Comparison of EuroQol-5D and Short Form-6D Utility Scores in Multiethnic Asian Patients with Psoriatic Arthritis: A Cross-sectional Study

Ying-Ying Leung, May-Ee Png, Hwee-Lin Wee, and Julian Thumboo

**ABSTRACT.** Objective. To compare EuroQol-5D (EQ-5D) and Short Form-6D (SF-6D) utility scores in multi-ethnic Asian patients with psoriatic arthritis (PsA).

Methods. Consecutive patients fulfilling the Classification Criteria for Psoriatic Arthritis attending a rheumatology outpatient clinic were recruited and completed the EQ-5D and SF-6D questionnaires. Comparisons were performed by score distribution, mean, median, and the Outcome Measures in Rheumatology filter: i.e., truth, discrimination, and feasibility.

Results. Eighty-six patients were enrolled (69 English-speaking and 17 Chinese-speaking; male:female ratio 0.91). The score distribution of SF-6D was normal, while that of EQ-5D was bimodal. A ceiling effect was observed in 20% of patients for EQ-5D and none for SF-6D. There were moderate correlations (Spearman’s rho = 0.59, p < 0.0001) between the 2 scores, but poor agreements on scatterplot, intraclass correlation (ICC 0.43 and standardized ICC 0.21), and Bland-Altman plots. EQ-5D generated lower utility scores than SF-6D in the poorer health subgroup. SF-6D had stronger correlation with the general health status and other external measures of health; and it distinguished better between good and poor general health status, with better effect size and relative efficiency statistics. EQ-5D demonstrated higher patient acceptability.

Conclusion. EQ-5D and SF-6D instruments generated different utility scores in PsA. SF-6D may be superior because of normal scaling distribution and the absence of ceiling and floor effects. SF-6D also had better construct validity and better discrimination of poor health status. More studies are required for cost-utility analysis in PsA. (First Release March 15 2013; J Rheumatol 2013;40:859–65; doi:10.3899/jrheum.120782)

**Key Indexing Terms:**

- PSORIATIC ARTHRITIS
- COMPARATIVE STUDY
- COST BENEFIT
- QUALITY-ADJUSTED LIFE-YEAR

Psoriatic arthritis (PsA) has deleterious effects on joints and skin, causing joint deformities, impaired physical function, and impaired health-related quality of life (HRQOL). The introduction of anti-tumor necrosis factor-α (anti-TNF) therapies has dramatically changed the treatment paradigms for PsA. Given that health resources are finite, this change highlights the importance of cost-utility analysis (CUA), the primary outcome of which is cost per quality-adjusted life-year (QALY). Indirect HRQOL measures, such as the EuroQol-5D (EQ-5D) and the Short form-6D (SF-6D), are commonly used to elicit health state values for calculating QALY. Both instruments measure health in terms of physical function, pain, and mental health. Also, both have a scoring function derived from the statistical modeling of preferences for multideficit health states elicited from the general population of the United Kingdom. Both instruments classify a respondent’s self-reported health status according to a specific descriptive or classification system and assign a utility score. A utility score of 1 represents a state of perfect health and a utility score of 0 represents being dead.

Differences between these 2 instruments arise from a combination of differences in descriptive systems andvaluations attached to the health states, and these may lead to different utility scores when applied to the same patient. These differences have been noted in other disease groups, including patients with rheumatic diseases. Although gaps have been noted in the existing literature, to date no conclusion has been drawn with regard to which instrument performs better. This highlights the necessity for
such comparisons to be made in a wider spectrum of diseases and sociocultural contexts.

The differences between the EQ-5D and SF-6D have also been observed in PsA cohorts that have poor health status and have undergone anti-TNF therapies. However, the difference between the 2 utility scores has not been demonstrated in the general population of patients with PsA. It should be noted that the SF-6D has an advantage in PsA because it can be derived from the Medical Outcomes Study Short Form-36 (SF-36), which has been commonly used and extensively validated in PsA cohorts. The aim of our study was to compare the utility scores EQ-5D and SF-6D in a population of multiethnic Asian patients who have PsA, particularly with regard to the aspects of truth, discrimination, and feasibility corresponding to the Outcome Measures in Rheumatology (OMERACT) filter. We compared the 2 instruments for their distribution and agreement, construct validity, discriminant capacity, and acceptability.

