The comparative ability of the EQ5D and SF6D health utility measures to assess change in patients with rheumatoid arthritis

Mark Harrison(1), Linda Davies(2), Nick Bansback(3) & Deborah Symmons(1)
1)arc Epidemiology Unit, The University of Manchester; 2)Health Economics Research at Manchester (HERMAN), The University of Manchester; 3)Centre for Health Evaluation and Outcome Sciences, Vancouver, BC, Canada

INTRODUCTION

Health-related quality of life measures with utility values such as the EQ-5D and SF-6D (based on the SF-36) are key outcome measures for use in clinical trials as they allow the cost-effectiveness of interventions to be calculated. Utility values, however, are not routinely used in clinical practice and therefore clinicians and health care professionals do not develop the necessary understanding through experience to gauge whether important change has occurred. Furthermore, interpretation of results based on statistical significance is influenced by sample size and variance, and a difference may be statistically significant, yet not clinically important. The definition of a clinically important difference aids study design by allowing researchers to calculate the sample sizes sufficient to detect important outcomes. It is therefore crucial to define the minimum important difference (MID) for utility measures to enable interpretation of the clinical importance of results.

A popular and frequently quoted definition of the MID is "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management."(1) This definition focuses solely on improvement in outcome, and studies aimed at defining the MID for utility measures in rheumatoid arthritis (RA) have reflected this narrow definition by either combining improvement and deterioration by ignoring the sign of change (2) or using statistical methods to calculate a MID which is then assumed to be constant irrespective of the direction of change(3). These estimates suggest the MID for the EQ-5D is 0.05 – 0.07(3;4) and for the SF-6D is 0.03 - 0.04,(2-4).

The limitations of Jaeschke's definition of MID in failing to recognize important deterioration as well as important improvement have been discussed recently by a number of authors. Cella et al(5) claim that the direction of clinically meaningful change is understudied, and that the validity of an equal MID for improvement and deterioration rests upon the symmetry of data on both sides of change, floor and ceiling effects of the measure, the severity and prognosis of disease, and the aim of a treatment or intervention. A study by the same authors in patients with cancer found that smaller gains in health-related quality of life were more meaningful and significant in value than larger losses.(6) The non-normal distribution and crude scoring levels of the EQ-5D, suggest that the assumption of equal MID in improving and worsening patients warrants further study. The MID for a measure almost certainly varies across patients and patient groups,(7) and the determination of MID is a cumulative process that depends on estimates from a range of sources.(6)

Outcome measures must also be able to measure change accurately. The terms sensitivity and responsiveness are frequently and interchangeably used to describe this aspect of a measure's validity. Liang distinguishes sensitivity, the ability to measure change regardless of whether it is relevant or meaningful, from responsiveness, the ability of an instrument to measure a meaningful or clinically important change.(8) Sensitivity might be tested using statistical measures based on assessing the signal to noise ratios such as effect sizes. Responsiveness is the degree of change in a measure associated with some criterion of important change.

The ability of the EQ-5D and SF-6D to measure change in RA patients have been compared against each other in a number of studies. In patients experiencing a deterioration in health, the EQ-5D has to date appeared to be sensitive,(9-11) whereas the SF-6D has less supportive evidence of sensitivity in these same patients. In improving patients the evidence for the sensitivity of the SF-6D was supportive, (10) whereas in other studies its relative sensitivity compared to other measures varied according to the definition of change and the sensitivity statistic employed. The evidence for the

sensitivity of the EQ-5D to improvement is also inconclusive, with some studies finding it insensitive(9) and others reporting evidence that it is the most responsive in detecting improvement.(11)

The aim of this paper was to estimate the MID for the EQ-5D and SF-6D MID in RA patients, focussing on the size of difference in patients reporting improving and deteriorating health. Secondly we explore the impact of the results in terms of sample size calculations. Finally, we assess whether the ability of the EQ-5D and SF-6D to detect change differs according to the direction of change.

METHODS

The study used the baseline and 1 year data from the British Rheumatoid Outcome Study Group (BROSG) randomised controlled trial (RCT) in patients with stable, established RA. The trial was conducted in 5 rheumatology centres in England (Stoke on Trent; Cannock; Truro; King's College Hospital, London; and Macclesfield), which include teaching and district general hospitals, serving urban and rural populations.

The BROSG Trial was conducted between 1998 and 2001 to compare the relative clinical and cost effectiveness and utility of symptomatic and aggressive treatment aimed at suppressing inflammation for established RA in a randomised, controlled, observer-blinded trial.(12;13). RA patients with more than 5 years duration were screened and invited to participate if they had been rheumatology outpatient attendees for at least 12 months, had been on stable therapy for at least 6 months, and had no evidence of systemic rheumatoid disease or serious co-morbidity.

