A comparison of health utility measures for the evaluation of multiple sclerosis treatments

J D Fisk, M G Brown, I S Sketris, L M Metz, T J Murray, K J Stadnyk

J Neurol Neurosurg Psychiatry 2005;76:58–63. doi: 10.1136/jnnp.2003.017897

See end of article for authors' affiliations

Correspondence to:
Dr J Fisk, Abbie J. Lane
Building, Queen Elizabeth
II Health Sciences Centre,
Veterans Memorial Lane,
Halifax, NS, B3H 2E3;
john.fisk@cdha.nshealth.ca

Received 1 May 2003
In revised form
4 March 2004
Accepted 2 April 2004

Objectives: To evaluate the practical application and psychometric properties of three health utility measures in a sample of MS patients with a broad range of neurological disability as measured by the Extended Disability Status Scale (EDSS).

Methods: Patients randomly selected from two MS clinic registries were assessed using standard clinical methods and completed three generic measures of health utility (EQ-5D, HUI Mark III, SF-6D). The proportion of missing data, test/retest reliability, and construct validity of each health utility measure were examined.

Results: The assessments were completed by 187 patients. Less than 10% of data were missing for the subscales of the SF-6D (<3.2%), HUI Mark III (<1.6%), and EQ-5D (≤7.5%). Severely disabled patients were more likely to omit physical function questions for the SF-6D (20%), and EQ-5D (43%). Retest reliability for the SF-6D (ICC=0.83), EQ-5D (ICC=0.81), and HUI Mark III (ICC=0.87) were adequate for population surveys. Correlations between assessment of clinical function and each health utility measure were strongest for the HUI Mark III (HUI Mark III EDSS $\rho = -0.77$, HUI Mark III ambulation index $\rho = -0.76$, HUI Mark III timed 25 foot walk $\rho = -0.73$, HUI Mark III nine hole peg test $\rho = -0.65$).

Conclusions: The health utility measures were generally feasible and reliable but the HUI Mark III demonstrated highest concordance with the EDSS across the full range of neurological disability. Of the three measures studied, the HUI Mark III may be the most appropriate for cost effectiveness evaluations of MS therapies.

Multiple sclerosis (MS) is the most common disabling neurological disease of young adults, causing enormous physical, economic, and psychosocial burden to individuals, their families, and society. While new drugs may, for the first time, modify the natural history of MS,^{1–2} clinical trial efficacy evidence is subject to trial design limitations, is short term (2–3 years) relative to the long term natural history of MS, and is less convincing when disability progression rather than relapse rate is considered as the outcome of interest.^{1–4} New MS drugs appear to reduce the frequency of symptom relapse by about 30%,¹ but the limited economic consequences of relapses has resulted in relatively high cost effectiveness estimates for these treatments.^{4–6} A reduced rate of disability progression in MS has much greater potential economic significance.^{4–8} Additional credible evidence of treatment efficacy and/or effectiveness regarding slowed disability progression would assist in the debate on the cost effectiveness and the public sector funding of these expensive treatments.⁹

To date, the Expanded Disability Status Scale (EDSS)¹⁰ remains the primary clinical outcome measure for MS related disability, despite recent introduction, validation, and use of the MS Functional Composite scale as an alternative.¹¹ Although the EDSS has the significant advantage of comparability across studies, it is far from an ideal instrument for use in economic analyses of outcomes, owing to limitations in its scaling properties.¹² At higher levels of disability, the EDSS primarily evaluates disability in ambulation and mobility. At lower levels of disability, the lack of attention to common symptoms such as pain and fatigue may limit the ability of the EDSS to detect small but clinically important changes, including changes that may affect employment and hence the cost of disability.

Alternatives to clinician based assessments are self report instruments of health related quality of life (HRQoL). A wide

variety of such instruments exists, however, and appropriate instrument selection for the specific application is essential.¹³ Of the generic profile measures of HRQoL, the Medical Outcomes Study Short Form-36 (SF-36)¹⁴ is most commonly used in studies of MS. While there is some evidence that the SF-36 subscales are responsive to early effects of MS treatment,¹⁵ such data are limited. Criticisms of the appropriateness of the SF-36 and other generic HRQoL scales for use by MS patients have centred on the lack of item content considered important for MS patients, and the result has been a proliferation of MS specific HRQoL assessments.^{16–18} While the combined use of both disease specific and generic HRQoL profile measures has been recommended for measuring health outcomes,¹³ neither have the measurement properties necessary for use in cost effectiveness and cost/utility analyses.¹⁹

