \\
([1 + \{0.4821\} \text{ x } 0.12 \text{ x } u'_4] \text{ x } & \\
([1 + \{0.4821\} \text{ x } 0.20 \text{ x } u'_5] - 1) &
\end{aligned}
$$

Where the values of u'_1, u'_2, u'_3, u'_4, u'_5 (the single-attribute mean disutilities) are selected from Table 5 depending on the individual's responses to the PBMSI items. The calculated disutility on the perfect health=0.0, dead = 1.0 scale can then be converted into a utility score on a dead = 0.0, perfect health = 1.0 scale:

PBMSI utility $_{(\text{dead = 0, perfect health =1})}$ = 1 – PBMSI disutility

Table 6 presents the MAUF_D based on the RS values. The procedure used to develop the RS MAUF_D was identical to the process described for the SG MAUF_D. Using the RS values, the sum of the corner states was equal to 3.65 and the scaling parameter was calculated to be equal to -0.9987. The full valuation function can be found in Table 7.

Figure 4 presents, for the development sample, a scatter plot to assess agreement between PBMSI scores obtained using the RS MAUF_D against scores obtained using the SG MAUF_D. As none of the data points were on the line of equality (red line) there was no agreement between the two methods. Scores produced by SG MAUF_D were consistently considerably higher than scores produced by the RS MAUF_D, yielding a strong correlation (0.8), but no agreement.

Figure 5 presents, for the development sample, the Bland-Altman plot between the SG MAUF_D and the RS MAUF_D. The *x* axis shows the mean of the results of the two methods ([SG MAUF_D + RS MAUF_D]/2), which is considered to represent the latent trait of "utility". The *y* axis is the absolute difference between the two methods ([SG MAUF_D − RS MAUF_D]). If the methods are concordant, the mean difference should be 0 with no pattern across the latent trait. The average difference between the methods was 0.46 (represented by the middle red line), and 95% of patients had a difference in scores between 0.24 and 0.68. A clinically meaningful difference on the SG or RS is 0.10; therefore the mean difference between the two methods was almost 5 times greater than the clinically meaningful difference. Additionally, there was a distinct pattern to the values such that, at the low end of the latent trait (poor health state) the differences were small; as latent health state improved, the difference between the methods increased.

Figure 6 presents a histogram of the distribution of differences between the SG MAUF_D and RS MAUF_D. As presented in the graph, for 90% of the sample, this difference was between 0.3 and

0.65, which was clinically meaningful. A paired t-test revealed that this difference in scores was statistically significant (p-value <0.0001).

Figure 7 presents, for the validation sample, a scatter plot of the PBMSI scores obtained using the RS MAUF$_V$ against scores obtained using the SG MAUF$_V$. Similar to the results obtained for the development sample, there was no agreement between scores produced by the two MAUF$_V$.

Figure 8 presents the Bland Altman plot for the validation sample, which shows that the mean difference between the SG MAUF$_V$ and RS MAUF$_V$ is 0.44, 4 times greater than the clinically meaningful difference of 0.1 points.

Figure 9 presents, for the validation sample, the distribution of the difference in scores between the SG MAUF$_V$ and RS MAUF$_V$. For almost 30% of the sample the difference in scores between the two scoring algorithms was between 0.3-0.4, and for 60% of the sample this difference was equal to 0.5. A paired t-test between scores indicated that the difference in scores between the SG MAUF$_V$ and RS MAUF$_V$ was statistically significant (p-value <0.0001).

Table 7 presents for the validation sample, the ability of the SG and RS MAUF$_V$ to discriminate between different clinical subgroups, assessed using the PDDS, 6MWT and the general health perception item of the RAND-36. Both the SG MAUF$_V$ and the RS MAUF$_V$ were able to differentiate between different levels of disability measured using the PDDS. However, the RS MAUF$_V$ had a wider range of values than the SG MAUF$_V$. The EQ-5D valuation function was not able to differentiate between moderate and severe levels of disability. For the 6MWT, both the SG and the RS MAUF$_V$ were able to differentiate between different levels of walking capacity, however, the values produced by the RS MAUF$_V$ were lower than the SG MAUF$_V$. The EQ-5D was also able to differentiate between different levels of walking capacity. As for general health perception, the SG MAUF$_V$ was able to differentiate between all levels of health perception (excellent, very good, good and fair). However, the RS MAUF$_V$ was only able to differentiate between excellent, very good and good health, but not between good and fair health. The EQ-5D also presented with problems discriminating between different levels of health perception, specifically between very good and good (p-value = 0.06).

DISCUSSION

This study compared two methods of valuing health states in people with MS, and revealed that the two methods produced considerably different results from each other. On a scale from 0 (dead) to 1 (perfect health), values produced by the SG were consistently higher than those produced by the RS. The median values for the corner state items were between 0.4 and 0.6 on the RS, and between 0.90 and 0.95 on the SG. With the SG, 50% of the sample rated having severe walking impairments, severe fatigue, severe impaired concentration and depression close or equivalent to perfect health (1.0). For these same items, none of the respondents gave a value of 1.0 on the RS.

Our results are similar to previous studies that have compared the SG and the RS. Jansen and colleagues[37] compared the two methods in 51 women with breast cancer. They asked patients to value a hypothetical chemotherapy scenario, and reported that values elicited using the SG (mean ~0.9) were consistently higher than the RS (mean ~0.6). Juniper and colleagues[38] compared the SG and RS in 40 patients with asthma. In their study, more than half of the patients (n=23) rated their current health equal to 1.0 (perfect health) on the SG, even though they represented patients at the more severe end of the spectrum (80% required inhaled steroids). Furthermore, among these 23 patients who rated their health equal to 1.0 on the SG, only 7 provided the same value on the RS. The remainder of patients provided values that were much lower (0.45). Sullivan and colleagues[39] interviewed 52 patients with diabetes mellitus on various health states describing different levels of disease severity in diabetic peripheral neuropathy. For all health states, the SG scores were considerably higher than the RS. The highest median preference score for the SG was 0.96 (mild neuropathy) and the lowest was 0.65 (below-knee amputation). On the RS, the highest median score was 0.89 (mild neuropathy) and the lowest was 0.23 (below-knee amputation).

In our study, correlations between the SG and the RS were very weak (r ~ 0.1), thus reinforcing the fact that there were considerable discrepancies in the values elicited by the two methods. These low correlations were similar to what others have reported in cancer (r=0.18),[37] chronic musculoskeletal pain (r= 0.21)[40] and liver cirrhosis (r = -0.07)[41] and asthma(r=0.18).[42]

The SG is a method that assesses the probability an individual would risk death to regain perfect health. As death is a highly undesirable state, patients may be inclined to stop the gambling earlier, thus resulting in an overestimate of the value associated with an impaired health state.[16;26;43] In the context of MS, the possible risk of dying after treatment is far from realistic as existing medical treatments are rarely life threatening. Instead treatment is directed at slowing the progression of disease or disability. As the RS does not involve risk or decision making under uncertainty, values elicited with this method tend to be systematically lower than the SG.

Fifty percent of our sample rated the SG at a higher level of difficulty than the RS. These findings are concordant with previous studies that have compared the SG with the RS. In patients with cancer, Dobrez and Calhoun[44] reported that 17% of their sample did not comprehend the SG method. Similarly in HIV/AIDS patients, Sakthong[45] and colleagues reported that the SG was more difficult for patients to understand compared to the RS (p=0.002), and that the completion time for the SG was much longer than the RS (average 5 minutes per health state vs 0.9 minutes per health state).

The SG method may be difficult for patients to comprehend because the concept of probabilities is a challenging one to grasp and far from everyday experience.[46] Lack of comprehension of the method is an important issue in the valuation of health states, as it can compromise the accuracy or reliability of the data collected.[47-48]

A MAUF was developed based on values obtained using each of the methods, and a PBMSI score was calculated for the development and validation samples. The average PBMSI score based on SG values was 0.62 for the development sample and 0.73 for the validation sample. Whereas the mean PBMSI score based on RS values was 0.16 for the development sample and 0.29 for the validation sample. Our results revealed that the type of valuation method used had a large impact on the MAUF. For the same health states, the MAUF developed based on the RS produced much lower utility values than the SG.

The SG is a classical method of measuring preferences, based on the axioms of expected utility theory proposed by von Neumann and Morgernstern.[10] It is the only available technique that measures preferences under conditions of both risk and uncertainty.[11;12] However, this study raises questions on the suitability of the SG in MS, as patients had difficulty understanding the task and were not willing to risk death for an improvement in health. On the SG, fifty percent of patients valued severely impaired health states (such as being wheelchair bound) close to or equal to 1.0 (perfect health). Conversely, on the RS, less than 5% of the sample valued severe health states equal to 1.0. Furthermore, the RS was reported to be fairly easy to complete and understand by more than two-thirds of the sample. However, despite its advantages in terms of simplicity and feasibility, the RS is criticized for not being a true measure of utilities because it does not meet the utility theory requirement of 'decisions under uncertainty'.

There were several notable features of this study. First, we used an internet based approach to value health states, rather than the traditional interviewer based approach. Traditionally the SG requires the use of a trained interviewer with props, where researchers must either go to the participant's home or offer sufficient incentives to bring the participant to the lab, which are both expensive. The advantage of using an online survey is that patients can complete the survey in the convenience of their home, resulting in greater recruitment or participation. Although other studies have used the internet to elicit preferences,[46;49;50] the validity and reliability of this approach requires further study. Second, in the SG, rather than alternating the proportion of success and death in a "ping-pong" manner we simply asked individuals to indicate the maximum risk of death they were willing to take with the hypothetical treatment. This may have resulted in a higher value of utilities than the former approach. Finally, alternate methods of valuation such as the time trade-off (number of years patients are willing to trade off for perfect health) were not assessed in this study.

In summary, this study elicited patient preferences for various items from a MS-specific classification system using two different valuation methods, the SG and RS. We compared these two methods in terms of the values they produced, their difficulty of use and impact on the MAUF. Our findings demonstrated that, the SG compared to the RS, produced higher utility and was more difficult for patients to understand. Although the SG is a classical technique of measuring preferences, similar to others, we did not find the SG practical in this patient population. Alternately; the RS may be a more suitable approach to elicit values in patients with MS. Furthermore, in the broader policy arena of allocating resources across multiple health conditions, the standard approach of using generic preference-based measures with general population weights would be difficult to disapprove. However, in the context of use here, which would be to evaluate the effect of interventions that are expected to impact widely on the health of individuals with MS, the PBMSI with patient preferences shows promise.

REFERENCE LIST

[1] Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG: Multiple sclerosis. N Engl J Med 2000;343:938-952.

[2] Noseworthy JH, Vandervoort MK, Wong CJ, Ebers GC: Interrater variability with the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) in a multiple sclerosis clinical trial. Neurology 1990;40:971.

[3] Francis DA, Bain P, Swan AV, Hughes RA: An assessment of disability rating scales used in multiple sclerosis. Archives of Neurology 1991;48:299-301.