The primary outcome measure was change in physical function, measured using the British version of the Health Assessment Questionnaire (HAQ)(14) which measures functional disability. Patients with a baseline HAQ score >2.5 were excluded. The BROSG trial showed no difference between treatment arms in the primary outcome measure (HAQ) at the end of the trial (adjusted for baseline HAQ, age at randomisation, gender, disease duration and treatment centre).(12;15) The patients had stable established

disease and were representative of approximately 30% of patients in routine rheumatology clinics.

Patient demographic data collected included age, gender, disease duration, smoking status and co-morbidity. Baseline measurements included a patient global assessment, physician global assessment, 28 tender joint count, 28 swollen joint count, HAQ. The patients also completed the SF-36(16), an internationally validated generic health status measure, and the EuroQol (EQ5D)(17), a generic measure of health status and health related utility(18). The DAS28 composite measure of RA disease activity (19), a visual analogue score for pain and the overall status in rheumatoid arthritis (OSRA) (20) disease activity (OSRA-A) and damage (OSRA-D) scores were also collected. All measures were repeated annually.

The MID was defined using question 2 (SF2) of the SF-36 which asks "compared to one year ago, how would you rate your health in general now?" The respondent answers the question using a 5-point likert scale where 1 is "much worse", 2 is "somewhat worse", 3 is "about the same", 4 is "somewhat better", and 5 is "much better". This method is similar to the approach described by Jaeschke et al (1). The 'anchor' of change has to be related to the outcome measure of interest, i.e. relate to change in the same aspect of health, and be interpretable in the context of identifying important change. The SF2 was chosen as the first indicator of retrospective self reported change in overall health, and the somewhat better or worse categories were used as the definition of MID. The question is framed over a 1 year period, and we related this change to the baseline and year 1 follow-up assessments in the BROSG trial. An alternative scale was also calculated using the EuroQol 'feelings thermometer'. The change in ratings between baseline and 1-year follow up was calculated as a percentage change based on the ACR response criteria (21;22). We defined small but important change as exceeding 20% but failing to achieve 50% change.

The MID for each outcome measure was calculated by summarising mean change in the groups defined above, with the MID for improving patients being mean change in the

"somewhat better" (retrospective assessment using the SF2 question) and the 20-50% improvement on the "feelings thermometer". Similarly, the MID for worsening patients was the mean change in the "somewhat worse" and 20-50% worsening (ACR response criteria) on the "feelings thermometer"

Sample size calculations were estimated using the methods described by Dupont and Plummer(23) for a one sample test of a mean difference equal to the MID. The calculation assumes we wish to have 90% power to detect the MID.

The responsiveness of the EQ-5D and SF-6D to change was assessed in the groups of patients grouped by self-assessed small change in health as described above. The effect size (ES) and standardised response mean were used to assess responsiveness. The ES and SRM determine the ratio of signal (size of change) to noise (variability in scores). The advantage of these statistics is that when expressed as a number of standard deviations it is possible to compare between measures with different scales. The ES was calculated using the mean change and the variance in baseline scores for the outcome measure in question, calculated as the difference between the mean score at follow-up and the mean score at baseline divided by the group mean at baseline. The SRM was calculated using the same mean change, but divided by the standard deviation of the change in scores over the follow up period. To compare the responsiveness statistics for the EQ-5D with those of the SF-6D, a ratio was calculated. The effect size ratio was the ES for the EQ-5D divided by the ES for the SF-6D, and similarly the standardised response mean ratio was the SRM for the EQ-5D divided by that of the SF-6D. Thus, for each measure a ratio in excess of 1 would indicate that the EQ-5D was the more sensitive, whilst a value less than 1 would indicate that the SF-6D was more sensitive.

RESULTS

A total of 440 patients in the BROSG trial attended a 12 month assessment and 26 (6%) were lost to follow up. Of these, 437 completed the SF2 question of the SF-36 and 436 completed the EQ-5D VAS. The patients of the BROSG trial generally deteriorated a little over the course of the first year of follow up (Table 1). There was statistically

significant deterioration in swollen joint count, patient global assessment, and pain during the 1 year of follow up. The EQ-5D had a bimodal distribution with groupings of patients around 0 and above 0.5. The group mean at baseline was 0.58, close to the edge of the distribution between 0.5 and 0.9.