Health utility measures fall within the broad construct of HRQoL measures, but rather than producing multiple values that represent different domains or concepts of subjective health, a single value is derived to represent an individual's health state on a scale from 0 (dead) to 1 (perfect health). The principal advantage of health utility measures is that they are designed to be interval measures that reflect person preferences. This allows for comparisons across studies, patient populations, and interventions, and for calculation of quality adjusted life years (QALYs) for cost effectiveness evaluations and resource allocation policy decisions.²⁰ To date, measurement of health utility in studies of MS are limited and the few published studies have focused on the

Abbreviations: EDSS, Extended Disability Status Scale; HRQoL, health related quality of life; MS, multiple sclerosis; HUI Mark III, Health Utilities Index Mark III; QALY, quality adjusted life year; SF-36, Medical Outcomes Study Short Form 36

costs of relapses rather than on the costs of disability progression.^{5 6 8}

While the importance of HRQoL measures is well recognised, there is no gold standard health utility measure, and comparisons of available measures are rare.^{21 22} Selecting the most appropriate health utility measure for a specific application is vital,²³ and evidence of the relation between health utilities and neurological disability is essential for determining the relative clinical appropriateness of these instruments as outcome measures for MS patients.²⁴ Appropriate health policy decisions require health outcome measures that are valid from the standpoint not only of those holding the “purse strings” of healthcare expenditures, but also from the standpoint of patients and clinicians. The challenge is to find such measures.

In this study, we examined the practical application and psychometric properties of three health utility measures in a sample of MS patients with a broad range of neurological disability. The selected instruments included the EQ-5D²⁵ and the Health Utilities Index Mark III (HUI Mark III),²⁶ because both have been used previously in studies of MS and are commonly used health utility measures. We also examined the SF-6D,²⁷ as it is derived from the SF-36, one of the most commonly used measures of HRQoL. Because of their limited item content and the scaling methods used to derive an interval measure from them, health utility measures are typically intolerant of missing item responses. Thus, our comparison of the three health utility measures was based on the feasibility of obtaining complete data for the instruments, evidence of test/retest reliability, and evidence of construct validity. For the latter, we compared the three health utility measures with clinician ratings of neurological disability and with objective measures of upper and lower extremity task performance.

METHODS

Design and subjects

This cross sectional study was conducted at MS clinics at two Canadian sites, the University of Calgary, in Calgary, Alberta, and the Queen Elizabeth II Health Sciences Centre, in Halifax, Nova Scotia. Ethics review board approval was obtained from both institutes. Subjects with clinically definite MS²⁸ were randomly selected from the registries of the clinics, both of which are the only specialised MS clinics serving their respective geographic regions (southern Alberta and Nova Scotia). Registration is maintained for all patients who have visited the clinics, and recent clinic visits were not required for recruitment eligibility. All subjects were characterised according to their current disease course (relapsing remitting (RR), primary progressive (PP), secondary progressive (SP), and progressive relapsing (PR)) using consensus definitions.²⁹ For this study, relapses were defined as the appearance of new signs and symptoms of MS or the reappearance of old signs and symptoms, lasting at least 72 hours, in the absence of fever, and preceded by 30 days of stability. Potential subjects were excluded if they were participating in a clinical trial. At the time of the study, 26 patients (14%) were taking one of the MS disease modifying agents: interferon beta-1b/Betaseron® (17 patients), glatiramer acetate/Copaxone® (5), interferon beta-1a/Rebif® (3), and interferon beta-1a/Avonex® (1). Two of the patients had previously taken such medications but had discontinued their use.

Protocol and measures

The EDSS was completed by trained nurses who administered a standardised neurological examination. This EDSS score determined disability group membership for the analyses described below. The study protocol also included

other measures commonly used in clinical trials of MS treatments: the ambulation index,³⁰ the nine hole peg test,³¹ and the timed 25 foot walk test from the MS Functional Composite.¹¹

Social and demographic data were obtained by interview. All subjects completed self assessed versions of the EQ-5D, the HUI Mark III, and the SF-6D. In order to avoid response biases, the order of completion of these consecutively administered questionnaires was randomised. For the initial completion of the questionnaires, the study nurses provided any assistance required by the subject to compensate for their physical limitations. To examine test/retest reliability of the health utility measures, subjects were provided with a second copy of the HRQoL assessments and were requested to complete and date these at home, 2 weeks after their initial assessment. These were returned to the study centre via prepaid courier. If the subjects had required assistance to complete the questionnaires because of physical limitations during the initial assessment, the study nurses ensured that such help would be available from a caregiver to complete the HRQoL measures at follow up.