[4] Goodkin DE, Cookfair D, Wende K, Bourdette D, Pullicino P, Scherokman B, Whitham R: Inter and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke Expanded Disability Status Scale (EDSS). Neurology 1992;42:859.

[5] Amato MP, Fratiglioni L, Groppi C, Siracusa G, Amaducci L: Interrater reliability in assessing functional systems and disability on the Kurtzke scale in multiple sclerosis. Archives of Neurology 1988;45:746-748.

[6] Kuspinar A, Rodriguez AM, Mayo NE: The effects of clinical interventions on health-related quality of life in multiple sclerosis: a meta-analysis. Mult Scler 2012;18:1686-1704.

[7] Kuspinar A, Mayo NE: A review of the psychometric properties of generic utility measures in multiple sclerosis. Pharmacoeconomics 2014;32:759-773.

[8] Peeters Y, Stiggelbout AM: Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities. Value Health 2010;13:306-309.

[9] Pickard AS, Tawk R, Shaw JW: The effect of chronic conditions on stated preferences for health. Eur J Health Econ 2013;14:697-702.

[10] Von Neumann J, Morgenstern O: Theory of games and economic behavior 2d rev. 1947.

[11] Torrance GW: Measurement of health state utilities for economic appraisal. J Health Econ 1986;5:1-30.

[12] Bennett KJ, Torrance GW: Measuring health state preferences and utilities: rating scale, time trade-off, and standard gamble technqiues; in Spilker B (ed): Quality of Life and Pharmaeconomics in Clinical Trials. Philadelphia, Lippincott-Raven Publishers, 1996, pp 253-267.

[13] Feeny D: Preference-based measures: utility and quality-adjusted life years; in Fayers P, Hays D (eds): Assessing quality of life in clinical trials. New York, Oxford University Press Inc., 2005, pp 405-429.

[14] Ryan M, Gerard K: Discrete choice experiments; in Fayers P, Hays D (eds): Assessing quality of life in clincal trials. New York, Oxford University Press, 2005, pp 431-445.

[15] Rosser R, Kind P: A scale of valuations of states of illness: is there a social consensus? Int J Epidemiol 1978;7:347-358.

[16] Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A: Measuring and valuing health benefits for economic evaluation. New York, Oxford University Press Inc., 2007.

[17] Doctor JN, Bleichrodt H, Lin HJ: Health utility bias: a systematic review and meta-analytic evaluation. Med Decis Making 2010;30:58-67.

[18] Krabbe PF, Tromp N, Ruers TJ, van Riel PL: Are patients' judgments of health status really different from the general population? Health Qual Life Outcomes 2011;9:31.

[19] Kuspinar A, Finch L, Pickard S, Mayo NE: Using existing data to identify candidate items for a health state classification system in multiple sclerosis. Qual Life Res 2014;23:1445-1457.

[20] Kuspinar A, Mayo NE: Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis? Health Qual Life Outcomes 2013;11:71.

[21] Mayo NE, Bayley M, Duquette P, Lapierre Y, Anderson R, Bartlett S: The role of exercise in modifying outcomes for people with multiple sclerosis: a randomized trial. BMC Neurology 2013;13:69.

[22] Gudex C. Standard Gamble user manual: props and self-completion methods. 1994.

[23] Kontodimopoulos N, Niakas D: Overcoming inherent problems of preference-based techniques for measuring health benefits: an empirical study in the context of kidney transplantation. BMC health services research 2006;6:3.

[24] Oliver A: Testing the internal consistency of the standard gamble in success and failure frames. Social science & medicine 2004;58:2219-2229.

[25] Dolan P, Sutton M: Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. Soc Sci Med 1997;44:1519-1530.

[26] Robinson A, Loomes G, Jones-Lee M: Visual analog scales, standard gambles, and relative risk aversion. Medical Decision Making 2001;21:17-27.

[27] Learmonth YC, Motl RW, Sandroff BM, Pula JH, Cadavid D: Validation of patient determined disease steps (PDDS) scale scores in persons with multiple sclerosis. BMC Neurol 2013;13:37.

[28] Marrie RA, Cutter G, Tyry T, Vollmer T, Campagnolo D: Does multiple sclerosis-associated disability differ between races? Neurology 2006;66:1235-1240.

[29] Hohol MJ, Orav EJ, Weiner HL: Disease steps in multiple sclerosis: a longitudinal study comparing disease steps and EDSS to evaluate disease progression. Mult Scler 1999;5:349-354.

[30] Paltamaa J, West H, Sarasoja T, Wikstrom J, Malkia E: Reliability of physical functioning measures in ambulatory subjects with MS. Physiother Res Int 2005;10:93-109.

[31] 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;68:150-156.

[32] Nortvedt MW, Riise T, Myhr KM, Nyland HI: Quality of life as a predictor for change in disability in MS. Neurology 2000;55:51-54.

[33] Parkin D, Rice N, Jacoby A, Doughty J: Use of a visual analogue scale in a daily patient diary: modelling cross-sectional time-series data on health-related quality of life. Soc Sci Med 2004;59:351-360.

[34] Kind P: The EuroQol instrument: an index of health-related quality of life; in Spilker B (ed): Quality of Life and Pharmaeconomics in Clinical Trials. Philadelphia, Lippincott-Raven Publishers, 1996, pp 191-201.

[35] Furlong W, Feeny D, Torrance G, Goldsmith C, DePauw S, Zhu Z, Denton M, Boyle M. Multiplicative multi-attribute utility function for the Health Utilities Index Mark 3 (HUI3) system: a technical report. 1998. Centre for Health Economics and Policy Analysis (CHEPA), McMaster University, Hamilton, Canada.

[36] Shahrbanian S, Duquette P, Kuspinar A, Mayo NE: Contribution of symptom clusters to multiple sclerosis consequences. Qual Life Res 17-9-2014.

[37] Jansen SJ, Kievit J, Nooij MA, Stiggelbout AM: Stability of Patients Preferences for Chemotherapy The Impact of Experience. Medical Decision Making 2001;21:295-306.

[38] Juniper EF, Norman GR, Cox FM, Roberts JN: Comparison of the standard gamble, rating scale, AQLQ and SF-36 for measuring quality of life in asthma. European Respiratory Journal 2001;18:38-44.

[39] Sullivan SD, Lew DP, Devine EB, Hakim Z, Reiber GE, Veenstra DL: Health state preference assessment in diabetic peripheral neuropathy. Pharmacoeconomics 2002;20:1079-1089.

[40] Goossens MlE, Vlaeyen JW, Rutten-van M, Âlken MP, van der Linden SM: Patient utilities in chronic musculoskeletal pain: how useful is the standard gamble method? Pain 1999;80:365-375.

[41] Adibi P, Akbari L, Kahangi LS, Abdi F: Health-state utilities in liver cirrhosis: A cross-sectional study. International journal of preventive medicine 2012;3:S94.

[42] Blumenschein K, Johannesson M: Relationship between quality of life instruments, health state utilities, and willingness to pay in patients with asthma. Ann Allergy Asthma Immunol 1998;80:189-194.

[43] Bleichrodt H: A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ 2002;11:447-456.

[44] Dobrez DG, Calhoun EA: Testing subject comprehension of utility questionnaires. Quality of Life Research 2004;13:369-376.

[45] Sakthong P, Schommer JC, Gross CR, Prasithsirikul W, Sakulbumrungsil R: Health utilities in patients with HIV/AIDS in Thailand. Value in Health 2009;12:377-384.

[46] Lenert LA, Sturley AE. Use of the internet to study the utility values of the public. Proceedings of the AMIA Symposium , 440. 2002. American Medical Informatics Association.

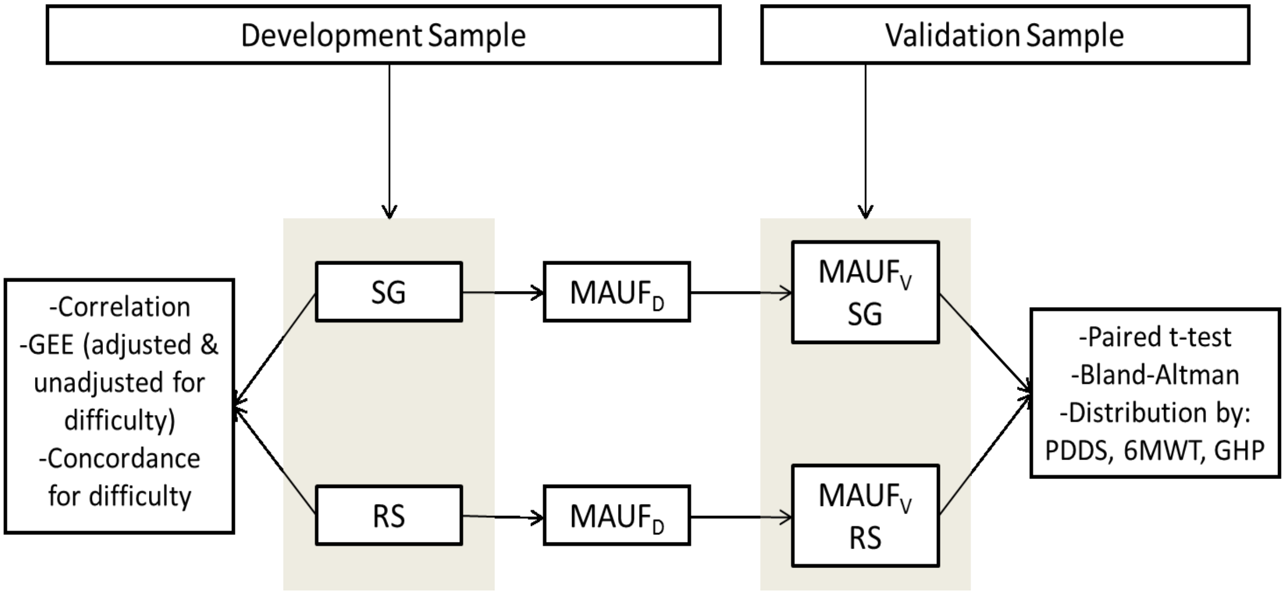
[47] Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA: Measuring preferences for health states worse than death. Med Decis Making 1994;14:9-18.

[48] Wittenberg E, Prosser LA: Ordering errors, objections and invariance in utility survey responses. Applied health economics and health policy 2011;9:225-241.

[49] Stein K, Dyer M, Crabb T, Milne R, Round A, Ratcliffe J, Brazier J: A pilot Internet "value of health" panel: recruitment, participation and compliance. Health and quality of life outcomes 2006;4.

[50] Chang WT, Collins ED, Kerrigan CL: An Internet-based utility assessment of breast hypertrophy. Plastic and reconstructive surgery 2001;108:370-377.

Figure 1 A flow diagram of the methodological steps involved in the study



SG, Standard Gamble; RS, Rating Scale; MAUF_D, Multi-Attribute Utility Function Development; MAUF_V, Multi-Attribute Utility Function Validation; GEE, Generalized Estimating Equations; PDDS, Patient Determined Disease Steps; 6MWT, 6 Minute Walk Test; GHP, General Health Perception.