Patients were classified into groups of change during the first year. 115 patients reported that their general health was worse than one year ago, and 77 patients reported that their health as better than one year ago. Of the patients reporting a change in health, 100 (87% of those reporting a worsening in health) reported that their health was 'somewhat' worse, and 53 (69% of those reporting improvement) reported that their health was 'somewhat better'. These patients form the basis of the responsive analysis based on the retrospective global rating of change. There were higher, but non-significant (p=0.127) proportions of patients from the aggressively treated arm of the BROSG trial in the patients reporting somewhat better health than in the groups reporting the same level of health or somewhat worse health.

Based on change defined using the percentage change between two EQ-5D VAS assessments, 81 patients (19%) had a VAS score worse ($\geq$20% & <50%) than baseline at the 1 year follow up and 62 patients (14%) had a VAS score 20-50% better than at baseline. These patients form the basis of the analysis based on change in VAS score. The proportion of patients from each treatment arm of the BROSG trial was similar in each of the three groups.

The MID based on mean change in the patients reporting somewhat worse health than the year previously was 0.10 (SD 0.26) for the EQ-5D, and 0.04 (SD 0.10) for the SF-6D (Table 2). Based on the change in EQ-5D VAS between baseline and 1 year assessments, there was a larger MID (0.13 [SD 0.24]) for the EQ-5D than the SF-6D (0.04 [SD 0.10]). The MID for the EQ-5D and SF-6D in patients reporting somewhat better health were both 0.04, although the EQ-5D had approximately 50% more variance around this estimate. Based on the EQ-5D VAS definition of improvement (Table 3), the MID in improving patients was 0.06 (SD 0.27) for the EQ-5D, double the MID of the SF-6D,

0.03 (SD 0.09), although both estimates were close to the estimates from the retrospective global assessment of change.

Almost all measures utilised in the BROSG study found greater change in patients classified as having worsening health compared to those defined as having an improvement in health. The exceptions to this rule, aside from the SF-6D, in groups defined using the SF2 question were the OSRA-D, the physician global assessment and the SF-36 physical composite score (PCS). These measures provided equal change in either direction. The SF-36 PCS changed more in improving patients than in worsening patients against the EQ-5D VAS change criterion.

The ES and SRM provided largely similar results in assessing the responsiveness of each measure to small but important change. The results did differ however, according to the definition of change. In measuring the change in patients reporting their health as somewhat better than the year previously, there was little difference between any of the responsiveness statistics, all ranged between 0.36 and 0.41. The ES ratio suggested the EQ-5D was marginally more responsive. However, the SRM ratio suggested that the SF-6D was slightly more responsive in worsening patients. This suggests both measures are equally responsive to deterioration in health, despite their very different mean change. However, against the criterion of change in EQ-5D VAS, the EQ-5D appeared to be the more responsive measure, as both the ES and SRM were considerably higher for the EQ-5D than the SF-6D. Conversely all of the results suggested that the SF-6D was to some extent more responsive than the EQ-5D in patients with improving health, although the difference between responsiveness statistics was small (range 0.03 -0.12). The EQ-5D and the SF-6D were both more responsive to deteriorations in health than improvements in health. The difference in responsiveness was again more pronounced for the EQ-5D, with responsiveness statistics up to three times larger in deteriorating patients, compared to more modest differences in responsiveness for the SF-6D. The EQ-5D and SF-6D were generally more responsive than measures of functional disability, namely the HAQ and in most cases the OSRA-D. The two utility measures were less responsive than patient

reported measures of global health, and disease activity measures such as the composite DAS28 measure and more specifically the pain VAS.

The power calculation based on the MID from the retrospective global rating of change measure suggested a sample size of 129 to detect a minimum important improvement in EQ-5D (Table 4). The power to detect a difference using the SF-6D in the same sample of patients should be higher as the measure appears to be the more responsive where patients were improving. Based on the SF-6D, the power calculation suggests a sample size of 66 would be needed to detect a minimum important difference. The equivalent power to detect the MID of the SF-6D in a sample size calculated to detect the MID of the EQ-5D with 90% power would be 99.5%. The MID estimates for the change in VAS anchor of change provided larger required sample sizes.

CONCLUSIONS

The results of this study show differences in the ability of the EQ-5D and SF-6D to measure change in RA patients according to different directions and magnitudes of change**.** The MID for the EQ-5D may be twice as large in worsening patients as in improving patients, and twice as large as the SF-6D in worsening patients. In contrast, the MID of the SF-6D appeared to be equal in both directions, and smaller than that of the EQ-5D. The SF-6D was more responsive than the EQ-5D to improvements in health due to treatment in patients with both very early and very severe disease, and also in detecting small but important improvement in patients with established disease. In contrast there was evidence to suggest that the EQ-5D might be more responsive to small deterioration in health in RA patients with stable, established disease. It is important to know how large deterioration in utility has to be in order to be considered important. Some treatment in RA may be termed a success if the progression of disease is slowed down or halted.

