The Impact of a Revised EQ-5D Population Scoring on Preference-Based Utility Scores in an Inflammatory Arthritis Cohort

Roisin Adams, MSc1,*, Benjamin M. Craig, PhD2, Cathal D. Walsh, PhD1,3, Douglas J. Veale, MD4, Barry Bresnihan, PhD4, Oliver FitzGerald, PhD4, Michael Barry, PhD1

1National Centre for Pharmacoeconomics, St. James Hospital, Dublin, Ireland; 2Health Outcomes and Behavior Program, Moffitt Cancer Center, Tampa, FL, USA; 3Department of Statistics, Trinity College Dublin, Dublin, Ireland; 4St. Vincents University Hospital, Dublin Academic Healthcare, Dublin, Ireland

ABSTRACT

Background and Objective: It is well established that there are problems with the EQ-5D. This is due to the original scoring methods used and how negative time trade-off (TTO) values were treated. A revised scoring method has been published. This article applies this to an inflammatory arthritis cohort. The objective is to examine the impact of a revised scoring system for the EQ-5D (UK) TTO on the utility estimates and in the case of rheumatoid arthritis, to explore the impact of using different utility metrics on the incremental cost-effectiveness ratio (ICER) results of an economic model. Methods: A total of 504 patients with inflammatory arthritis were rescored using revised EQ-5D scoring, which uses an episodic random utility model to deal with negative TTO values. Differences in utility scores were compared and the new mapping coefficients were obtained. These were then used in an economic model to examine the impact on the ICER. Results: In rheumatoid arthritis, the overall change is less for the revised EQ-5D scoring than with the original EQ-5D (TTO) but greater than the SF-6D: EQ-5D UK −0.22 (95% confidence interval [CI] −0.30 to −0.15), revised EQ-5D UK −0.16 (95% CI −0.21 to −0.10) and SF-6D −0.08 (95% CI −0.11 to −0.05). A similar trend is seen in the psoriatic arthritis group. The economic model produced different ICERs, when different utility measures were used; EQ-5D (TTO) €42,402, SF-6D €111,788, and revised EQ-5D (TTO) €57,747. Conclusion: In the context of inflammatory arthritis, this article demonstrates that a revised scoring for EQ-5D may have a significant impact on utility estimates and on the output of the economic model. Keywords: EQ-5D, scoring, SF-6D, utility.

Copyright © 2011, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

Introduction

Rheumatoid arthritis (RA) and psoriatic arthritis (PsA) are part of a group of conditions described as inflammatory arthritis. They are chronic, progressive conditions that place a substantial burden on patients, their caregivers, and the health service. The most common of these conditions are RA and PsA. RA, in particular, has a negative effect on quality of life (QOL), including physical, psychological, and social functioning, and is associated with premature mortality [1]. PsA presentation ranges from mild impairment of function that can be adequately treated with mild therapeutic interventions to severe disease with erosive arthropathy that may result in significant functional disability and increased mortality [2].

Measurement of patients’ response to treatment via QOL instruments is now one of the recommended methods for quantifying the effectiveness of a technology in economic evaluations around the world [3–6]. QOL is quantified using a single index figure anchored between 0 and 1, with 0 indicating immediate “death” and 1 indicating “perfect health.” This figure has been termed a utility value. QOL can be measured using direct and indirect methods. Direct measurement asks a person to value directly his or her own health or a relevant disease-specific state using a valuation task such as time trade-off (TTO) or standard gamble (SG). Indirect measurement involves QOL questionnaires, for example, EQ-5D or SF-6D, in which the patient completes questions on his or her current health state, and these responses are scored using weights or preferences obtained from the general population. There are documented differences in the utility values produced in these ways with indirect methods giving consistently lower levels of utility than direct methods [7]. Indirect generic measures such as the EQ-5D [8] and the SF-6D [9] are the most commonly used, and a single index measurement of QOL can be derived.
A variety of instruments have been used to measure both disease severity in RA and PsA and the impact of this severity on QOL, particularly in inflammatory arthritis trials [10–13]. These measures include clinical instruments that measure disease activity, such as the European League Against Arthritis disease activity score (DAS) [14], disease-specific instruments such as the Health Assessment Questionnaire (HAQ) [15], and generic instruments such as the EQ-5D [8] and SF-36 [16].