The SF-6D was developed from the SF-36 items on the basis of expert panel review.²⁷ A subsample of SF-36 items were restructured into six health dimensions representing physical functioning, role limitations, social functioning, pain, mental health, and vitality. Using both visual analogue scaling and standard gamble techniques, preference weights were obtained for a selected sample of health states from 165 individuals, and the responses were modelled using ordinary least squares regression to produce a health utility measure. Despite its limited item content, the SF-6D assesses symptom domains that have been identified both by MS patients (mental health, vitality) and by MS clinicians (physical functioning, physical role limitations) as important determinants of HRQoL.³²

The EQ-5D is a utility measure that was designed for survey research and as an adjunct to “more detailed condition specific or treatment specific measures”.²⁵ The health dimensions sampled in the EQ-5D are mobility, self care, usual activities (role limitations), pain/discomfort, and anxiety/depression. Each dimension contains just three levels, yielding a total of 245 health states that are defined by all of the possible patterns of item responses. Although limited in its item content and response options, the EQ-5D has the practical advantage of being easy to administer and score.

In contrast to the relative simplicity of the EQ-5D, the HUI Mark III can derive a total of 972 000 possible health states representing the responses in eight health dimensions: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain.²⁶ The HUI Mark III has been described as an “in the skin” measure of HRQoL in that it purposefully does not enquire about functioning or role limitations that may be influenced by factors such as the physical or social environment. One potential advantage of the HUI Mark III for use in MS patients is its item content, which emphasises sensory and motor functioning, and its subscales, which are somewhat analogous to the functional systems assessed by the EDSS. The developers of the HUI have argued that the multi-attribute utility theory and standard gamble methodology used to establish the HUI Mark III preference weights offers advantages over measures that rely on time trade off or visual analogue scaling methods.³³ Nevertheless the HUI Mark III has been less widely used, presumably in part due to the costs associated with its licensing.

Analyses

All data were analysed using the SAS statistical package.³⁴ The EQ-5D preference weights used in this study came from

the UK general population survey.³⁵ The SF-6D utility was based on the visual analogue scaling procedure.²⁷ HUI Mark III utility was calculated using the algorithms published by the developers of this scale.²⁶ Proportions and means (SD) are reported for categorical and continuous data respectively. All analyses except those examining test/retest reliability were based on the first administration of the health utility measures.

Feasibility

While the measures had obvious differences in feasibility of administration related to issues of their availability, cost, length, and ease of utility calculation, our primary concern was item completion rate. We examined the frequency of missing item responses for the EQ-5D, HUI Mark III, and SF-6D while potential systematic effects of sex, education (<12 years versus >13 years), and neurological disability were examined using Fisher's exact test. For the latter analyses, subjects were grouped according to their EDSS score into mild (EDSS 0.0–2.5; from normal neurological results to minimal disability), moderate (EDSS 3.0–5.5; moderate disability but retaining independent ambulation), severe (EDSS 6.0–8.0; from requiring assistance for ambulation to restricted to bed or chair but out of bed for most of the day and retaining self care) and very severe (EDSS 8.5–9.5; from essentially restricted to bed most of day to helpless bed patient) disability categories. Floor and ceiling effects were also examined for each utility and their subscales.

Test/retest reliability

Intraclass correlation coefficients (ICC) were used to assess the retest performance of the utility measures between time 1 (t_1) and t_2 . ICC coefficients between 0.80 and 0.89 are considered acceptable for population surveys,³⁶ while a minimum of 0.90 has been proposed for clinical application of an instrument.³⁷

Construct validity

As constructs such as HRQoL cannot be measured directly, instruments that purportedly measure the construct(s) in question and instruments that address dissimilar constructs can be compared using a correlation matrix in order to provide evidence of construct validity. We produced two Spearman correlation matrices, one to compare correlations between the health utility measures, and a second to compare the health utility measures with clinical ratings (EDSS, ambulation index) and performance based measures of disability (timed 25 foot walk test, nine hole peg test). Lower correlations are expected for instruments that measure different constructs using different data collection methods, higher correlations are expected for similar constructs measured by similar methods, with moderate correlations expected for similar constructs measured by different methods or similar methods that measure different constructs. Using these guidelines, we considered correlations less than 0.30 as weak evidence of validity, correlations between 0.30 and 0.59 as moderate, and correlations above 0.59 as strong.³⁶