The MIDs previously reported for the SF-6D were approximately 0.03 to 0.04 in studies in the UK and Canada. (2-4). Estimates for the EQ-5D range from 0.05 (3) up to 0.07 for the EQ-5D based on eleven patient groups from eight longitudinal studies from a range of conditions (one of which was early RA).(4) The estimate for the EQ-5D in the early RA

group was 0.13.(4) The MID estimates from this study for the SF-6D correspond closely with the estimates of Marra et al and Walters & Brazier, and suggest that the MID for the SF-6D in RA is consistently around 0.04. The estimates for the EQ-5D were more interesting. In improving patients, the MID estimate agrees with the estimates previously reported for the EQ-5D MID in RA. However, where patients deteriorated, the MID estimates are as much as twice the size of those reported previously. As a proportion of the range of possible scores for the EQ-5D (-0.59 - 1) and SF-6D (0.3 – 1), the previous estimates represent a change of 3-4% of the range for the EQ-5D and 4-6% of the range for the SF-6D. The results for the SF-6D from this study are entirely consistent; however my estimates for the EQ-5D represent 3-4% of the range in improving patients and 6-8% of the range in deteriorating patients.

The finding of different MID estimates for the EQ-5D in improving and deteriorating patients is novel. Previous studies estimating the MID for these measures have reported values that either can not account for the direction of change(3) or treat change uniformly.(2) However our data suggest that this assumption does not hold. In deteriorating patients the required sample sizes to detect an important and statistically significant change for both the EQ-5D and SF-6D would be comparable, as the EQ-5D has a larger MID but also greater variance in the estimate, which increases the sample size required to detect that level of change statistically. It is possible that the larger MID and variance are due to the bi-modal distribution of the EQ-5D, which is well documented (10;11;25;26). If a minority of the patients reporting a worsening in health are moving between the distributions i.e. from above 0.5 to around 0, this might explain the large MID and large variance around this estimate. The SF-6D is more normally distributed and the group was close to the middle of the range of possible scores (0.30 – 1.00), providing opportunity for more equal change in either direction.

In previous studies of responsiveness, the EQ-5D was consistently the most responsive measure in detecting deterioration in health and almost twice as responsive as alternative measures (Table 5).(9;11) The evidence suggested that responsiveness in improving patients was, however, conflicting. Russell et al, in assessing response to Biologics

treatment, and Marra et al using methods similar to this study found the SF-6D was more responsive than the EQ-5D.(9;10) However, Conner-Spady et al reported that the EQ-5D was more responsive than the SF-6D in detecting both improvement and deterioration.(11) It is difficult to speculate why the Conner-Spady study found such a large effect for the EQ-5D in improving patients. Their patients were recruited from a rheumatology clinic and change assessed retrospectively using a 5 point scale collapsed into improvement, no change and deterioration. Therefore the recruitment and assessment of change was essentially analogous to the self-report change study by Marra et al.

Despite concerns about scaling potentially limiting its scope to detect change, the EQ-5D was also able to detect patient reported improvement and deterioration in this study and other studies.(9;11) The SF-6D was, however, more responsive in these improving patients and it is likely that the smaller increments between scoring levels in this measure allow patients to report smaller, but still important improvements. Some patients experiencing noticeable change may not feel their improvement in an aspect of health warrants a change from ‘severe’ to ‘moderate’ or from ‘moderate’ to ‘no problems’, i.e. EQ-5D states. This is reflected by the fact that the SF-6D shows relatively small absolute change but has a small standard deviation,(10;11) which leads to good responsiveness statistics in longitudinal studies or studies assessing change/differences. The SF-6D appears to be more ‘efficient’ in groups where an improvement may be anticipated, requiring fewer patients to be able to detect a clinically important and statistically significant difference.

An important limitation of our results relates to the use of anchors. The Likert scale taken from question 2 of the SF-36 has not been well validated as an external anchor of change, with only one study to date reporting the test-retest reliability of the question over 1-4 days. This study found reliability to be moderate to strong (weighted Kappa 0.64-0.73)(27). However, the question is employed as a reference of change over a 1 year period in this study. Nevertheless, this scale has also been used as an anchor in previous estimates of the EQ-5D and SF-6D in the field of rheumatology (2;4) and therefore provides a useful common criterion against which to compare our results.