MATERIALS AND METHODS

Data source and collection. From June 2010 through October 2011, we recruited consecutive patients with PsA [based on the Classification Criteria for Psoriatic Arthritis (CCP)] who attended an outpatient clinic at the rheumatology center at Singapore General Hospital. Patients of different ethnicities (Chinese, Malay, Indian, and others) were recruited. All participants completed identical and validated questionnaires, either the Singapore (English) or the Singapore (Simplified Chinese) version of the SF-36 Health Survey, version 2.0 (SF-36v2), and the EQ-5D, and provided sociodemographic data. The study was reviewed and approved by the Institutional Review Board of SingHealth. Before entry into the study, participants were informed of its nature and purpose and each participant signed an informed consent form.

Instruments. The EQ-5D is a standardized tool to measure health state. It has a 20 cm visual analog scale (EQ-VAS), which records the respondent’s self-rated health as a score from 0 to 100 (0 representing worst imaginable health state and 100 best imaginable health state), and a descriptive system comprising 5 health domains (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression). The 5-domain descriptive system classifies 243 different health states using 3 levels of severity for each domain (no problems, some problems, extreme problems). The EQ-5D tariff was estimated using the time tradeoff (TTO) method in a sample of 3395 respondents from the UK general population. It consists of a set of numbers that indicates the level of HRQOL for each EQ-5D health state, on a scale from 1 (full health) to 0 [dead; range −0.594 to 1, where negative values are valued as worse than dead (WTD)]. The EQ-5D is commonly used in quantifying the influence of medical interventions on HRQOL. By comparing the difference in EQ-5D health states before and after treatment, analysts calculated the treatment effectiveness in QALY.

Both the English and Chinese versions of the EQ-5D have been validated in Singaporean patients with rheumatic diseases and measurement equivalence has been demonstrated for the Singapore English and Chinese versions. Favorable CUA for anti-TNF therapy in PsA has been reported.

The SF-6D is a classification tool for describing health as derived from the 7 of the 8 health domains that are covered by the SF-36v2 health survey. It has 6 multilevel domains: physical functioning, role participation (combined role-physical and role-emotional), social functioning, bodily pain, mental health, and vitality. Each of the 6 domains has 4–6 levels of response, thus the SF-6D describes 18,000 health states. A preference tariff was estimated using the standard gamble (SG) method to obtain utility values from the UK general population on 249 of the possible health states. The resulting SF-6D index, which ranges from 0.296 (worse health state) to 1.0 (best health state), was used in the assessment of QALY and the cost-effectiveness of various healthcare interventions. A few studies in PsA have demonstrated improvement in SF-6D scores with anti-TNF treatment beyond the minimal clinically important difference (MID) of 0.0320,21. The English and Chinese versions of the SF-6D have been demonstrated to be equivalent in the Singaporean population including patients with rheumatic diseases.

Statistical analysis. Collected data were entered into a Microsoft Excel spreadsheet and analyzed using SPSS software, version 17.0 (SPSS Inc.) and Stata/SE software version 11.0 (StataCorp LP). All statistical tests were 2-tailed and conducted at a 5% level of significance.

EQ-5D and SF-6D utility scores were calculated using the scoring algorithms developed by Dolan and Brazier, et al, respectively. Missing data in SF-36v2 were estimated according to protocol. We compared the distribution, mean, median, and agreement between the utility scores. For distribution, we compared the mean ± SD and median [interquartile range (IQR)] utility scores generated by the 2 instruments. Ceiling and floor effects were considered present if > 15% of participants responding gave the highest and lowest possible scores. The within-subject differences in the 2 utility scores were compared by t test. A limit of 0.03 between scores was chosen based on the smallest estimate of the published results of MID for the EQ-5D and SF-6D. To assess the degree of agreement between the 2 utility scores, we used the interclass correlation coefficient (ICC; 2-way random-effects model with absolute agreement) and a Bland-Altman plot. An ICC > 0.7 suggests an acceptable level of agreement. As the lower bound of the 2 utility scores differs, we standardized the utility scores linearly to fit the range of 0–1 (based on the theoretical possible range).