Construct validity was also examined by comparing the ability of the HRQoL utility measures to distinguish groups of subjects at different levels of neurological disability. This was based on the assumption that MS subjects with mild, moderate, severe, and very severe disability would report different levels of HRQoL, and that more disabled subjects would report lower HRQoL. For these analyses, one way analysis of covariance was used with EDSS group stratification (0.0 to 2.5 = mild, 3.0 to 5.5 = moderate, 6.0 to 8.0 = severe, 8.5 to 9.5 = very severe disability) as the inter-group factor, and age, sex, and disease course used as

covariates. The dependent variables were the three health utility measures. The statistical probability levels that are reported were corrected for multiple comparisons within each individual health utility measure by using Tukey's procedure, and the F ratios were adjusted for the covariates.

RESULTS

Consent to participate was received from 198 patients, but eight individuals could not be scheduled for a clinical assessment within the 8 month time frame of the study and three did not complete the HRQoL utility measures. Thus, data from 187 patients, all of whom completed the HRQoL utility measures on two occasions, were analysed (table 1).

Feasibility

Missing data were <5% for all of the individual subscales of the HUI Mark III (<1.6%) and the SF-6D (<3.2%), and for EQ-5D subscales (<0.5%) other than mobility (7.5%). Missing data for total scores of the health utility instruments were all <10% (HUI Mark III = 5.9%; SF-6D = 6.4%; EQ-5D = 8.6%). There were no systematic effects of sex or education on missing data for subscale and total scores of the EQ-5D and HUI Mark III, but males were more likely to have missing data for the total score of the SF-6D (males = 14.9%, females = 3.6%, $p = 0.01$). Level of neurological disability affected the proportion of missing data for the SF-6D and EQ-5D. Subjects with very severe disability (EDSS 8.5–9.5) were likely to omit physical function questions for the SF-6D (20.0%, $p = 0.01$) and skip the EQ-5D mobility item (42.9%, $p < 0.01$). Less disabled subject groups were unlikely to skip these items from the SF-6D (0.0–2.4%) and the EQ-5D (0.0–9.1%). The result was more missing SF-6D (40.0%, $p < 0.01$) and EQ-5D (42.9%, $p < 0.01$) utility scores for the very severely disabled subjects.

Floor (poorest self rated health) and ceiling (optimal self rated health) effects were also examined for the individual utility subscales and for the utilities themselves. For the EQ-5D, floor effects were $\leq 10\%$ for all of the subscales, and only 4% of subjects had an EQ-5D utility ≤ 0 . Ceiling effects were common between the individual subscales, however, ranging from 32% (mobility) to 68% (self care), and 15% of subjects reported an EQ-5D utility score of 1.0. Calculations of floor and ceiling effects for the SF-6D utility were more difficult because the full range of utility (0–1) is not represented in this scale.²⁷ However, only 3% and 1% of subjects reported the

Table 1 Clinical and demographic characteristics of subjects

Demographics	MS patients (n = 187)
Female	140 (75%)
Age (years), mean (SD) (range)	51 (10) (26–84)
Years of education, mean (SD) (range)	14 (2) (7–22)
Disease course subtypes, n (%)	
Relapsing remitting	74 (40)
Primary progressive	55 (29)
Secondary progressive	44 (24)
Progressive relapsing	14 (7)
Exacerbation of symptoms at time 1, n (%)	12 (6)
Exacerbation of symptoms at time 2, n (%)	0 (0)
EDSS score, median (interquartile range)	6 (4)
EDSS disability classification, n (%)	
Mild (0.0–2.5)	41 (22)
Moderate (3.0–5.5)	43 (23)
Severe (6.0–8.0)	88 (47)
Very severe (8.5–9.5)	15 (8)
Ambulation index score, mean (SD) (range)	4 (3) (0–9)
9 hole peg test, mean (SD) (range)*	87 (105) (12–300)
25 foot walk test, mean (SD) (range)*	65 (81) (3–180)

*Measured in seconds

lowest and highest possible scores respectively. For individual subscales, however, floor and ceiling effects were more common. Floor effects were reported by 41% of subjects in physical function and 16% reported floor effects in role limitations, while the remaining subscales were <10%. Ceiling effects were reported by 84% for role limitations, 58% for mental health, 39% for social functioning, 29% for bodily pain, 14% for vitality, and only 6% for physical function. For the HUI III, ceiling effects were present in only 3% of subjects for the overall utility as well as for each of the individual subscales. There were no floor effects on the subscales but 10% of subjects reported HUI III utilities that were ≤ 0 .