Retrospective scales are also subject to a number of possible sources of bias, which may include recall bias, whereby recent or relevant/significant events may influence a patient's response to a question (28), or at least make them less likely to have forgotten the true course of their disease.(29;30) A further bias is adaption to disease and coping(31) and response shift (32), and both may influence patient assessment of health or change in health. Adaptation and coping may be the development of new skills to cope with and overcome the restrictions of disease, or a change in activities to avoid limitation by disease. Adaption may also be psychological such as denial of the severity of health state, and restricted recognition of full health.(31) Response shift is a change in the way a patient values their health due to a change in their internal reference points over time, and can lead to an apparent improvement where no underlying change has occurred.(32) These problems may potentially be exacerbated by the long interval (1 year) over which the rating is made in this study.

A further limitation of the retrospective assessment of change used as an anchor in this study may be that it is only a 5 point likert scale. Other authors have suggested using likert scales with 7 or more points(1;33), for example ranging from almost the same; hardly any better/ worse through to a very great deal worse/better.(33) Our chosen 5 point scale may lack the sensitivity of these extra categories. However in defining MID authors have tended to collapse categories. For example, Guyatt et al chose to collapse the category of almost the same, hardly any better/worse through to somewhat better/worse,(33) The phrasing of the cut-off point is, therefore, similar to that of the scale used in the first part of this study.

In view of the criticisms of retrospective evaluation of health change, we chose to use an alternative classification based on the change between two serial valuations of overall health on a VAS scale. Change was classified as important if it fell between a difference of 20% and 50% from baseline using terms for response based on the ACR response criteria.(21) Again, this anchor has not been validated. However, we based the lower boundary on a similar method described previously in important research validating

utility measures in RA by Marra et al (9) and (4). The strengths of the anchors employed in our study are the ease of interpretation of results and the emphasis on the change that the patient perceives to be important. Liang recommends that for measures of quality of life the subject's valuation should be used to assess the magnitude of change and its importance.(8)

In the absence of substantial evidence about the validity of the anchor, it is difficult to conclusively state whether patients value small improvements in health more than small deteriorations in health, or whether the results reflect differential interpretation of the descriptive language of "somewhat better" compared to "somewhat worse". Evidence in our results suggest that either scenario could be true. The magnitude of change in disease activity measures (DAS 28, 28 swollen joint count), patient and physician global assessments, the pain VAS and the HAQ all indicate a larger change in patients reporting somewhat worse health than somewhat better health. In contrast, the EQ-5D feelings thermometer VAS scale, OSRA activity and damage scores, and the 28 tender joint count all provide evidence of equal magnitude of change. However, a strength of our study was the use of two anchors of change. The results from the alternative anchor of change, the difference between VAS ratings of health today at baseline and 1 year, support the findings of an MID more than twice as large in patients retrospectively reporting worsening health compared to those in improving health.

A further limitation is that the anchors of important change we used in this study were both taken from the questionnaires on which the utility measures we are evaluating are based, the SF-36 and the EuroQol EQ-5D. This might potentially affect the MID estimates of the utility measure based on the questionnaire from which the anchor was taken. However, neither scale of the anchors is used in the calculation of the SF-6D or EQ-5D utility measures, and both anchors provided consistent results. The influence of the anchor on the estimate of the MID therefore appears unlikely. It is also likely that there is some misclassification of change between the anchors and some of the outcome measures, and this may be particularly true for the change in VAS scale. This measure frames a patient's health on the day in question, and the change in VAS in theory assesses

the difference between a persons health on 2 days, 1 year apart. The framing of the question allows considerable potential for transient and possibly trivial factors to influence the rating of change in health. Despite this possibility, the design of the study aimed to classify patients on the basis of important change and only compare the responsiveness of the EQ-5D and SF-6D in patients changing in the same direction.

There are weaknesses in the use of responsiveness statistics. There is an array of such statistics, and different methods may lead to different conclusions.(8;35) To date, no measure has been proven conclusively to be preferable or superior to another. Furthermore responsiveness statistics are limited in the information they convey. The responsiveness statistics used in this study, the ES and SRM, standardise change in terms of the standard deviation. They were favoured because they provide a useful indication of the relative sample size that might be needed to detect a significant difference between groups.(8) The basis of the ES and SRM are that relevant change should exceed random noise or the variability in unchanged patients. These measures used without an anchor of important change, however, give no information about the validity of the measure to measure change in the underlying construct,(35) and essentially are measures of sensitivity. In this study, the responsiveness measures were not used without some external reference of change.