All instruments display some shortcomings in assessing health-related QOL (HR-QOL) in inflammatory arthritis [17,18]. Although using generic measures should theoretically allow us to compare results for a variety of conditions, disparities have been shown to exist in the utilities derived from the EQ-5D and SF-6D, and this is attributed to the different descriptive systems, the values attached to the health states, or a combination of these [19]. This has important implications for economic analyses of treatments such as biological therapy in inflammatory arthritis, which is more likely to be used for patients in severe health states than those in mild health states. If we cannot measure the change in these categories accurately, the full potential of the treatment may not be adequately measured.

It is well established that the EQ-5D and the SF-6D produce different utility values in the same cohort [20,21]. This is in part due to different definitions of perfect health. According to the 1995 Health Survey of England, the EQ-5D considers more than half of the population to be in perfect health, whereas the SF-6D considers less than 3% to be in perfect health [22].

Therefore, the SF-6D has a different criterion for perfect health than the EQ-5D. This presents decision makers with a challenge in comparing results of economic evaluations that have used different methods to calculate utility. In an RA cohort, the utility gain produced by the EQ-5D is twice that produced by the SF-6D [20,23]. This discrepancy between the measures has been the subject of a number of recent publications that highlight [20,24–28] the methodology of the original scoring of the EQ-5D (UK) and the manner in which worse-than-dead (WTD) values were adjusted [29]. Using an inflammatory arthritis cohort, a recent study found that the floor utility value measured by the EQ-5D was –0.43 and that 17% of utility values in this cohort were WTD. The prevalence of negative values had a profound effect on the burden of disease estimates in this article [20]. The lowest measured value in the group for SF-6D was 0.29.

In an attempt to lessen the heterogeneity between these two measures, we used a revised scoring method for the EQ-5D and rescored a rheumatology cohort receiving biological therapy [26]. We compared the scores to the utility values produced by the original scoring methods for both RA and PsA. We then used the mapping coefficients calculated for RA in a RA model and compared the incremental cost-effectiveness ratio (ICER) results using different measures. The aim of our analysis is to examine the impact of a revised scoring system on the utility estimates and, in the case of RA, on the results of an economic model.

Methods

Data source
Utility data were derived from a database of 504 patients from a tertiary referral center in the Irish health-care setting that records the clinical and QOL outcomes of patients receiving biological therapy for RA and PsA. Patients included in this study have a diagnosis of either RA according to the American College of Rheumatology criteria or PsA according to CASPAR (Classiﬁcation criteria for Psoriatic Arthritis) criteria and commenced biological therapy (anti–tumor necrosis factor–α monoclonal antibodies, B-cell antagonists, or T-cell modulators).

Instruments used
The QOL instruments used were the paper versions of the EQ-5D (3 level), SF-36 (version 1), and the modified HAQ. The EQ-5D, SF-36, and HAQ were collected as part of normal clinical practice for monitoring the impact of treatment on QOL. All questionnaires were measured at baseline before the start of therapy and at follow-up at 12 months.

The Disease Activity Score 28 (DAS28) was collected as one of the clinical outcomes in monitoring response to treatment and disease activity [14]. This instrument incorporates the number of both swollen and tender joints, a laboratory measure of inflammation such as the erythrocyte sedimentation rate and C-reactive protein (CRP), and a patient-assessed global disease impact measure. The DAS28 first developed for RA incorporates just 28 of the 68 joints. The DAS28 is an index that can assess disease activity and also be used to derive a response measure [30].

The HAQ is the physical disability scale (Modified HAQ) that measures function in relation to the degree of difficulty experienced in performing activities of daily living such as dressing, rising, personal hygiene, walking, eating, and the ability to carry out chores. The HAQ contains 20 items across 8 domains that are scored from 0 (no difficulty) to 3 (unable to do) [15]. The HAQ Disability Index (HAQ-DI) is the single index score derived from scoring the HAQ.

The SF-6D is derived from the SF-36 and uses 11 items from the 36 items of the RAND Medical Outcomes Study short-form health survey (SF-36) [9,16]. Scoring data for the SF-6D were collected using the SG valuation technique on a random sample (n = 836) of the general population in the United Kingdom [9]. The SF-6D scoring algorithm was revised in 2007 using nonparametric bayesian analysis [31]. The bayesian version overcomes some of the bias of the original regression models when assigning values to the worst health states (e.g., it yields a value of 0.203 for the worst SF-6D state compared to 0.301 using the original parametric algorithm). The bayesian estimates of the SF-6D utilities were calculated in Microsoft Excel (further details of the methodology are available from the University of Sheffield) [32].