Test/retest reliability

Intra-class correlation (ICC) coefficients were 0.81 for the EQ-5D, 0.83 for the SF-6D, and 0.87 for the HUI Mark III. All ICC coefficients met the criterion for application in population surveys.³⁶

Construct validity

Spearman correlation coefficients provided strong evidence of construct validity between the health utility instruments. The correlation of the SF-6D and the EQ-5D was 0.70, while the correlations of the SF-6D and EQ-5D with the HUI Mark III were 0.69 and 0.80, respectively.

The correlation matrix presented in table 2 was used to examine construct validity between the health utilities and clinical measures. As a higher score on clinical measures represents poorer neurological status, while a higher health utility score represents better perceived health, all correlations between the clinical and utility measures are negative. With few exceptions, the HUI Mark III had the highest correlations with the clinical measures and only the HUI Mark III had strong correlations with the EDSS total score, the nine hole peg test, the ambulation index, and the timed 25 foot walk test. The EQ-5D demonstrated strong evidence of construct validity with most clinical measures, but moderate evidence with the nine hole peg test. Only moderate evidence of construct validity was evident between the SF-6D and each clinical measure. The correlations for all health utility measures with the individual functional systems scores of the EDSS were lower than the correlations with the EDSS total score, but the pattern of the strength of these relations was similar across the health utilities. For all three, their relations were highest with the pyramidal, bowel/bladder, and cerebellar functional systems, and lowest for the cerebral and visual functional systems.

For illustrative purposes, fig 1 presents the three health utility measures at each EDSS point. The planned analysis of covariance revealed overall significant effects of EDSS group

for all health utility measures (SF-6D $F = 7.14$, $p < 0.001$; EQ-5D $F = 16.49$, $p < 0.001$; HUI Mark III $F = 26.19$, $p < 0.001$), indicating that each instrument was able to distinguish between the groups of mildly, moderately, severely, and very severely disabled patients. Further pairwise comparisons indicated that the individual health utility measures differed in their ability to distinguish between the EDSS groups. For the EQ-5D, the decline in utility between the mildly and moderately impaired groups was not statistically significant ($p = 0.30$) although all other pairwise comparisons were significant. In contrast, the SF-6D was only able to distinguish the mildly disabled group from the moderately and severely disabled groups ($p < 0.001$), and was unable to distinguish between any other groups owing to a "flattening" of the decline in utility scores beyond moderate disability levels. Only the HUI Mark III demonstrated a significant decrease across increasing EDSS disability groups for all pairwise comparisons.

DISCUSSION

Head to head comparisons of health utility measures are necessary for selecting the most sensitive measure of the health construct of interest.²¹ The importance of comparative studies in specific patient populations is increasingly evident,²² but studies such as ours have been exceedingly rare, making the selection of utility measures for studies of MS difficult. As a result, investigators are likely to select instruments based on their familiarity, ease of administration, cost, and availability, rather than their psychometric properties in a given subject population. Another consideration could be the time referents for the instrument. For example, while the HUI Mark III asks individuals to relate each health state to "the past 2 weeks", the reference point is "a typical day" for the SF-6D, and "today" for the EQ-5D.

In our analyses, while each health utility that we examined performed well in one or more respect, only the HUI Mark III consistently met or exceeded the criteria we set out for selecting a clinically relevant measure for pharmacoeconomic studies of MS treatments. Both the HUI and our clinical comparator, the EDSS, primarily assess function, and it is not surprising that strong concordance was observed between them. However, the HUI Mark III also assesses emotional health and, most importantly, incorporates patients' perspectives of their health status. The HUI Mark III seems best suited for pharmacoeconomic studies of MS treatments in which strong concordance between primary clinical outcomes of neurological function and health utility are desired.

The EQ-5D and SF-6D too assess anxiety/depression and mental health respectively, and both also measure role limitations in addition to function. Thus, while their item content may result in less concordance with measures of neurological function, they may be better suited for studies of broader psychosocial health issues.