CONCLUSION

The EQ-5D and SF-6D respond to important change and improvement in response to treatment in patients with arthritis, but the SF-6D is more responsive in improving arthritis patients, whilst the EQ-5D may be more responsive to deteriorations in health. The EQ-5D has an unequal MID according to the direction of change. Small but important deteriorations lead to larger mean change than small but important improvements. The SF-6D detects change more symmetrically and has an equal MID in both directions of change.

ISSUES & FUTURE RESEARCH

1. Use of alternate methods to estimate the MID in view of limitations of the current methodology. Possibilities may include the combination of anchor and distribution based methods such as those proposed by Crosby (37).
2. Test the relative responsiveness to change of the EQ-5D and SF-6D prospectively in samples of patients, receiving efficacious treatment, large enough to detect the MID.

REFERENCES

(1) Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10(4):407-15.

(2) Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003;1(4):4-12.

(3) Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ5D) and disease specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60(7):1571-82.

(4) Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005 Aug;14(6):1523-32.

(5) Cella DA, Bullinger M, Scott C, Barofsky I. Group vs Individual Approaches to Understanding the Clinical Significance of Differences or Changes in Quality of Life. Mayo Clin Proc 2002 Apr 1;77(4):384-92.

(6) Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002 May;11(3):207-21.

(7) Guyatt GH, Osoba DH, Wu AW, Wyrwich KW, Norman GR. Methods to Explain the Clinical Significance of Health Status Measures. Mayo Clin Proc 2002;77(4):371-83.

(8) Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care 2000 Sep;38(9 Suppl):II84-II90.

(9) Marra CA, Rashidi AA, Guh D, Kopec JA, Abrahamowicz M, Esdaile JM, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res 2005;14:1333-44.

(10) Russell AS, Conner-Spady B, Mintz A, Mallon C, Maksymowych WP. The responsiveness of generic health status measures as assessed in patients with rheumatoid arthritis receiving infliximab. J Rheumatol 2003 May;30(5):941-7.

(11) Conner-Spady B, Suarez-Almazor ME. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. Med Care 2003;41(7):791-801.

(12) Symmons DPM, Tricker K, Roberts C, Davies L, Dawes P, Scott DL. The British Rheumatoid Outcome Study Group (BROSG) randomised controlled trial to compare the effectiveness and cost-effectiveness of aggressive versus symptomatic therapy in established rheumatoid arthritis. Health Technol Assess. 2004.

(13) Davis M, Tricker K, Roberts C, Dawes P, Hassell A, Knight S, et al. Aggressive therapy with conventional disease modifying anti-rheumatic drugs (DMARD) does not prevent disease progression in patients with stable established rheumatoid arthritis (RA): Results of a randomised observer-blinded controlled clinical trial. Rheumatology (Oxford) 2004;43 (Supplement 2):ii44.

(14) Kirwan JR, Reeback JS. Stanford Health Assessment Questionnaire modified to assess disability in British patients with rheumatoid arthritis. Br J Rheumatol 1986 May;25(2):206-9.

(15) Symmons D, Tricker K, Harrison M, Roberts C, Davis M, Dawes P, et al. Patients with stable long-standing rheumatoid arthritis continue to deteriorate despite intensified treatment with traditional disease modifying anti-rheumatic drugs - results of the British Rheumatoid Outcome Study Group randomized controlled clinical trial. Rheumatology (Oxford) 2006;45:558-65.

(16) Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992 Jun;30(6):473-83.

(17) The EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy 1990 Dec;16(3):199-208.

(18) Ware JE, Jr., Kosinski MA, Keller SD. SF-36 physical and mental health summary scales: a user's manual. 5th ed. Boston, MA: New England Medical Center; 1994.

(19) Tuttleman M, Pillemer SR, Tilley BC, Fowler SE, Buckley LM, Alarcon GS, et al. A cross sectional assessment of health status instruments in patients with rheumatoid arthritis participating in a clinical trial. J Rheumatol 1997;24(10):1910-5.

(20) Symmons DP, Hassell AB, Gunatillaka KA, Jones PJ, Schollum J, Dawes PT. Development and preliminary assessment of a simple measure of overall status in rheumatoid arthritis (OSRA) for routine clinical use. Q J Med 1995 Jun;88(6):429-37.

(21) Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993 Mar;20(3):557-60.

(22) Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Arch Intern Med 1993 Jun 14;153(11):1337-42.

(23) Dupont WD, Plummer WD, Jr. Power and sample size calculations. A review and computer program. Control Clin Trials 1990 Apr;11(2):116-28.