Figure 2 Rating scale values by quantiles for PBMSI corner states in the development sample

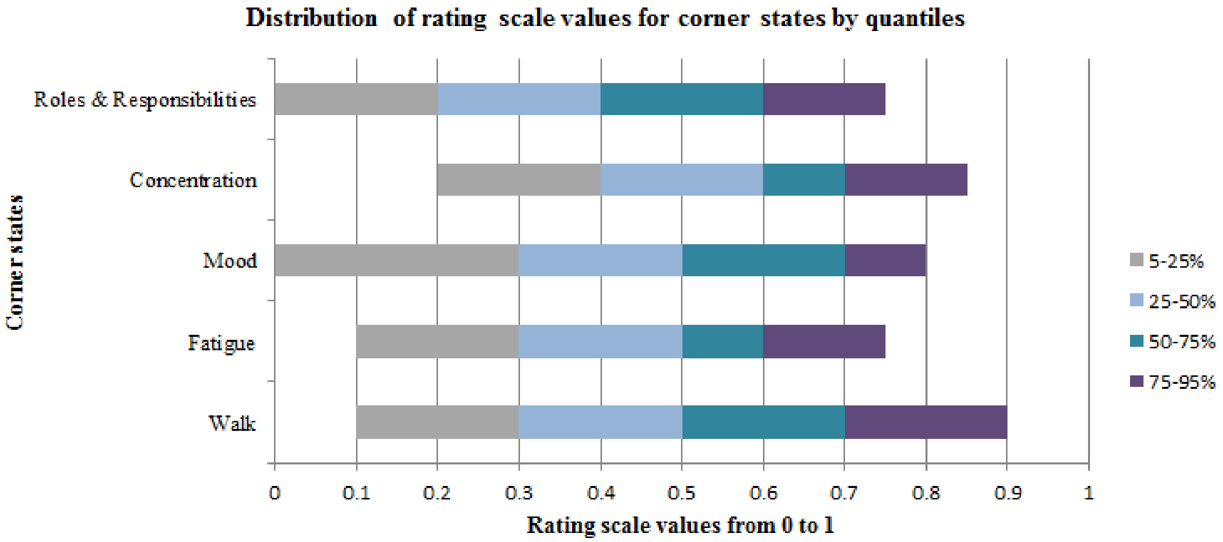


Figure 3 Standard gamble values by quantiles for PBMSI corner states in the development sample

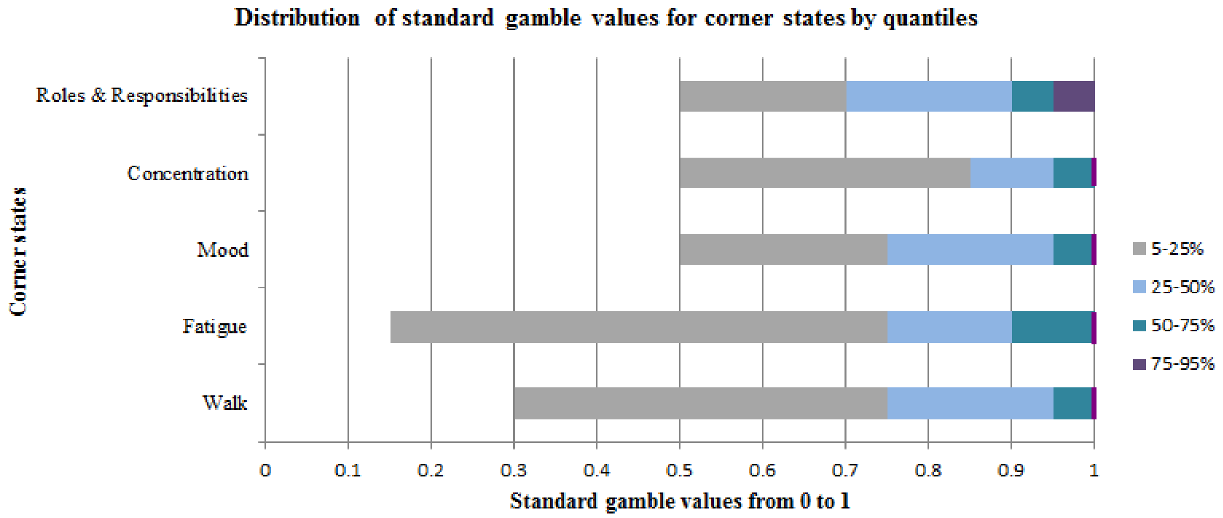
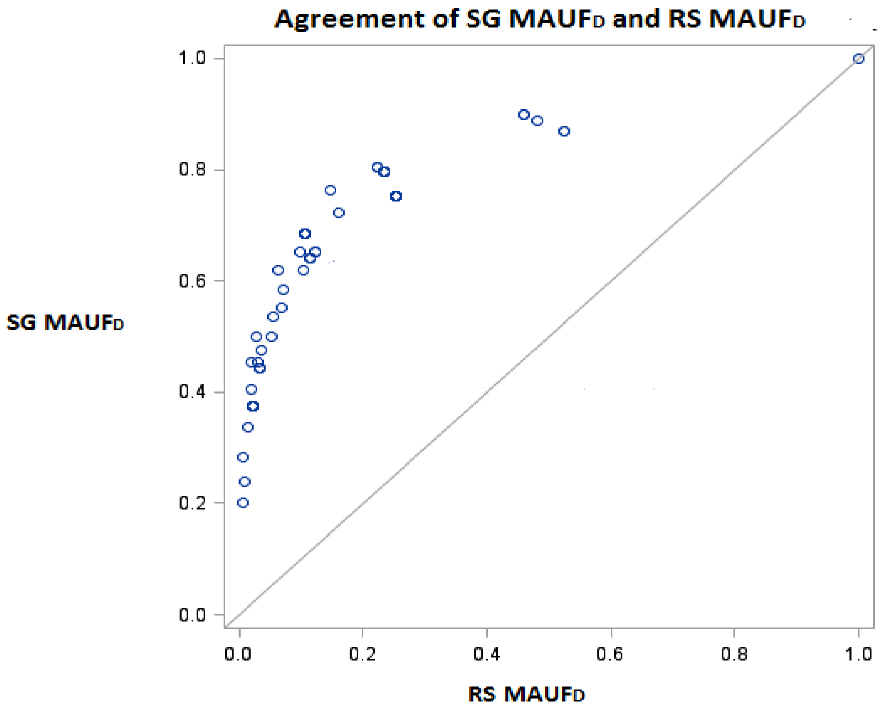
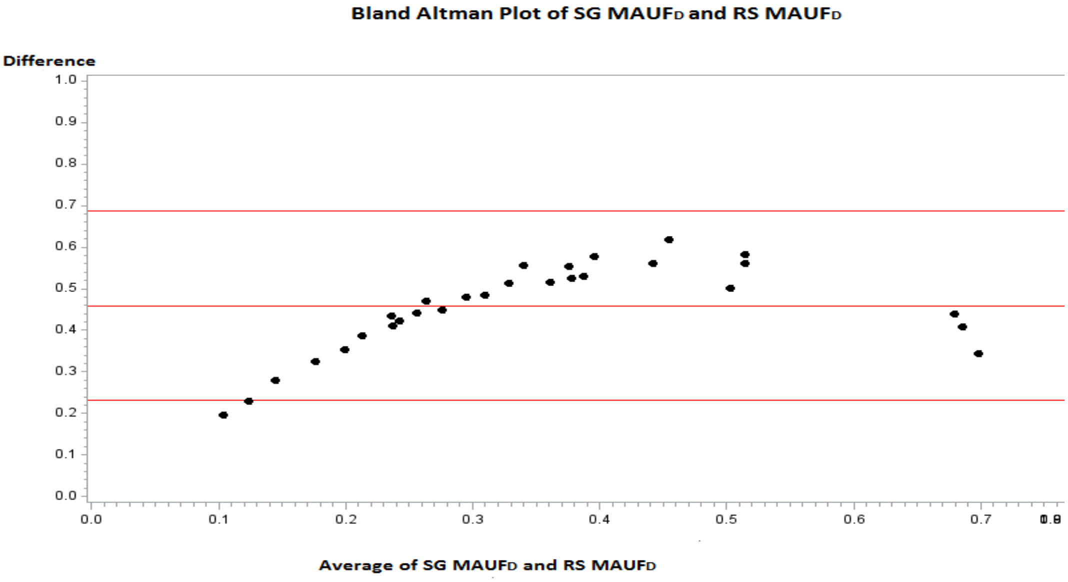


Figure 4 Scatter plot to assess agreement between the SG MAUF$_D$ and the RS MAUF$_D$ for the development sample.



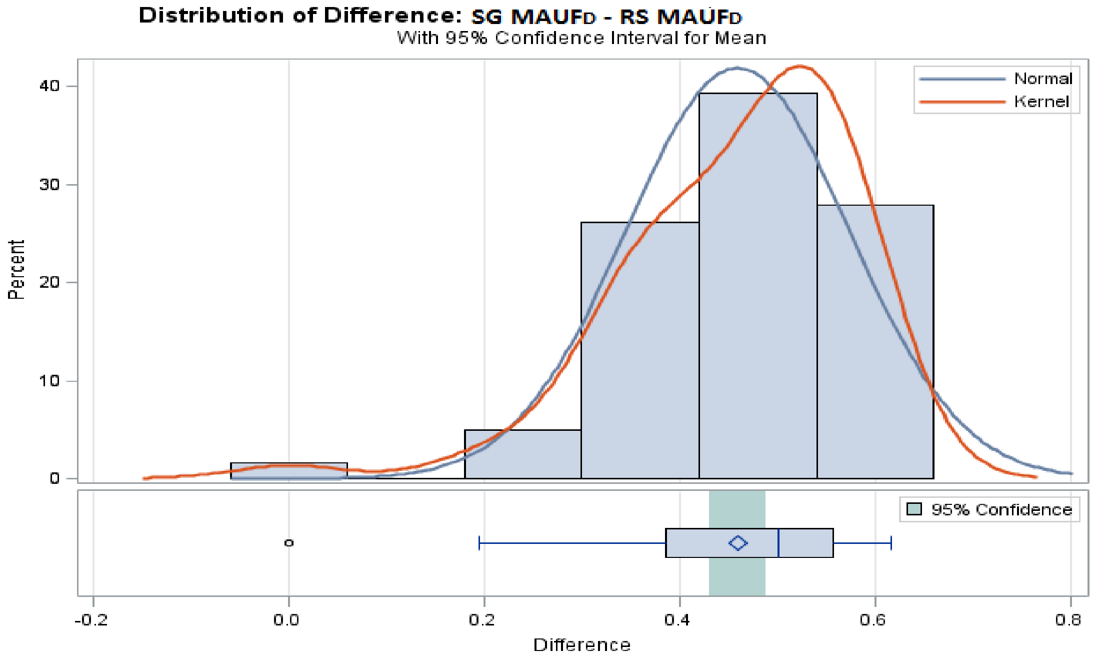
SG, Standard Gamble; RS, Rating Scale; MAUF$_D$, Multi-Attribute Utility Function Development.