Table 2 Spearman correlations between the health utility measures and clinical measures of neurological disability and functioning

Clinical measures	SF-6D	EQ-5D	HUI Mark III
EDSS	-0.48	-0.66	-0.77
EDSS Functional Systems			
Cerebral	-0.29	-0.32	-0.28
Visual	-0.16	-0.24	-0.35
Brainstem	-0.29	-0.42	-0.40
Pyramidal	-0.45	-0.63	-0.71
Sensory	-0.34	-0.38	-0.46
Cerebellar	-0.38	-0.45	-0.63
Bowel & bladder	-0.45	-0.50	-0.64
Ambulation index	-0.52	-0.68	-0.76
9 hole peg test	-0.41	-0.56	-0.65
25 foot walk test	-0.49	-0.63	-0.73

Figure 1 Mean health utility scores for MS patients by each EDSS point. Squares, EQ-5D; triangles, SF-6D; diamonds, HUI Mark III. Error bars represent the standard error of the mean.

Despite the differences between the health utility measures in their item content, number and type of health dimensions, response time frame, and scaling methods, there were also numerous similarities. All were sufficiently reliable for population survey research,³⁶ and missing item responses were relatively rare. SF-6D total scores were missing more often for males in our sample, but the reason for this remains unclear. The HUI Mark III did not differ from the other measures in the proportion of missing data by subscale or total utility score, but it was less affected by EDSS disability group membership. While the small size of our very severely disabled group requires cautious interpretation, these subjects more often failed to respond to the physical functioning and mobility items of the SF-6D and EQ-5D. Similar omission of physical function data by severely disabled MS patients has also been observed in postal surveys using the SF-36.³⁸ As patients in the very severe disability group were at a minimum restricted to bed for much of the day,¹⁰ these omissions may reflect a lack of relevant response options for bed restricted patients on the SF-6D and EQ-5D. Regardless, it would appear prudent to limit collection of EQ-5D and SF-6D data to MS patients with mild to severe disability.

Together, all three of the health utility measures demonstrated strong evidence of construct validity, with correlations of a magnitude similar to those reported in a population based study.²¹ Nevertheless, statistically and clinically significant differences between the measures emerged when we examined their concordance with clinical measures. The EQ-5D and HUI Mark III demonstrated strong concordance with clinical measures of neurological disability, particularly with the pyramidal, cerebellar, and bowel and bladder functional system scores of the EDSS. This same pattern was seen for the SF-6D, but evidence of construct validity was moderate. Presumably, this reflects the more limited decline in SF-6D utility with increased disability illustrated in fig 1. Similar compression of the range of the SF-6D utility has been noted previously,²¹ and may represent a response shift in more disabled subjects. As individuals adapt to disability, illness, or ageing, their reference for comparison of disability states can shift from young healthy individuals to peers who are older or disabled, with the result being reports of a higher self reported health than expected.³⁹ This potential for response shift warrants consideration when selecting a health utility measure for pharmacoeconomic studies of MS, and our cross sectional data suggest that this may be a particular concern for the use of the SF-6D. An accurate understanding of the influence of response shift will require extensive longitudinal follow up of MS patients, however.

In contrast to the strong relationship between the HUI Mark III and EDSS in our sample, Grima and colleagues⁶ reported only a moderate relationship ($\rho = -0.54$) between the EDSS and an earlier, Mark II version of the Health Utility Index. They suggested that a strong relationship could not be expected because the EDSS predominately reflects ambulation and the Health Utility Index consists of multiple health attributes, but their exclusion of subjects with an EDSS >6.0 undoubtedly limited the range of health utility that they obtained. The broad range of neurological impairment in our sample was important as it allowed us to assess the limits of the health utilities and to compare them across the full range of neurological impairment. The full range of utility was represented in the responses to both the HUI Mark III and the EQ-5D, and both had overall strong relations with the clinical measures. However, unlike the HUI Mark III, the EQ-5D was unable to distinguish mildly from moderately impaired patients. The higher proportion of ceiling effects in the EQ-5D may, at least in part, explain this finding. Using a different approach and population, Hawthorne and colleagues²¹ also suggested a lack of sensitivity of the EQ-5D

in distinguishing “between those with full health and those with some health problems” (p. 368 of that article). Such distinctions are important for evaluations of MS treatments that are likely to continue to target less disabled patients.