(24) Pagano M, Gauvreau K. Principles of biostatistics. 2nd edition ed. Pacific Grove, CA: Brooks/Cole; 2000.

(25) Wolfe F, Hawley DJ. Measurement of the quality of life in rheumatic disorders using the EuroQol. Br J Rheumatol 1997 Jul;36(7):786-93.

(26) Marra CA, Esdaile JM, Guh D, Kopec JA, Brazier JE, Koehler BE, et al. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. Med Care 2004;42(11):1125-31.

(27) Wyrwich KW, Metz SM, Babu AN, Kroenke K, Tierney WM, Wolinsky FD. The reliability of retrospective change assessments. Qual Life Res 2002;11(7):636.

(28) Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. J Clin Epidemiol 1997;50(8):869-79.

(29) Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003 May;56(5):395-407.

(30) Jordan K, Dunn KM, Lewis M, Croft P. A minimal clinically important difference was derived for the Roland-Morris Disability Questionnaire for low back pain. J Clin Epidemiol 2006 Jan;59(1):45-52.

(31) Menzel P, Dolan P, Richardson J, Olsen JA. The role of adaptation to disability and disease in health state valuation: a preliminary normative analysis. Soc Sci Med 2002 Dec;55(12):2149-58.

(32) Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med 1999 Jun;48(11):1507-15.

(33) Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. Thorax 1987 Oct;42(10):773-8.

(34) Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res 1992 Feb;1(1):73-5.

(35) Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Qual Life Res 2003;12(4):349-62.

(36) Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: A critical review and recommendations. J Clin Epidemiol 2000;53(5):459-68.

(37) Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003; 56(5):395-407.

TABLES

Table 1: Change over the first year of follow up in the BROSG trial

		(n=440)
Utility Measure	EQ-5D	-0.03 (0.23)
	SF-6D	-0.01 (0.11)
	EQ-5D VAS	-1.3 (18.9)
OSRA	OSRA-A	0.2 (1.9)
	OSRA-D	0.1 (1.3)
Disease Activity	DAS28	0.2 (1.1)
	28 Swollen Joint Count	0.6 (4.6)
	28 Tender Joint Count	0.8 (5.7)
	Patient Global Assessment (mm)	3.4 (21.7)
	Physician Global Assessment (mm)	1.7 (20.5)
	Pain VAS (mm)	3.7 (23.1)
	Fatigue VAS (mm)	1.3 (24.6)
Physical Function	HAQ	0.09 (0.40)
SF-36	MCS	-1.4 (10.0)
	PCS	-0.1 (7.8)

Table 2: Minimum important difference and responsiveness based on 1 year change summary using the SF2 question (mean[sd])

		Somewhat worse (n = 100)			Somewhat better (n = 53)		
		MID	ES	SRM	MID	ES	SRM
Utility Measure	EQ-5D	-0.10 (0.26)	-0.41	-0.38	0.04 (0.14)	0.25	0.27
	SF-6D	-0.04 (0.10)	-0.36	-0.41	0.04 (0.10)	0.32	0.39
	ES/SRM Ratio (EQ-5D: SF-6D)		1.14	0.93		0.78	0.69
	EQ-5D VAS	-8.9 (18.5)	-0.56	-0.48	6.8 (21.1)	0.43	0.32
OSRA	OSRA-A	0.98 (1.95)	0.53	0.51	-0.79 (1.7)	-0.48	-0.46
	OSRA-D	0.38 (1.42)	0.28	0.27	-0.34 (1.1)	-0.23	-0.30
Disease Activity	DAS28	0.62 (1.1)	0.49	0.55	-0.31 (0.9)	-0.25	-0.33
	28 Swollen Joint Count	0.7 (5.1)	0.16	0.14	-0.4 (4.2)	-0.09	-0.09
	28 Tender Joint Count	2.1 (6.6)	0.33	0.33	-1.7 (5.3)	-0.31	-0.32
	Patient Global Assessment (mm)	9.1 (24.1)	-0.55	0.38	-2.0 (20.1)	0.13	-0.10
	Physician Global Assessment (mm)	8.4 (21.9)	0.45	0.38	-8.0 (15.3)	-0.45	-0.52
	Pain VAS (mm)	10.7 (20.4)	0.54	0.52	-2.6 (22.5)	-0.12	-0.12
	Fatigue VAS (mm)	7.1 (24.8)	0.30	0.28	-3.7 (20.1)	-0.14	-0.19
Physical Function	HAQ	0.15 (0.43)	0.24	0.35	<0.01 (0.43)	<0.01	<0.01
SF-36	MCS	-3.8 (10.6)	-0.32	-0.36	1.8 (8.7)	0.17	0.21
	PCS	-3.1 (7.0)	-0.38	-0.45	3.2 (7.6)	0.34	0.42