Figure 5 Bland-Altman plot to assess agreement between the SG MAUF$_D$ and the RS MAUF$_D$ in the development sample



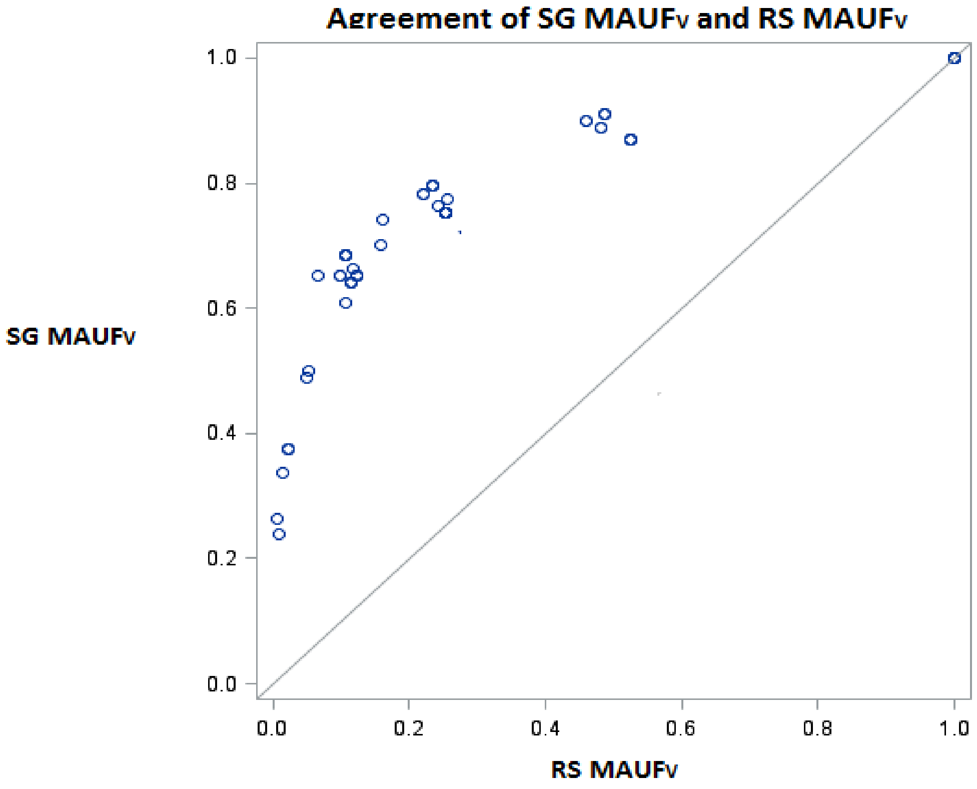
SG, Standard Gamble; RS, Rating Scale; MAUF$_D$, Multi-Attribute Utility Function Development.

Figure 6 Histogram of the differences in values between the SG MAUF$_D$ and the RS MAUF$_D$ in the development sample



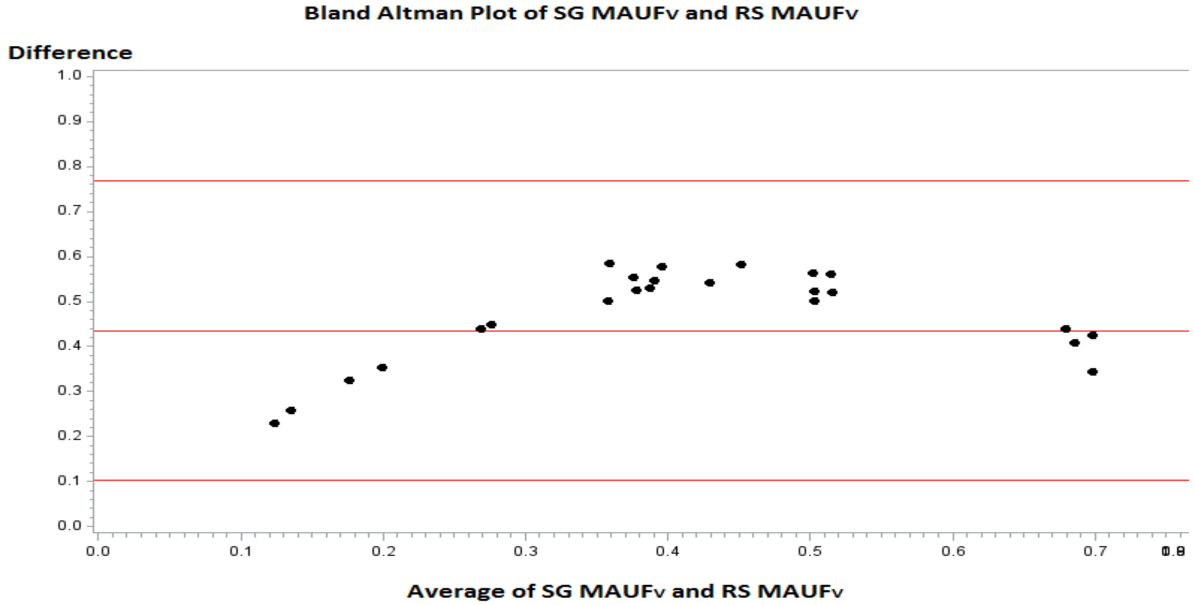
SG, Standard Gamble; RS, Rating Scale; MAUF$_D$, Multi-Attribute Utility Function Development.

Figure 7 Scatter plot to assess agreement between the SG MAUF$_V$ and the RS MAUF$_V$ for the validation sample.



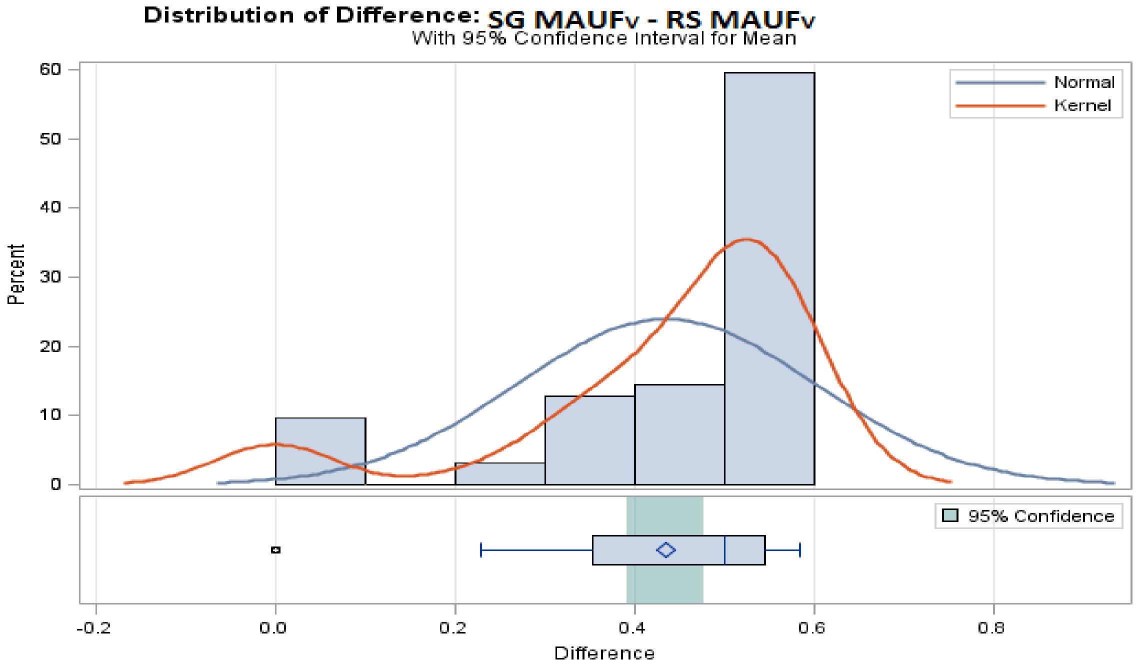
SG, Standard Gamble; RS, Rating Scale; MAUF$_D$, Multi-Attribute Utility Function Validation.

Figure 8 Bland-Altman plot to assess agreement between the SG MAUF$_V$ and the RS MAUF$_V$ in the validation sample



SG, Standard Gamble; RS, Rating Scale; MAUF$_D$, Multi-Attribute Utility Function Validation.

Figure 9 Histogram of the differences in values between the SG MAUF$_V$ and the RS MAUF$_V$ in the validation sample



SG, Standard Gamble; RS, Rating Scale; MAUF$_D$, Multi-Attribute Utility Function Validation.

Table 1 Demographic and clinical characteristics of the development and the validation sample

Characteristics	**Mean (SD) or N (%)**	
	Development Sample	**Validation Sample**
Age (y)	46.6 (11.5)	47.3 (9.97)
Women / Men	48 / 13 (79 / 21)	48 / 16 (75 / 25)
English/French*	44 / 17 (72 / 28)	14 / 50 (22 / 78)
University/College/High School	36 / 17 / 8 (59 / 28 / 13)	47 / 13 / 4 (73 / 20 / 6)
VAS health state (0-100)	66.1 (16.4)	73.0 (14.0)
PBMSI Health State		
11111	1 (2)	6 (8)
12121	5 (8)	4 (6)
12221	6 (10)	5 (8)
22111	8 (13)	9 (14)
22222	8 (13)	3 (5)
Walking		
1	23 (38)	29 (48)
2	29 (48)	30 (49)
3	9 (15)	2 (3)
Fatigue		
1	10 (16)	20 (33)
2	49 (80)	35 (57)
3	2 (3)	6 (10)
Mood		
1	29 (48)	37 (61)
2	30 (49)	22 (36)
3	2 (3)	2 (3)
Concentration		
1	20 (33)	28 (44)
2	35 (57)	34 (54)
3	6 (10)	1 (2)
Roles & Responsibilities		
1	37 (61)	19 (31)
2	21 (34)	42 (68)
3	3 (5)	1 (2)

DMT, Disease Modifying Therapy, VAS, Visual Analogue Scale.

*Language survey completed in.

Percentages were rounded to the largest integer.

Table 2 Mean standard gamble and rating scale values derived from the development sample

Item and level	SG* Mean (SD)	RS* Mean (SD)	Correlation Coefficient
Walking			
Intermediate	0.87 (0.24)	0.65 (0.22)	0.07
Worst	0.82 (0.24)	0.49 (0.24)	0.11
Fatigue			
Intermediate	0.89 (0.21)	0.62 (0.19)	-0.09
Worst	0.81 (0.25)	0.46 (0.22)	-0.11
Mood			
Intermediate	0.90 (0.20)	0.62 (0.19)	0.15
Worst	0.84 (0.22)	0.46 (0.28)	-0.29
Concentration			
Intermediate	0.91 (0.19)	0.64 (0.20)	0.13
Worst	0.88 (0.21)	0.53 (0.22)	-0.006
Roles & Responsibilities			
Intermediate	0.87 (0.20)	0.65 (0.22)	0.09
Worst	0.80 (0.22)	0.39 (0.23)	0.18
All intermediate	0.84 (0.20)	0.48 (0.20)	0.12
All worst	0.60 (0.28)	0.20 (0.22)	0.07

RS, Rating Scale; SG, Standard Gamble

*Rating Scale (RS) values were measured on a worst imaginable-best imaginable scale, Standard Gamble (SG) utilities were measured on a dead-perfect health scale.