We examined the properties of the health utility measures in a sample of MS patients with a broad range of disability that were drawn from regional MS speciality clinic registries, as it is within this context that MS treatment cost effectiveness studies are likely. Despite its strengths, however, our study is subject to a number of limitations. Among them is our limited sample size, which was insufficient to allow us to look for differences between specific subpopulations of patients, such as those with current exacerbation versus stability of symptoms and those taking new therapies or not. We did not attempt to examine the measurement properties of alternative methods of administration of these instruments, such as telephone based administration, nor did we debrief patients to determine why specific questions were not answered. In addition, further longitudinal studies will clearly be required in order to explore issues such as the potential response shift of MS patients over time and the responsiveness of health utility measures to treatment effects and changes in neurological disability. The limitations of our study illustrate the impact of time constraints and fiscal challenges on the ongoing process of validating health utility measures. However, we concur that it is “...far better (to choose) an apparently vague measure of outcome which is valid, reproducible, and important to patients than an apparently exact measure which only partially reflects the concerns of patients and which ignores the side effects of treatments”⁴⁰.

Funding for MS treatments is subject to reassessment in ongoing health policy debate and the relative effectiveness and cost effectiveness of competing health care services is central to such debate. Treatment consequences must consider outcomes beyond disease specific measures in order to contribute meaningfully to considerations of ‘cost/utility’ and ‘opportunity costs’ in health service delivery and programme funding deliberations. The appropriate selection and use of health utility measures will allow comparisons of treatment and programme outcomes on a common metric that reflects individuals’ health preferences.²² While our current findings provide greatest support for the use of the HUI Mark III in future clinical trials, further prospective longitudinal studies are needed to broaden the scope of health outcome assessments and to clarify the choice of utility measures for comparative cost/utility analyses of MS treatments.

ACKNOWLEDGEMENTS

This research was supported by AstraZeneca and by the Multiple Sclerosis Society of Canada. I S Sketris holds a CHSRF/CIHR/NSHRF Chair in Health Services Research.

Authors’ affiliations

J D Fisk, T J Murray, Department of Medicine, Dalhousie University, Halifax, NS, Canada

J D Fisk, Department of Psychiatry, Dalhousie University, Halifax, NS, Canada

J D Fisk, M G Brown, K I Stadnyk, Capital District Health Authority, Halifax, NS, Canada

M G Brown, I S Sketris, T J Murray, Department of Community Health and Epidemiology, Dalhousie University, Halifax, NS, Canada

M G Brown, I S Sketris, College of Pharmacy, Dalhousie University, Halifax, NS, Canada

L M Metz, Department of Clinical Neurological Sciences, University of Calgary, Calgary, AB, Canada