For the aspect of truth, we compared the construct validity using the Spearman’s rank correlation between the 2 utility scores and the SF-6D summary scores, physical component summary (PCS), mental component summary (MCS), and SF-6D general health (SF-GH). The SF-GH is the participant’s response to the first question of the SF-36: “In general, would you say your health is —?” and it is not included in the calculation of the SF-6D.

For the aspect of discrimination or discriminant capacity, we examined the discriminatory capacities of the 2 utility scores to distinguish between participants with contrasting health states. Participants with different levels of impairment were classified according to the SF-GH. The SF-GH was analyzed with the following categories: “excellent/very good,” “good,” and “fair/poor.” The ability to differentiate between SF-GH “excellent/very good” versus “good” and “good” versus “fair/poor” subgroups was calculated using 1-way ANOVA. The effect size was calculated as the standardized mean difference described by Cohen (i.e., the difference in mean scores divided by the pooled SD). The effect size was categorized as small (0.2–0.5), moderate (0.5–0.8), or large (0.8). The influence of sociodemographic factors was evaluated by t test or 1-way ANOVA. The ability of the utility scores to detect difference between health status of SF-GH was tested by the relative effective statistics (RE). This is calculated as the ratio of the standard deviation of the SF-6D utility score to the standard deviation of the SF-5D. An RE > 1.0 indicates that SF-6D is more efficient than the EQ-5D in detecting the difference. The reverse is true if the RE is < 1.0. We evaluated the RE of the 2 utility scores in differentiating health status according to SF-GH, “excellent/very good” versus “good” and “good” versus “fair and poor.” For “feasibility,” we reported the proportion of missing data for each utility score.

RESULTS

Characteristics of participants. Eighty-six participants (69 English-speaking and 17 Chinese-speaking) were enrolled into the study. Characteristics of participants in each ethnic group and in the total sample are shown in Table 1.
HRQOL among participants with PsA was much lower than that found among the healthy “normal” population: both the norm-based SF-36 PCS and MCS were below the norm mean of 50. The characteristics of PsA participants from different ethnicities were generally similar, except that the EQ-V AS was poorer among Indian participants with PsA.

**EQ-5D and SF-6D distribution and agreement.** The score distribution for the SF-6D was normal (skewness = 0.27, kurtosis = –0.37, p = 0.10 by the Kolmogorov-Smirnov test), while that of the EQ-5D was bimodal (skewness = –1.62, kurtosis = 2.80, p < 0.001 by Kolmogorov-Smirnov test; Figure 1). The 3 most commonly reported EQ-5D profiles were 11111 (20%), 11121 (29.4%), and 11122 (10.6%), whereas the reported SF-6D profiles were spread across all states, none of which was reported by > 2 participants (2.4%). A ceiling effect was observed in EQ-5D (range –0.014 to 1.0), where 20% of participants responded with the highest possible score and 2.3% of participants had negative scores for EQ-5D corresponding to the WTD state. No ceiling or floor effects were observed for the SF-6D (range 0.355 to 1.0). The mean (± SD) EQ-5D and SF-6D utility scores were 0.74 (± 0.24) and 0.68 (± 0.13), respectively (p = 0.001). The median EQ-5D and SF-6D utility scores were 0.8 (IQR 0.09) and 0.64 (IQR 0.18). There was a mean difference of 0.05 (± 0.2) between the utility scores, which was higher than the MID of 0.03 that we chose for comparison. The paired utility scores are presented graphically as a scatterplot for the entire population (Figure 2); the Spearman product-moment correlation coefficient between EQ-5D and SF-6D was 0.59 (p < 0.0001). The deviation from the 45-degree line was evident, particularly in the low end of the utility scores. In Bland-Altman plots, EQ-5D scores were systematically lower than the SF-6D in subjects with lower averaged utility scores (Figure 2). Poor agreement between EQ-5D and SF-6D utility scores was demonstrated with the low ICC (0.43, 95% CI 0.23 to 0.59) for the entire population. The ICC was even lower when it was standardized (0.21, 95% CI –0.09 to 0.50). There were wide limits of agreement on the Bland-Altman plot (1.96 SD from –0.35 to 0.46; Figure 2).