Table 3 Concordance between the levels of difficulty between the RS and the SG in the development sample.

Standard Gamble	Rating Scale				
	Very easy	Fairly easy	Fairly difficult	Very difficult	Total
Very easy	3 (5%)	2 (3%)	1 (2%)	0 (0%)	6 (10%)
Fairly easy	2 (3%)	12 (20%)	4 (7%)	1 (2%)	19 (31%)
Fairly difficult	1 (2%)	16 (26%)	8 (13%)	0 (0%)	25 (41%)
Very difficult	1 (2%)	3 (5%)	7 (11%)	0 (0%)	11 (18%)
TOTAL	7 (12%)	33 (54%)	20 (33%)	1 (2%)	61 (100%)

Simple Kappa: 0.09 (95%CI -0.08 to 0.25); Weighted Kappa: 0.13 (-0.08 to 0.34)

Table 4 Calculation of parameters in the estimation of the PBMSI MAUF$_D$ in the development sample

Item & level	Mean utility		Mean disutility		Rescaled mean utility [a]		Rescaled mean disutility [b]	
	SG	RS	SG	RS	SG	RS	SG	RS
Walking								
1	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
2	0.87	0.53	0.13	0.47	0.28	0.29	0.72	0.71
3[c]	0.82	0.33	0.18	0.67	0.00	0.00	1.00	1.00
Fatigue								
1	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
2	0.89	0.48	0.11	0.52	0.42	0.25	0.58	0.75
3[c]	0.81	0.31	0.19	0.69	0.00	0.00	1.00	1.00
Mood								
1	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
2	0.90	0.46	0.10	0.54	0.38	0.31	0.63	0.69
3[c]	0.84	0.22	0.16	0.78	0.00	0.00	1.00	1.00
Concentration								
1	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
2	0.91	0.49	0.09	0.51	0.25	0.26	0.75	0.74
3[c]	0.88	0.31	0.12	0.69	0.00	0.00	1.00	1.00
Roles & Responsibilities								
1	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
2	0.87	0.43	0.13	0.57	0.35	0.30	0.65	0.70
3[c]	0.80	0.18	0.20	0.82	0.00	0.00	1.00	1.00

[a] Rescaled mean utility score = (person mean utility score Level X – person mean utility score Level 3) / (person mean utility score Level1 - person mean utility score Level3)

[b] Rescaled mean disutility score = 1 – (rescaled utility score)

[c] Corner states

Table 5 PBMSI MAUF$_D$ developed based on standard gamble values obtained from the development sample

Walking		**Fatigue**		**Mood**		**Concentration**		**Roles & Responsibilities**	
Level	u'_1	**Level**	u'_2	**Level**	u'_3	**Level**	u'_4	**Level**	u'_5
Single attribute mean disutilities									
1	0.00	1	0.00	1	0.00	1	0.00	1	0.00
2	0.72	2	0.58	2	0.63	2	0.75	2	0.65
3	1.00	3	1.00	3	1.00	3	1.00	3	1.00
Scaling parameter and corner state disutilities									
$c =$	0.4821	$c_1 =$	0.18	$c_3 =$	0.16	$c_5 =$	0.20		
		$c_2 =$	0.19	$c_4 =$	0.12				

Valuation function

$$\textit{PBMSI disutility}_{\text{(perfect health = 0, dead = 1)}} = (1/0.4821) \text{ x } ([1 + \{0.4821\} \text{ x } 0.18 \text{ x } u'_1] \text{ x } ([1 + \{0.4821\} \text{ x } 0.19 \text{ x } u'_2] \text{ x } ([1 + \{0.4821\} \text{ x } 0.16 \text{ x } u'_3] \text{ x } ([1 + \{0.4821\} \text{ x } 0.12 \text{ x } u'_4] \text{ x } ([1 + \{0.4821\} \text{ x } 0.20 \text{ x } u'_5] - 1)$$

$$\textit{PBMSI utility}_{\text{(dead = 0, perfect health =1)}} = 1 - \textit{PBMSI disutility}_{\text{(perfect health = 0, dead = 1)}}$$

u', disutility; c, scaling parameter; $c_{1\text{-}5}$, corner state disutility for items 1 to 5; PBMSI, Preference-Based Multiple Sclerosis Index.

Table 6 PBMSI MAUF$_D$ developed based on rating scale values obtained from the development sample

Walking		Fatigue		Mood		Concentration		Roles & Responsibilities	
Level	u'_1	**Level**	u'_2	**Level**	u'_3	**Level**	u'_4	**Level**	u'_5
Single attribute mean disutilities									
1	0.00	1	0.00	1	0.00	1	0.00	1	0.00
2	0.71	2	0.75	2	0.69	2	0.74	2	0.70
3	1.00	3	1.00	3	1.00	3	1.00	3	1.00
Scaling parameter and corner state disutilities									
c =	-0.9987	c_1 =	0.67	c_3 =	0.78	c_5 =	0.82		
		c_2 =	0.69	c_4 =	0.69				

Valuation function

PBMSI disutility (perfect health = 0, dead = 1) = (1/-0.9987) x
([1 + {-0.9987} x 0.67 x u'_1] x
([1 + {-0.9987} x 0.69 x u'_2] x
([1 + {-0.9987} x 0.78 x u'_3] x
([1 + {-0.9987} x 0.69 x u'_4] x
([1 + {-0.9987} x 0.82 x u'_5] – 1)

PBMSI utility (dead = 0, perfect health =1) = 1 – *PBMSI disutility*(perfect health = 0, dead = 1)

u', disutility; c, scaling parameter; $c_{1\text{-}5}$, corner state disutility for items 1 to 5; PBMSI, Preference-Based Multiple Sclerosis Index.

Table 7 Known-groups validity of the PBMSI and the EQ-5D index against external measures of disease severity in the validation sample.

Measure	SG MAUFv Mean (SD)	RS MAUFv Mean (SD)	EQ-5D Mean (SD)
PDDS			
0-1 (mild)	0.79 (0.15)*	0.63 (0.41)*	0.77 (0.08)*
2-3 (moderate)	0.67 (0.19)	0.23 (0.19)	0.66 (0.12)
4-5 (severe)	0.58 (0.23)	0.10 (0.08)	0.69 (0.12)
6MWT			
600 + m	0.89 (0.14)*	0.38 (0.38)*	0.78 (0.08)*
300 to 599m	0.70 (0.17)	0.22 (0.18)	0.71 (0.12)
0 to 299m	0.53 (0.25)	0.12 (0.10)	0.50 (0.20)
General Health Perception			
Excellent	0.88 (0.21)*	0.71 (0.51)*	0.77 (0.15)
Very Good	0.79 (0.15)	0.36 (0.31)	0.73 (0.13)
Good	0.70 (0.16)	0.21 (0.15)	0.72 (0.12)
Fair	0.62 (0.31)	0.31 (0.46)	0.59 (0.12)
Poor	---	---	---

PDDS, Patient Determined Disease Steps; 6MWT, 6-Minute Walk Test; PBMSI, Preference-Based Multiple Sclerosis Index; m, Meters; SD, Standard Deviation; MAUF, Multi-Attribute Utility Function.
*Linear test for trend, p-value < 0.05

CHAPTER 14: Conclusion

The overall objective of this PhD thesis was to take important steps towards developing a Preference-Based Multiple Sclerosis Index (PBMSI) for use as a global outcome in clinical and cost-effectiveness studies for MS. To operationalize this important objective, a series of manuscripts were prepared. Four manuscripts, to present background information and foundation work are published; one manuscript about revising items using patient input has been submitted. The final manuscript, representing the most critical piece of the doctoral thesis is in preparation for submission.

The first manuscript was a meta-analysis of the effects of clinical interventions on HRQL in persons with MS. In preparing this manuscript we faced a challenge of how to pool and combine the HRQL results together. As mentioned in the relevant paper, there are two types of HRQL measures: (i) health profiles, and (ii) preference-based measures. Among the included studies, heath profiles were the most commonly-used method of measuring HRQL. The most commonly-used health profile was the SF-36,[35] consisting of 36 items that are divided into 8 domains. Each domain is scored on a scale from 0 to 100, with higher scores being representative of better functioning and well-being. Health profiles provide no information on the relative importance attached to each domain. As a result, the domains cannot be combined into an overall score, nor a trade-off can be made between them when evaluating the effectiveness of interventions. For example, when a treatment had a positive effect on physical health and a negative one on mental health, it was difficult determining whether the intervention resulted in a net improvement or decline in HRQL. Preference-based measures, on the other hand, do attach a value to each described health state. Not only do these measures provide descriptive information on the various dimensions of health, but also provide a value for each one. Preference-based measures have the advantage of yielding a single number that balances gains in one domain of HRQL against losses in another. In clinical research, they can be administered pre and post intervention to evaluate the effectiveness of a treatment and to track change in HRQL over time. An additional advantage of these measures lies in their ability to be applied in health economic research. The single value produced by preference-based measures can be used to calculate QALYs and determine the cost per QALY associated with different treatment options.

In the second manuscript, we identified the domains that were most important for the quality of life of patients with MS. The Food and Drug Administration (FDA)[86] guidelines explicitly state that patients must be involved in the development of patient-reported outcomes. Therefore, in close proximity with these guidelines, we conducted semi-structured interviews with MS patients to determine the domains that should be included in the PBMSI. In this manuscript, we also assessed the content validity of existing generic preference-based measures in MS. There was no single preference-based measure that captured all domains of health relevant to MS. In fact, important domains such as fatigue and cognition were missing in these measures. The three measures that were compared in this manuscript (PGI, SF-6D and EQ-5D) were correlated with each other, as they were all administered on the same individual (n=185). Instead of the traditional paired t-test, we used generalized estimating equations (GEE) to compare between the measures. This approach allowed us to simultaneously compare between the three measures, whereas the paired t-test would have allowed comparison between only two of the measures. An effect size (ES) was calculated to compare the magnitude of the difference in standardized units.

The third manuscript was a comprehensive review of the literature on the psychometric properties of generic preference-based measures in MS. In this review, we also incorporated the data that we had on hand from the Gender and Life Impact of Multiple Sclerosis Study. Convergent validity was examined by estimating the extent to which generic preference-based measures were correlated with other measures of HRQL. The Schmidt-Hunter method, which is a random-effects model that weighs each study by its sample size, was used to combine the correlation coefficient values. To our knowledge, this was the first study that evaluated the psychometric properties of generic preference-based measures in MS. Generic preference-based measures were able to explain 36% of the variance in disease specific health profiles. A large of proportion of the variance (64%) remained unexplained, which questioned the validity of generic preference-based measures in people with MS.