Competing interests: none declared

REFERENCES

- 1 Goodkin DE, Fisher JS. Treatment of multiple sclerosis with methotrexate. In: Goodkin DE, Rudick RA, eds. *Multiple sclerosis advances in clinical trial design. Treatment and future perspectives*. London: Springer, 2002:251–7.
- 2 The PRISMS Study Group and the University of British Columbia MS/RMR Analysis Group. PRISMS-4: long-term efficacy of interferon- β 1a in relapsing MS. *Neurology* 2001;**56**:1628–136.
- 3 Rice G, Ebers G. Interferons in the treatment of multiple sclerosis: do they prevent the progression of disease? *Arch Neurol* 1998;**55**:1578–80.
- 4 Brown MG, Murray TJ, Sketris IS, et al. Cost-effectiveness of interferon beta-1b in slowing multiple sclerosis disability progression. First estimates. *Int J Technol Assess Health Care* 2000;**16**:751–67.
- 5 Parkin D, McNamee P, Jacoby A, et al. A cost-utility analysis of interferon beta for multiple sclerosis. *Health Technol Assess* 1998;**2**:54.
- 6 Grima DT, Torrance GW, Francis G, et al. Cost and health related quality of life consequences of multiple sclerosis. *Mult Scler* 2000;**6**:91–8.
- 7 Kendrick M, Johnson KI. Long-term treatment of multiple sclerosis with interferon-beta may be cost-effective. *Pharmacoeconomics* 2000;**18**:45–53.
- 8 Parkin D, Jacoby A, McNamee P, et al. Treatment of multiple sclerosis with interferon β : an appraisal of cost-effectiveness and quality of life. *J Neurol Neurosurg Psychiatry* 2000;**68**:144–9.
- 9 Little R. NHS to fund treatment for 1000 patients with MS. *BMJ* 2002;**324**:316.
- 10 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983;**33**:1444–52.
- 11 Rudick R, Antel J, Confavreux C, et al. Recommendations from the National Multiple Sclerosis Society Clinical Outcomes Assessment Task Force. *Ann Neurol* 1997;**42**:379–82.
- 12 Herndon RM, Goodkin DE. *Handbook of neurologic rating scales*. New York: Demos Vermande, 1997.
- 13 Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;**118**:622–9.
- 14 Ware JE Jr, Sherbourne CD. The MOS 36-item short form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;**30**:473–83.
- 15 Arnoldus JHA, Killestein J, Pfenning LEMA, et al. Quality of life during the first 6 months of interferon- β treatment in patients with MS. *Mult Scler* 2000;**6**:338–42.
- 16 Vickrey BG, Hays RD, Harooni R, et al. A health-related quality of life measure for multiple sclerosis. *Qual Life Res* 1995;**4**:187–206.
- 17 Ford HL, Gerry E, Tennant A, et al. Developing a disease-specific quality of life measure for people with multiple sclerosis. *Clin Rehabil* 2001;**15**:247–58.
- 18 Hobart J, Lamping D, Fitzpatrick R, et al. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. *Brain* 2001;**124**(Pt 5):962–73.
- 19 Drummond MF, O'Brien B, Stoddart GL, et al. *Methods for the economic evaluation of health care programmes*, 2nd ed. Oxford: Oxford University Press, 1997.
- 20 Cairns J. Measuring health outcomes. *BMJ* 1996;**313**:6.
- 21 Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. *Ann Med* 2001;**33**:358–70.
- 22 Siderowf A, Ravina B, Glick HA. Preference-based quality-of-life in patients with Parkinson's disease. *Neurology* 2002;**59**:103–8.
- 23 Revicki DA, Ehreth JL. Health-related quality-of-life assessment and planning for the pharmaceutical industry. *Clin Ther* 1997;**19**:1101–15.
- 24 Freeman JA, Hobart JC, Langdon DW, et al. Clinical appropriateness: a key factor in outcome measure selection: the 36 item short form health survey in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2000;**68**:150–6.
- 25 Williams A. *The role of the EuroQol instrument in QALY calculations*. York: University of York, 1995.
- 26 Feeney DH, Torrance GW, Furlong WJ. Health Utilities Index. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*, 2nd ed. Philadelphia: Lippincotte-Raven Pub, 1996:239–52.
- 27 Brazier J, Usherwood T, Harper R, et al. Deriving a preference-based single index from the UK SF-36 Health Survey. *J Clin Epidemiol* 1998;**51**:1115–28.
- 28 Poser CM, Paty DW, Scheinberg L, et al. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol* 1983;**13**:227–31.
- 29 Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. *Neurology* 1996;**46**:907–11.
- 30 Hauser SL, Dawson DM, Leirich JR, et al. Intensive immunosuppression in progressive multiple sclerosis. A randomized, three-armed study of high-dose intravenous cyclophosphamide, plasma exchange, and ACTH. *N Engl J Med* 1983;**308**:173–80.
- 31 Goodkin DE, Hertsgaard D, Seminary J. Upper extremity function in multiple sclerosis: improving assessment sensitivity with box-and-block and nine-hole peg tests. *Arch Phys Med Rehabil* 1988;**69**:850–4.
- 32 Rothwell PM, McDowell Z, Wong CK, et al. 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.
- 33 Torrance GW, Furlong W, Feeny D, et al. Multi-attribute preference functions. Health Utilities Index. *Pharmacoeconomics* 1995;**7**:503–20.
- 34 SAS. *SAS version 8.2. User Manual*. Cary, NC: The SAS Institute Inc, 2001.
- 35 Dolan P, Gudex C, Kind P, et al. *A social tariff for EuroQol: results from a UK general population survey*. Discussion paper 138. Centre for Health Economics, York Health Economics Consortium, NHS Centre for Reviews & Dissemination. York: University of York, 1995.
- 36 Nunnally JC, Bernstein IH. *Psychometric theory*. New York: McGraw-Hill Inc, 1994.
- 37 McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health surveys adequate? *Qual Life Res* 1995;**4**:293–307.
- 38 Hobart J, Freeman J, Lamping D, et al. The SF-36 in multiple sclerosis: why basic assumptions must be tested. *J Neurol Neurosurg Psychiatry* 2001;**71**:363–70.
- 39 Groot W. Adaptation and scale of reference bias in self-assessments of quality of life. *J Health Econom* 2000;**19**:403–20.
- 40 Rothwell PM. Quality of life in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 1998;**65**:433.