**Aspect of truth.** For construct validity, the Spearman’s rho correlation between the EQ-5D and SF-36v2 summary scores and SF-GH were moderate (0.37–0.45), while the correlations for the SF-6D with SF-GH status were good (0.57–0.84; Table 2). There were similar correlations between the 2 utility scores and the EQ-V AS. At the domain level, there were moderate correlations between the EQ-5D and SF-6D domains that measure similar constructs. The Spearman’s rho were 0.53 (p < 0.001) between SF-6D pain and EQ-5D pain/discomfort, and 0.48 (p < 0.001) between SF-6D mental health and EQ-5D anxiety/depression. SF-6D physical functioning correlated weakly with EQ-5D mobility, self-care, and usual activities (r = 0.33–0.39), which reflects that these domains measured different aspects of HRQOL.

**Aspect of discrimination.** The utility scores did not differ by sex, age, marital status, ethnicity, languages, or education level. For discriminant capacity, the SF-6D distinguished between participants with “better” or “poor” health status, with strong effect sizes. The EQ-5D distinguished partici-
pants with “good” versus “fair/poor” health status with only moderate effect size (Table 3). Considering the difference between “excellent/very good” and “good” of the SF-general health, the RE score was 0.98, implying that the SF-6D was as efficient as the EQ-5D in differentiating these 2 health status. Also, the RE was 1.95 for differentiating “good” from “fair/poor” of the SF-GH. This implied that SF-6D was 95% more efficient than EQ-5D in identifying patients with “good” or “poor” general health status.

Aspect of feasibility. The EQ-5D has 5 items, which is much fewer than the SF-6D, which is derived from the 36 items of the SF-36. The SF-6D individual items are rarely presented in isolation. The acceptability in terms of completion rate is higher in the EQ-5D. Missing raw data was 1.2% for the EQ-5D. It was 9.3% for SF-6D, and was then reduced to 3.5% by estimating the missing data according to the SF-36v2 protocol.

DISCUSSION

Using an appropriate and valid utility index is a major issue in cost-utility analysis. Therefore, it is important to understand the performance of different indirect utility instruments in various diseases. In this study, we presented and compared utility data in terms of distribution, agreement, and the OMERACT filter, which comprises truth, discrimination, and feasibility, in a cohort of multiethnic Asian patients with PsA recruited from a secondary and tertiary referral rheumatology center. We found a 0.05 difference in the utility scores generated by the 2 instruments, which is...
Significantly lower utility scores with EQ-5D compared to with EQ-5D, which corresponded to a status of WTD.

12% of the PsA participants reported a negative utility score.

...generate lower utilities than SF-6D in subgroups with poorer health, and that the reverse is true in the healthier subgroups. Our study also revealed similar findings in...subgroups.8,32,33. Our study also demonstrated poor agreement and ICC...between good versus poor subgroup.

The results of an economic model36. Although the HRQOL of our cohort was much worse than that of the general population, the general health status of our participants was better than that reported by Adams, et al, and the proportion of participants with the WTD state in our study cohort was only 2.3%. However, the large difference between the utilities generated by these 2 instruments was still present. This implies that the difference between the 2 utility scores is not just related to the preference-based weights and methods of handling the WTD status. In Singapore, where there is no primary healthcare that supports the care of chronic inflammatory arthritis, the majority of patients with PsA are cared for in secondary and tertiary centers. Hence, we believe that our sample represents the whole spectrum of illness in PsA.

It is well known that different indirect utility instruments may yield different utility scores37,38. Many attributed this to the different methods by which these health preference utility scores were derived (TTO for EQ-5D and SG for SF-6D). The SF-6D describes more health states than does EQ-5D (18,000 vs 243 health states) and therefore may capture more health states at the extreme ends of the range and may capture smaller health changes39. However, only a minority of these states have been valued by SG, and not all the states were valued by TTO when weights of the EQ-5D were calculated. This means that most of the health states carry utility values that were estimated from the utility function rather than being measured directly. Moreover, it should be emphasized that the utility scales are anchored at 0 and 1, while they are by no means representing variables lying on interval scales. The weak interval properties of various utility scores generated from indirect HRQOL instruments have been illustrated in a large-scale comparison40. Indeed, the descriptive systems in different HRQOL differ widely in their coverage of different dimensions of HRQOL and thus the reported differences in utility scores are attributable, in part, to these differences.