In the fourth manuscript, we developed a prototype-PBMSI. Preference-based measures usually consist of one item per domain. Therefore, selecting the item that is most representative of the construct at hand can be a challenging one. Following the recent work of Brazier and colleagues[38;87] on condition specific measures, we used Rasch analysis to select one item per domain. Based on the threshold map, items that captured people at mild, moderate and severe levels of disease severity were selected for inclusion in the prototype PBMSI. Furthermore, because multi-attribute utility theory (MAUT) required the items to have a certain degree of structural independence between them, we also assessed the correlation between items. Items that were redundant or highly correlated with each other were removed. The final prototype PBMSI included 5 items, with 3 response levels each. The discriminative capacity of the response options were assessed twice: first through observation of the thresholds using Rasch analysis, and the second by mapping onto a visual analogue scale (VAS) of health rating. For each item, regression coefficient values were observed and the linear test for trend was used to assess if the response options provided the same discriminative ability within the magnitude of their capacity. The prototype PBMSI demonstrated good convergent and discriminative validity.

In the fifth manuscript, we took the 5 items and had them revised by an expert panel of clinicians and researchers in both English and French, and undergo cognitive interviewing with patients. This was a small yet important phase of the project, as several changes were made to the items in terms of the recall period and the phrasing of the items. Unlike the widely used preference-based measure EQ-5D, which asked patients to fill out the questionnaire based on their health state ‘today’, patients with MS stated that ‘over the past week’ was a more representative time frame of their health. During the qualitative review process, items were revised so that there was consistency between them in terms of phrasing and response options. Conducting cognitive interviews with patients not only helped increase the accuracy of reporting, but also helped reduce measurement error in the PBMSI.

In the sixth manuscript we elicited patient preferences for the different items in the PBMSI using two standard valuation methods; the standard gamble (SG) and the rating scale (RS). As far as our research goes, this was the first study to administer the SG and the RS in patients with MS. The SG is directly based on the axioms of utility theory and involves decision making under risk and uncertainty. As has been demonstrated in the sixth manuscript, there are challenges for utilizing this technique in patients with MS. The SG was not only a difficult technique for patients to understand, but also patients did not want to take a risk of dying in return for an improvement in health. All of these raise doubt on the validity of the SG technique in patients with MS. The RS, on the other hand, was found to be a much simpler method for patients to understand and use. One of the main criticisms with the RS is that it does not include an element of choice or decision making under uncertainty. However, despite this limitation, the RS has a long-standing history in economic research.[38] It was first identified as a possible measure for use in economic evaluation purposes more than three decades ago and has now become one of the most widely used measures for these purposes.[38] It has been used to develop preference-based measures like the Quality of Well-Being Scale and the 15-D. In the end, therefore, we are of the strong opinion that compared to the SG; the RS appears a more appropriate valuation method in patients with MS and hence should be used in the development of the MAUF.

During the third year of my PhD studies, Versteegh and colleagues[88] derived a MS specific preference-based measure from the Multiple Sclerosis Impact Scale-29 (MSIS-29). However, there are important differences in the methods used to develop the PBMSI and those used by Versteegh and colleagues to develop a scoring algorithm for selected items from the MSIS-29. First, the two samples were different. To develop the domains of the PBMSI, we purposely sampled patients with MS (n=185) diagnosed after 1995, during the era of Magnetic Resonance Imaging (MRI) technology and availability of disease modifying drugs. Prior to 1995, diagnosis was mainly based on abnormal neurological signs and symptoms, and management was aimed at reducing the severity of acute relapses through the use of steroids. However, over the past 20 years MRI has played a pivotal role in the early diagnosis of the disease. Furthermore, the introduction of disease modifying therapies has allowed for a better management of the progression of MS. This was important because this is the population faced with treatment decisions.

The MSIS-29, on the other hand, was developed before this era with more severe patients (n=30). In fact, more than 60% of the sample were wheelchair bound or ambulating with an aid, were not working, and had progressive MS. Moreover, among the 8 items that Versteegh and colleagues[88] selected for inclusion in their preference-based measure, only 2 items (work and concentration) were identified as important by our sample of MS patients. Walking, fatigue and mood were missing. In deciding the right outcome measure for a study, it is essential to select one with items that are important for the health condition.

Peferences for the PBMSI items were obtained from patients with MS whereas the one based on the MSIS-29 obtained preference values from the general population. We found the the SG, a decision based approach incorporating uncertainity and risk of death, very difficult for our population to understand. The MSIS-29 based measure, used the Time-Trade-Off (TTO) methods, also based on decision but without this risk element, although the trade off is years of life (death). The experience of the patients with both SG and RS, support the use of the RS to capture the impact of MS.

Clinical and Economic Applications of the PBMSI

As it stands now, the PBMSI is ready for further testing of its applications in (a) clinical practice, (b) clinical research and (c) economic research.

In clinical practice, clinicians need measures that are easy to score and simple to administer. The scoring algorithm of the PBMSI could be simplified so that a value of 0, 1, and 2 is assigned to the first, second and third response levels, respectively. As the preference weights of the worst and intermediate levels did not vastly differ across items, a simple un-weighted sum would be valid producing a quick profile from 0 to 10 of how the patient is. The PBMSI could help clinicians evaluate the overall impact of a new treatment on patients' health and track change over time. The items could also be provided to patients for self-monitoring of their disease. This approach could be tested in targeted research asking clinicians to use and comment on the acceptability and feasibility of use in their practice.

In clinical research, the PBMSI could be employed to evaluate the clinical effectiveness of different interventions in MS. As the PBMSI attaches explicit weights to the various dimensions of health, a single index that ranges from 0 (death) to 1 (perfect health) can be produced. The PBMSI will overcome one of the major challenges concerning health profiles, such as the SF-36 - that domains cannot be combined into an overall indicator of health. For example, in comparing two types of therapies (Therapy A and B), one may perform better against one domain (e.g. physical health), but worse against another (e.g. mental health). At the end of a clinical trial, health profiles would not be able to provide any information on whether a therapy was most-effective or not. However, preference-based measures would easily be able to trade off gains in one domain against losses in another, and determine the overall net effect of the intervention on HRQL.[3;75;89;90]

The PBMSI can be used for economic evaluation to contrast different interventions for people with MS. For example, the PBMSI would be a good measure for contrasting physical therapy vs.

Fampridine[91;92] for improving gait. Physical therapy has many benefits, through exercise, not only for gait, but also for fatigue, depression, etc. A drug like Fampridine is targeted specifically to conduction of action potentials in demyelinated nerve fibres and its effects are on improvement of performance shown by increased speed of walking.[91;92] However, this drug may have negative effects not captured by measuring gait alone, and hence a measure like the PBMSI would detect these differences in therapeutic approach. It is only reasonable in this context that people with MS are the ones valuing the disability dimensions.

For the economic purpose of allocating resources across the population, a disease focused approach would not be helpful and hence the use of general population weights in creating metrics that when linked to life expectancy yield a quality adjusted life year (QALY). However, there are many other methods for adjusting survival for quality of life (QAS)[93-97] which may yield important information for contrasting treatments within a specific disease context. A challenge has often been to get the correct value for the adjustment variable (Q); here for MS, the PBMSI would provide the Q.

Directions for Future Research

I am involved in a trial of exercise for people with MS (MSTEP©)[98] in which the PBMSI is part of the assessment package. This trial involves 240 people tracked over 2 years. In addition, 120 people from this study provided preferences using the RS. A final MAUF for the PBMSI will be developed based on this large sample of MS patients. In addition to the MAUF, alternative methods of modeling the data such as generalized estimating equations (GEE) will be employed. The predictive validity of the two mathematical models, GEE and MAUF, will be compared with each other. Furthermore, this data set will provide rich data for further validation cross-sectionally and longitudinally. Within the next 2 to 3 years, the validation process should be completed, yielding sufficient data to make a decision about its future in MS research and clinical practice.

REFERENCE LIST

(1) Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG. Multiple sclerosis. *N Engl J Med* 2000;343:938-952.

(2) Beck CA, Metz LM, Svenson LW, Patten SB. Regional variation of multiple sclerosis prevalence in Canada. *Mult Scler* 2005;11:516-519.

(3) Kind P. Values and valuation in the measurement of HRQoL. In: Fayers P, Hays D, eds. *Assessing quality of life in clinical trials*. 2 ed. New York: Oxford University Press Inc.; 2005;391-404.

(4) Smith KJ, McDonald WI. The pathophysiology of multiple sclerosis: the mechanisms underlying the production of symptoms and the natural history of the disease. *Philos Trans R Soc Lond B Biol Sci* 1999;354:1649-1673.

(5) Dean M.Wingerchuk, John H.Noseworthy, Claudia F.Lucchinetti. Multiple Sclerosis: Current Pathophysiological Concepts. *Laboratory Investigation* 2001;81:263-281.

(6) Fred D.Lublin, Stephen C.Reingold, National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Defining the clinical course of multiple sclerosis: Results of an international survey. *Neurology* 1996;46:907-911.

(7) Edward J.Fox. Immunopathology of multiple sclerosis. *Neurology* 2004;63:S3-7.

(8) Lublin FD. New multiple sclerosis phenotypic classification. *Eur Neurol* 2014;72 Suppl 1:1-5.

(9) Linda Coulthard-Morris. Clinical and rehabilitation outcome measures. In: Jack S.Burks, Kenneth P.Johnson, eds. *Multiple Sclerosis Diagnosis, medical Management, and Rehabilitation*. New York: Demos Medical Publishing; 2000;221-290.

(10) Kenneth P.Johnson. Therapy of Relapsing Forms. In: Jack S.Burks, Kenneth P.Johnson, eds. *Multiple Sclerosis: Diagnosis, Medical Management, and Rehabilitation*. New York: Demos Medical Publishing; 2000;167-175.

(11) Paul W.O'Connor. Reason for hope: the advent of disease-modifying therapies in multiple sclerosis. *Canadian Medical Association Journal* 2000;162:83.

(12) Lawrence D.Jacobs, Roy W.Beck, Jack H.Simon et al. Intramuscular Interferon Beta-1A Therapy Initiated during a First Demyelinating Event in Multiple Sclerosis. *The New England Journal of Medicine* 2000;343:898-904.

(13) Kieseier BC, Hartung HP. Current disease-modifying therapies in multiple sclerosis. *Semin Neurol* 2003;23:133-146.

(14) Tanasescu R, Ionete C, Chou IJ, Constantinescu CS. Advances in the treatment of relapsing-remitting multiple sclerosis. *Biomed J* 2014;37:41-49.

(15) Polman CH, O'Connor PW, Havrdova E et al. A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *N Engl J Med* 2006;354:899-910.

(16) Rudick RA, Stuart WH, Calabresi PA et al. Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *N Engl J Med* 2006;354:911-923.

(17) Cohen BA, Mikol DD. Mitoxantrone treatment of multiple sclerosis Safety considerations. *Neurology* 2004;63:S28-S32.

(18) Marriott JJ, Miyasaki JM, Gronseth G, O'Connor PW. Evidence Report: The efficacy and safety of mitoxantrone (Novantrone) in the treatment of multiple sclerosis: Report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. *Neurology* 2010;74:1463-1470.