Abbreviations: HAQ = Health Assessment Questionnaire; VAS = Visual analogue scale

Table 3: Minimum important difference and responsiveness based on 1 year change summary using EQ-5D VAS (mean[sd])

		>20% & <50% worse (n=81)			>20% & <50% better (n=62)		
		MID	ES	SRM	MID	ES	SRM
Utility Measure	EQ-5D	-0.13 (0.24)	-0.62	-0.51	0.06 (0.27)	0.22	0.21
	SF-6D	-0.04 (0.10)	-0.35	-0.37	0.03 (0.09)	0.25	0.29
	ES/SRM Ratio (EQ-5D: SF-6D)		1.94	1.38		0.88	0.72
	EQ-5D VAS	N/a	N/a		N/a	N/a	N/a
OSRA	OSRA-A	0.73 (1.95)	0.43	0.38	-0.37 (1.67)	-0.21	-0.22
	OSRA-D	0.49 (1.31)	0.42	0.38	-0.10 (1.24)	-0.06	-0.08
Disease Activity	DAS28	0.53 (0.99)	0.48	0.53	-0.03 (0.92)	-0.24	-0.37
	28 Swollen Joint Count	1.3 (4.53)	0.29	0.28	0.2 (5.0)	0.03	0.03
	28 Tender Joint Count	2.4 (5.7)	0.44	0.41	-0.6 (6.9)	-0.07	-0.08
	Patient Global Assessment (mm)	8.8 (22.3)	0.52	0.39	-0.5 (21.8)	-0.03	-0.02
	Physician Global Assessment (mm)	6.0 (17.5)	0.35	0.34	-3.0 (19.0)	-0.17	-0.16
	Pain VAS (mm)	9.8 (22.4)	0.45	0.44	-1.1 (20.1)	-0.05	-0.06
	Fatigue VAS (mm)	6.2 (26.8)	0.25	0.23	-0.3 (29.0)	-0.01	-0.01
Physical Function	HAQ	0.13 (0.48)	0.21	0.27	0.07 (0.33)	0.11	0.23
SF-36	MCS	-5.3 (11.4)	-0.48	-0.47	0.29 (8.4)	0.02	0.03
	PCS	-1.2 (7.8)	-0.14	0.16	2.0 (6.7)	0.23	0.30

Abbreviations: HAQ = Health Assessment Questionnaire; VAS = Visual analogue scale

Table 4: Power calculation for the prospective responsiveness study(24)

			MID		
		H_0	H_1	SD H_1	Sample size
SF2	SF-6D	0	0.04	0.10	66
	EQ-5D	0	0.04	0.14	129
VAS	SF-6D	0	0.03	0.09	95
	EQ-5D	0	0.06	0.27	213

Power 0.90; H_0 = Null hypothesis, H_1 = alternative hypothesis; SD = standard deviation

Table 5: Comparison of results from this study with reported effect sizes and standardised response mean from the literature

		Responsiveness Statistics			
		Effect Size[a]		Standardized response mean[b]	
Definitions	Study/Author	EQ-5D	SF-6D	EQ-5D	SF-6D
Improvement					
Self-report	BROSG	0.25	**0.32**	0.27	**0.39**
Change in VAS	BROSG	0.22	**0.25**	0.21	**0.29**
Treatment	Russell et al	0.67	**1.40**	0.64	**0.87**
Self-report	Conner-Spady et al	**0.53**	0.36	-	-
Self-report	Marra et al	0.15	**0.31**	0.20	**0.36**
Patient global	Marra et al	0.36	**0.54**	0.43	**0.62**
Deterioration					
Self-report	BROSG	**-0.41**	-0.36	-0.38	**-0.41**
Change in VAS	BROSG	**-0.62**	-0.37	**-0.51**	-0.37
Self-report	Conner-Spady et al	**-0.58**	-0.24	-	-
Self-report	Marra et al	**-0.16**	-0.08	**-0.19**	-0.13
Patient global	Marra et al	**-0.55**	-0.24	**-0.63**	-0.35

Figures in **bold** denote the most responsive measure; [a] mean difference divided by standard deviation at baseline; [b] mean difference divided by the standard deviation of difference; [c] $({}^{t}\text{statistic}_1/{}^{t}\text{statistic}_2)^2$ where ${}^{t}\text{statistic}_1$ = alternative measure, and ${}^{t}\text{statistic}_2$ = gold standard (in this case defined as RAQoL)