There has been debate whether the valuation of a health state may be a better reflection of the true welfare value that is associated with health41, while criticisms are that utility scores and general health evaluations are basically measure-
ments of different constructs. In AS, Boonen, et al8 have shown that the EQ-5D, SF-6D, and EQ-VAS correlated well with external health measurements, with only moderate agreement; the disagreement between EQ-5D and EQ-VAS was more prominent in the subgroup with poorer health. There is, however, growing interest using the general health measure as a composite measurement of disease activity in PsA. The PsA index, which consisted of patient global assessment, skin global assessment, and physician global assessment, was found to explain 90% of the variance in the baseline scores of the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) Composite Exercise project42, and was taken forward for further evaluation in the GRAPPA 2010 and OMERACT 11 meetings43. In PsA, patient global assessment (PGA) has been shown to have reasonable reliability44 and construct validity.45 We observed better ICC and limits of agreement for the SF-6D with EQ-VAS (data not shown), which may imply that the SF-6D provides a better reflection of general health. Although the EQ-VAS is a similar general health measurement, it has not been evaluated as a valid measure of general health status or composite disease activity score in PsA. Therefore, this finding requires further validation.

The measurement of HRQOL in PsA is under investigation, and there have been limited data on using indirect HRQOL measures for CUA in PsA. Our study adds information to the literature on the comparison of performance of these 2 utilities in PsA. In terms of “truth,” the SF-6D utility score performed better, with higher correlation with SF-36 summary scores and SF-GH. Caution in interpretation is needed, in that the SF-6D and summary scores were derived from the same instrument, the SF-36. However, SF-GH was not included in the calculation of the SF-6D. In terms of “discrimination,” the SF-6D utilities performed better in differentiating participants with poorer health status in terms of effect sizes and the RE scores. Gaujoux-Viala, et al30 demonstrated similar systematic disagreement between EQ-5D and SF-6D in a prospective study on 813 patients with early arthritis, especially in patients with worse clinical outcomes. The SF-6D was shown to have better RE statistics30. In their longitudinal study over 2 years, the same group of authors also recently demonstrated better responsiveness for improvement with the SF-6D in standardized response mean and effect sizes31.

For “feasibility,” the SF-6D adds more missing values than the EQ-5D, although the missing values could be minimized by substituting with the average from the other items in the same subscale. For distribution, the SF-6D was better in terms of having normal scaling distribution and the absence of ceiling and floor effects. However, further research is required to determine which instrument performs better.

There are several limitations to our study. First, the construct validity assessments are limited to the SF-GH. No other assessments of disease activity, such as joint count or Health Assessment Questionnaire data, were collected. Second, this was a cross-sectional study and the discrimination evaluation is limited to how the utility scores differentiate different health states, instead of a change in status with treatment over time. Third, we did not address the reliability for these utility scores in PsA. Fourth, our study cohort consisted of PsA patients with a long duration of illness, recruited from a single center, which limits the study’s generalizability to patients with early PsA. Our sample size was relatively small and this may affect the interpretation of ceiling and floor effects. Moreover, HRQOL measures are heavily influenced by comorbidities that may introduce inaccuracy. Finally, the scoring algorithms used for both instruments were developed from the UK general population, because no such algorithm was available in Singapore at the time of the study.

Despite these limitations, we have demonstrated in a cohort of patients with PsA that the SF-6D performed slightly better in terms of construct validity and discrimination than the EQ-5D, which is shorter and more feasible in clinical practice. The SF-6D also had normal distribution and lack of ceiling effects. However, the 2 instruments yielded different utility scores in PsA. This would have a great effect on QALY estimates, and it highlights the importance of choosing the appropriate instrument for cost-effectiveness evaluation. Additional research is needed to determine whether the EQ-5D or the SF-6D is the better instrument for cost-utility analysis in PsA.

ACKNOWLEDGMENT
The authors appreciate the support of the Duke-NUS/SingHealth Academic Medicine Research Institute and the editorial services of Jon Kilner, MS, MA, Pittsburgh, PA, USA, and Taara Madhavan, Associate, Duke-NUS Graduate Medical School, Singapore.

REFERENCES