(19) Mehling M, Kappos L, Derfuss T. Fingolimod for multiple sclerosis: mechanism of action, clinical outcomes, and future directions. *Curr Neurol Neurosci Rep* 2011;11:492-497.

(20) Cohen JA, Barkhof F, Comi G et al. Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. *N Engl J Med* 2010;362:402-415.

(21) O'Connor PW, Li D, Freedman MS et al. A Phase II study of the safety and efficacy of teriflunomide in multiple sclerosis with relapses. *Neurology* 2006;66:894-900.

(22) Freedman MS, Wolinsky JS, Wamil B et al. Teriflunomide added to interferon-beta in relapsing multiple sclerosis: a randomized phase II trial. *Neurology* 2012;78:1877-1885.

(23) Confavreux C, Li DK, Freedman MS et al. Long-term follow-up of a phase 2 study of oral teriflunomide in relapsing multiple sclerosis: safety and efficacy results up to 8.5 years. *Mult Scler* 2012;18:1278-1289.

(24) Kappos L, Gold R, Miller DH et al. Efficacy and safety of oral fumarate in patients with relapsing-remitting multiple sclerosis: a multicentre, randomised, double-blind, placebo-controlled phase IIb study. *Lancet* 2008;372:1463-1472.

(25) Kappos L, Gold R, Miller DH et al. Effect of BG-12 on contrast-enhanced lesions in patients with relapsing--remitting multiple sclerosis: subgroup analyses from the phase 2b study. *Mult Scler* 2012;18:314-321.

(26) Polman C, Barkhof F, Sandberg-Wollheim M, Linde A, Nordle O, Nederman T. Treatment with laquinimod reduces development of active MRI lesions in relapsing MS. *Neurology* 2005;64:987-991.

(27) Grima DT, Torrance GW, Francis G, Rice G, Rosner AJ, Lafortune L. Cost and health related quality of life consequences of multiple sclerosis. *Mult Scler* 2000;6:91-98.

(28) Karampampa K, Gustavsson A, Miltenburger C, Kindundu CM, Selchen DH. Treatment experience, burden, and unmet needs (TRIBUNE) in multiple sclerosis: the costs and utilities of MS patients in Canada. *Journal of population therapeutics and clinical pharmacology= Journal de la therapeutique des populations et de la pharamcologie clinique* 2011;19:e11-e25.

(29) Breslow L. A quantitative approach to the World Health Organization definition of health: physical, mental and social well-being. *Int J Epidemiol* 1972;1:347-355.

(30) Ware JE, Jr. Standards for validating health measures: definition and content. *J Chronic Dis* 1987;40:473-480.

(31) Wood-Dauphinee S. Assessing quality of life in clinical research: from where have we come and where are we going? *J Clin Epidemiol* 1999;52:355-363.

(32) Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual model of health-related quality of life. *J Nurs Scholarsh* 2005;37:336-342.

(33) Guyatt GH, Veldhuyzen Van Zanten SJ, Feeny DH, Patrick DL. Measuring quality of life in clinical trials: a taxonomy and review. *CMAJ* 1989;140:1441-1448.

(34) Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;118:622-629.

(35) Ware JE, Kosinski M, Dewey JE, Gandek B. *SF-36 health survey: manual and interpretation guide*. Quality Metric Inc., 2000.

(36) Ware JE, Jr. The SF-36 Health Survey. In: Spilker B, ed. *Quality of Life and Pharmaeconomics in Clinical Trials*. 2 ed. Philadelphia: Lippincott-Raven Publishers; 1996;337-345.

(37) Kind P. Values and valuation in the measurement of HRQoL. In: Fayers P, Hays D, eds. *Assessing quality of life in clinical trials*. 2 ed. New York: Oxford University Press Inc.; 2005;391-404.

(38) Brazier J. *Measuring and valuing health benefits for economic evaluation*. Oxford University Press, 2007.

(39) Poissant L, Mayo NE, Wood-Dauphinee S, Clarke AE. The development and preliminary validation of a Preference-Based Stroke Index (PBSI). *Health Qual Life Outcomes* 2003;1:43.

(40) Feeny D. Preference-based measures: utility and quality-adjusted life years. *Assessing quality of life in clinical trials* 2005;405-429.

(41) Torrance GW. Designing and conducting cost-utility analyses. *Quality of life and pharmacoeconomics in clinical trials Philadelphia: Lippincott-Raven Publishers* 1996;1105-1121.

(42) Hawthorne G, Richardson J. Measuring the value of program outcomes: a review of multiattribute utility measures. 2001.

(43) Torrance GW. Measurement of health state utilities for economic appraisal: a review. *Journal of health economics* 1986;5:1-30.

(44) Bennett KJ, Torrance GW. Measuring health state preferences and utilities: rating scale, time trade-off, and standard gamble techniques. *Quality of life and pharmacoeconomics in clinical trials* 1996;2:253-265.

(45) Patrick DL, Erickson P. Applications of health status assessment to health policy. *Quality of Life and Pharmacoeconomics in Clinical Trials Second ed Philadelphia: Lippincott-Raven Publishers* 1996;717-727.

(46) Bleichrodt H. A new explanation for the difference between time tradeGÇÉoff utilities and standard gamble utilities. *Health economics* 2002;11:447-456.

(47) Van Osch SM, Wakker PP, Van Den Hout WB, Stiggelbout AM. Correcting biases in standard gamble and time tradeoff utilities. *Medical Decision Making* 2004;24:511-517.

(48) Poissant L. The Development of a Preference-based Health Index for Stroke. 2002.

(49) Torrance GW. Utility approach to measuring health-related quality of life. *Journal of chronic diseases* 1987;40:593-600.

(50) O'Leary JF, Fairclough DL, Jankowski MK, Weeks JC. Comparison of Time-tradeoff Utilities and Rating Scale Values of Cancer Patients and Their Relatives Evidence for a Possible Plateau Relationship. *Medical Decision Making* 1995;15:132-137.

(51) Krabbe PF, Tromp N, Ruers TJ, van Riel PL. Are patientsGÇÖ judgments of health status really different from the general population. *Health Qual Life Outcomes* 2011;9:31.

(52) Kaplan RM, Anderson JP, Wu AW, Mathews WC, Kozin F, Orenstein D. The Quality of Well-being Scale: applications in AIDS, cystic fibrosis, and arthritis. *Medical care* 1989;S27-S43.

(53) Kaplan RM, Ganiats TG, Sieber WJ, Anderson JP. The Quality of Well-Being Scale: critical similarities and differences with SF-36. *International Journal for Quality in Health Care* 1998;10:509-520.

(54) McDowell I. *Measuring health: a guide to rating scales and questionnaires*. Oxford University Press, 2006.

(55) Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. *Oper Res* 1982;30:1043-1069.

(56) Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. *Medical care* 1996;34:702-722.

(57) Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. *J Clin Oncol* 1992;10:923-928.

(58) Feeny D, Furlong W, Torrance GW et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care* 2002;40:113-128.

(59) Sintonen H, Pekurinen M. A fifteen-dimensional measure of health-related quality of life (15D) and its applications. *Quality of life assessment: key issues in the 1990s*. Springer; 1993;185-195.

(60) Sintonen H. The 15D instrument of health-related quality of life: properties and applications. *Annals of medicine* 2001;33:328-336.

(61) Kind P, Brooks R, Rabin R. *EQ-5D concepts and methods:: a developmental history*. Springer, 2006.

(62) Kind P. The EuroQoL instrument: an index of health-related quality of life. *Quality of life and pharmacoeconomics in clinical trials* 1996;2:191-201.

(63) Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. *Medical Decision Making* 2001;21:7-16.

(64) Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Medical care* 2005;43:203-220.

(65) Johnson JA, Luo N, Shaw JW, Kind P, Coons SJ. Valuations of EQ-5D health states: are the United States and United Kingdom different? *Medical care* 2005;43:221-228.

(66) Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. *Qual Life Res* 1999;8:209-224.

(67) Richardson J, Atherton Day N, Peacock S, Iezzi A. Measurement of the quality of life for economic evaluation and the Assessment of Quality of Life (AQoL) Mark 2 instrument. *Australian Economic Review* 2004;37:62-88.

(68) Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ* 2002;21:271-292.

(69) Gudex C. The descriptive system of the EuroQOL instrument. *EQ-5D concepts and methods: a developmental history*. Springer; 2005;19-27.

(70) Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI-«): concepts, measurement properties and applications. *Health and quality of life outcomes* 2003;1:54.

(71) Dolan P. Whose preferences count? *Med Decis Making* 1999;19:482-486.

(72) Kind P, Lafata JE, Matuszewski K, Raisch D. The use of QALYs in clinical and patient decision-making: issues and prospects. *Value Health* 2009;12 Suppl 1:S27-S30.

(73) Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. *Qual Life Res* 2003;12:599-607.

(74) Food and Drug Administration. Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register* 2009;74:65132-65133.

(75) Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. *Measuring and valuing health benefits for economic evaluation*. New York: Oxford University Press Inc., 2007.

(76) Brazier J, Akehurst R, Brennan A et al. Should patients have a greater role in valuing health states? *Appl Health Econ Health Policy* 2005;4:201-208.

(77) Tengs TO, Wallace A. One thousand health-related quality-of-life estimates. *Med Care* 2000;38:583-637.

(78) Ratcliffe J, Brazier J, Palfreyman S, Michaels J. A comparison of patient and population values for health states in varicose veins patients. *Health Econ* 2007;16:395-405.

(79) McPherson K, Myers J, Taylor WJ, McNaughton HK, Weatherall M. Self-valuation and societal valuations of health state differ with disease severity in chronic and disabling conditions. *Med Care* 2004;42:1143-1151.

(80) Kind P. Beyond economic evaluation: an appropriate scoring system for EQ-5D based on real values for health [abstract]Kind P. *Quality of Life Research* 2010;19 (supp.1):21

(81) Insinga RP, Fryback DG. Understanding differences between self-ratings and population ratings for health in the EuroQOL. *Qual Life Res* 2003;12:611-619.

(82) McPherson K, Myers J, Taylor WJ, McNaughton HK, Weatherall M. Self-valuation and societal valuations of health state differ with disease severity in chronic and disabling conditions. *Med Care* 2004;42:1143-1151.

(83) Peeters Y, Stiggelbout AM. Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities. *Value Health* 2010;13:306-309.

(84) Noseworthy JH, Vandervoort MK, Wong CJ, Ebers GC. Interrater variability with the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) in a multiple sclerosis clinical trial. The Canadian Cooperation MS Study Group. *Neurology* 1990;40:971-975.

(85) Kuspinar A, Rodriguez AM, Mayo NE. The effects of clinical interventions on health-related quality of life in multiple sclerosis: a meta-analysis. *Mult Scler* 2012;18:1686-1704.

(86) Food and Drug Administration. Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register* 2009;74:65132-65133.

(87) Young T, Yang Y, Brazier JE, Tsuchiya A, Coyne K. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. *Quality of Life Research* 2009;18:253-265.

(88) Versteegh MM, Leunis A, Uyl-de Groot CA, Stolk EA. Condition-specific preference-based measures: benefit or burden? *Value in Health* 2012;15:504-513.

(89) Poissant L, Mayo NE, Wood-Dauphinee S, Clarke AE. The development and preliminary validation of a Preference-Based Stroke Index (PBSI). *Health Qual Life Outcomes* 2003;1:43.

(90) Feeny DH, Torrance GW. Incorporating utility-based quality-of-life assessment measures in clinical trials. Two examples. *Med Care* 1989;27:S190-S204.

(91) Jensen H, Ravnborg M, Mamoei S, Dalgas U, Stenager E. Changes in cognition, arm function and lower body function after Slow-Release Fampridine treatment. *Mult Scler* 2014.

(92) Hobart J, Blight AR, Goodman A, Lynn F, Putzki N. Timed 25-foot walk: direct evidence that improving 20% or greater is clinically meaningful in MS. *Neurology* 2013;80:1509-1517.

(93) Wang JD. *Basic principles and practical applications in epidemiological research*. Singapore: World Scientific, 2007.

(94) Revicki DA, Feeny D, Hunt TL, Cole BF. Analyzing oncology clinical trial data using the Q-TWiST method: clinical importance and sources for health state preference data. *Qual Life Res* 2006;15:411-423.

(95) Schwartz CE, Cole BF, Gelber RD. Measuring patient-centered outcomes in neurologic disease. Extending the Q-TWiST method. *Arch Neurol* 1995;52:754-762.

(96) Hwang JS, Tsauo JY, Wang JD. Estimation of expected quality adjusted survival by cross-sectional survey. *Stat Med* 1996;15:93-102.

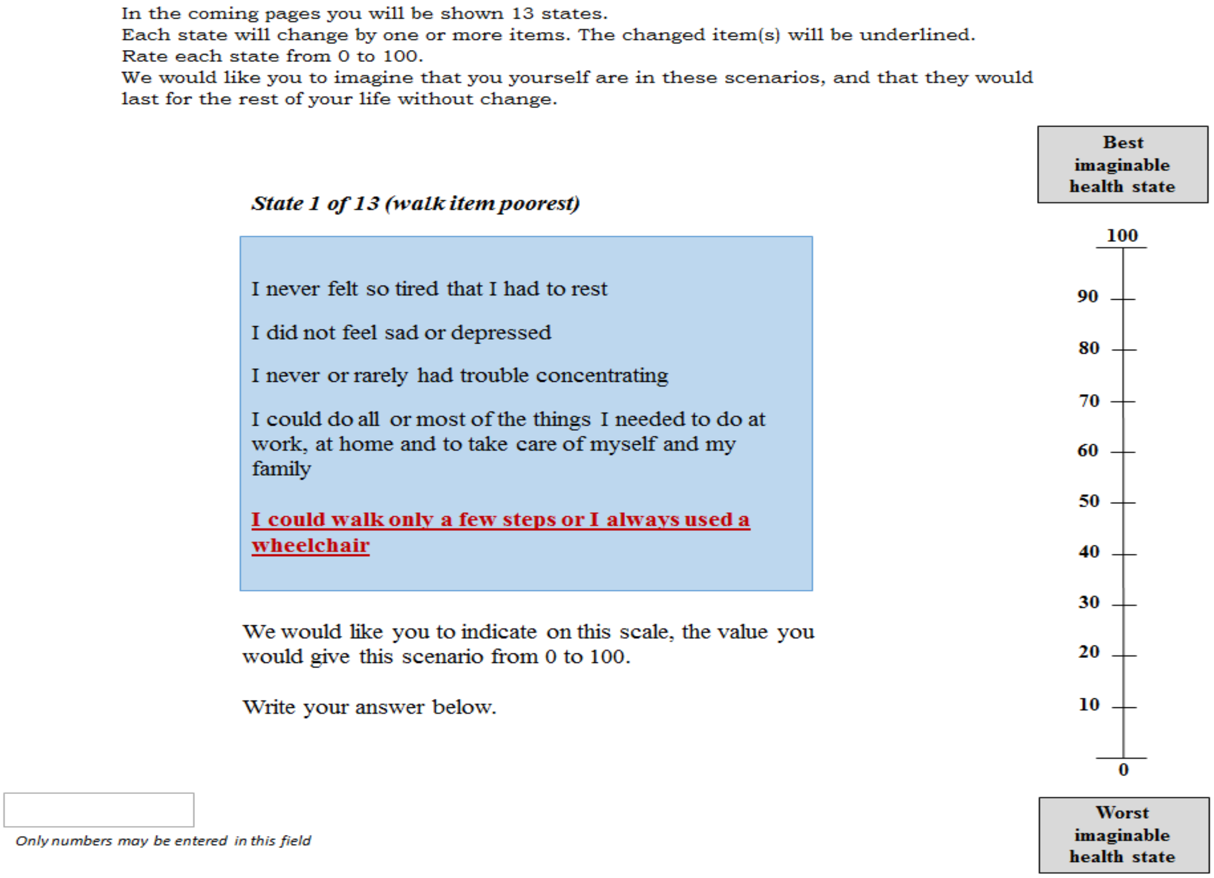
(97) Goldhirsch A, Gelber RD, Simes RJ, Glasziou P, Coates AS. Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *Journal of Clinical Oncology* 1989;7:36-44.

(98) Mayo NE, Bayley M, Duquette P, Lapierre Y, Anderson R, Bartlett S. The role of exercise in modifying outcomes for people with multiple sclerosis: a randomized trial. *BMC neurology* 2013;13:69.

APPENDICES

APPENDIX 1

Online Survey: Rating Scale Method



In the coming pages you will be shown 13 states.
Each state will change by one or more items. The changed item(s) will be underlined.
Rate each state from 0 to 100.
We would like you to imagine that you yourself are in these scenarios, and that they would last for the rest of your life without change.

State 1 of 13 (walk item poorest)

I never felt so tired that I had to rest

I did not feel sad or depressed

I never or rarely had trouble concentrating

I could do all or most of the things I needed to do at work, at home and to take care of myself and my family

I could walk only a few steps or I always used a wheelchair

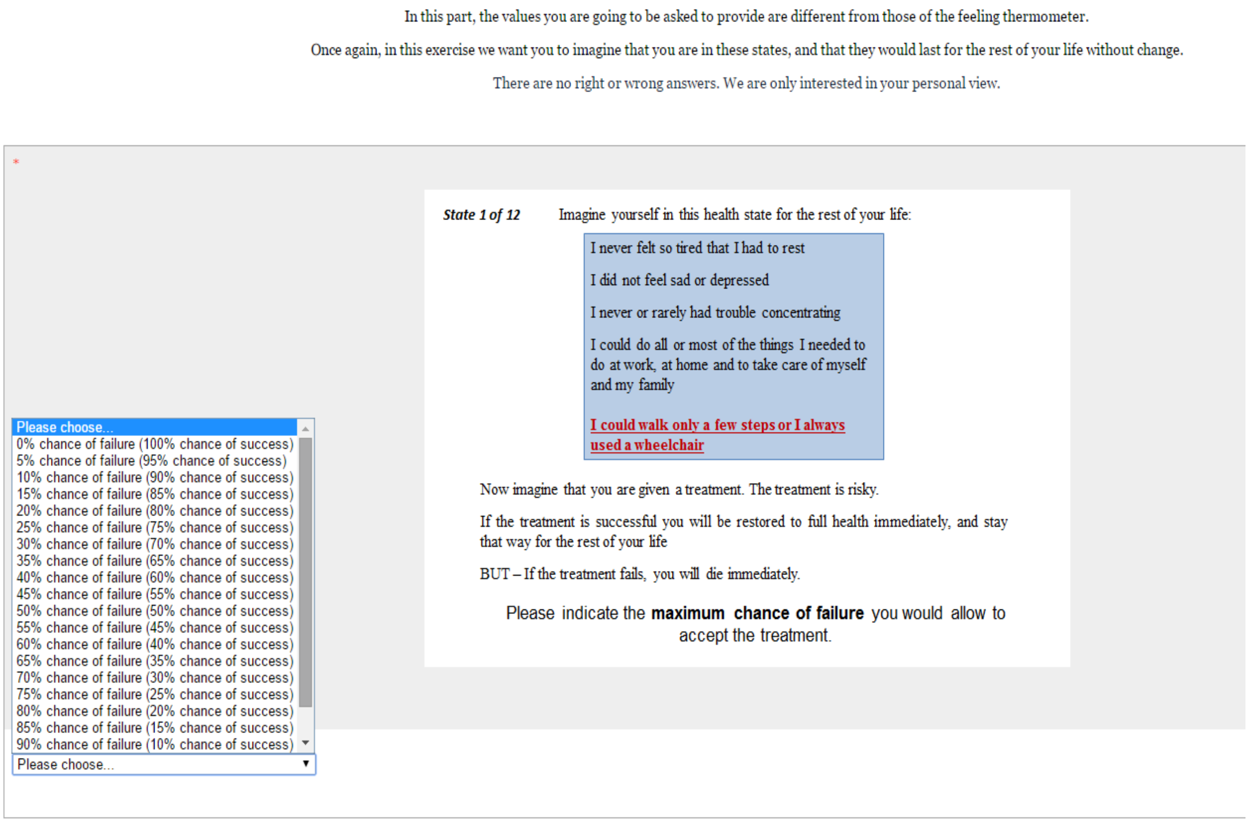
We would like you to indicate on this scale, the value you would give this scenario from 0 to 100.

Write your answer below.

Only numbers may be entered in this field

APPENDIX 2

Online Survey: Standard Gamble Method



In this part, the values you are going to be asked to provide are different from those of the feeling thermometer.

Once again, in this exercise we want you to imagine that you are in these states, and that they would last for the rest of your life without change.

There are no right or wrong answers. We are only interested in your personal view.

*

State 1 of 12 Imagine yourself in this health state for the rest of your life:

I never felt so tired that I had to rest

I did not feel sad or depressed

I never or rarely had trouble concentrating

I could do all or most of the things I needed to do at work, at home and to take care of myself and my family

I could walk only a few steps or I always used a wheelchair

Now imagine that you are given a treatment. The treatment is risky.

If the treatment is successful you will be restored to full health immediately, and stay that way for the rest of your life

BUT – If the treatment fails, you will die immediately.

Please indicate the **maximum chance of failure** you would allow to accept the treatment.

Please choose...
0% chance of failure (100% chance of success)
5% chance of failure (95% chance of success)
10% chance of failure (90% chance of success)
15% chance of failure (85% chance of success)
20% chance of failure (80% chance of success)
25% chance of failure (75% chance of success)
30% chance of failure (70% chance of success)
35% chance of failure (65% chance of success)
40% chance of failure (60% chance of success)
45% chance of failure (55% chance of success)
50% chance of failure (50% chance of success)
55% chance of failure (45% chance of success)
60% chance of failure (40% chance of success)
65% chance of failure (35% chance of success)
70% chance of failure (30% chance of success)
75% chance of failure (25% chance of success)
80% chance of failure (20% chance of success)
85% chance of failure (15% chance of success)
90% chance of failure (10% chance of success)

Please choose...