Methods used to calculate utilities
The EQ-5D index is a preference-based index measure in which an individual provides an assessment of each component of his/her health status according to a structured health-status classification system, and a single preference-based score is derived for each individual based on societal preferences [29]. It is used extensively to measure QOL in inflammatory arthritis [33–35]. The EQ-5D has five dimensions, each with three levels of severity (level 1, no problems to level 3, extreme problems). Therefore, the instrument can produce 243 health states, 35. An additional two health states are included: “dead” and “unconscious.”

Scoring method for the UK TTO
The preferences for the scoring function were measured using the TTO technique on a random sample of 2997 adults of the UK population (Measurement of Health [MVH] study) [29]. Dolan [29] devised a scoring method that assigned a single index utility value for each health state described. Forty-five of the health states were scored directly from the population using TTO valuation, and the values of the remaining states were predicted using regression estimates.

To anchor the scale, perfect health and dead were assigned scores of 1 and 0, respectively. For states described as better than dead (BTD) (>0) on the TTO, scores were calculated using the formula $^x/10$, where $x$ is the number of years spent in perfect health equal to 10 years in the health state. For states scored as WTD (<0), the formula given is $^x/(10 - x)$, where immediate death equates to a scenario of $x$ years in perfect health followed by $(10 - x)$ years in...
the health state. For states BTD, the ratios range from 1 to 0, but ratios for WTD states lie between 0 and −39 (the WTD x has an upper bound at 9.75 years). The asymmetry seen between the positive and negative ratios seem to inflate the influence of the WTD responses; therefore, Dolan transformed the negative ratios to $-x/\alpha$, replacing 34% of the TTO responses [36]. By bounding the negative ratio at $-1$, the influence of these WTD responses on the mean slope lessened and improved face validity of mean ratio estimates.

Revised scoring method for the EQ-5D UK

To provide an alternative method for handling the challenges posed in valuing the WTD states in the MVH study, Craig and Busschbach [26] re-examined the original data using an episodic regression model instead of a ratio regression model. The health state valuations have been published and these are provided in Appendix 1 in found at doi:10.1016/j.jval.2011.03.002 [37]. The theoretical basis for both models was presented in a previously published paper [38]. The utility of a health state over time $t$ for an individual $i$ is random and can be represented by:

$$U_{ij}(t) = \begin{cases} \mu_{ij} + e_{ij} & \text{Episodic RUM} \\ \mu_{ij} + e_{ij} t & \text{Instant RUM} \end{cases}$$

The main distinguishing factor between these models is how the WTD TTO responses are interpreted, and this differs greatly between the episodic and the instant. In the episodic random utility model (RUM), the error represents variability in the value of an episode (error associated with time). In the instant RUM, the error represents variability of an instantaneous state, not the episode, and suggests a random slope with respect to time. The regression model of the episodic RUM treats the time in perfect health as the dependent variable and the time in the health state as the independent variable. The coefficient is the value estimator, not a mean ratio. The central advantage of the episodic RUM over the original approach is that this procedure does not involve arbitrary transformations of the WTD responses.

**Statistical analysis**

Descriptive statistics were used to describe the baseline demographics; mean values, range, and SD are given. A paired-sample $t$ test was used to compare the mean utility at baseline and at follow-up and the change measured by the original EQ-5D UK TTO, the revised EQ-5D UK, and the SF-6D. Confidence intervals (CIs) (95%) are presented around the change in utility and the ICER estimates. Statistical analysis was performed using Statistical Package for Social Sciences (SPSS) Version 16 (SPSS Inc., Chicago, IL).

**Statistical models used**

General linear models were fitted for each of the measures with HAQ-DI and DAS28. Quadratic and higher dimensional models were examined but did not have a statistically significantly improved fit than the linear model. For each of the regression models, standard errors, 95% CIs and $R^2$ are shown.

**ICER calculation**

To calculate an ICER for the RA group, we used the mapped coefficients of the three dependent variables (original EQ-5D (TTO), revised EQ-5D (TTO), and the SF-6D) to populate an RA model. The model was run separately for each of the three methods. In the RA model, disease changes are driven by changes to patients’ HAQ-DI scores, and therefore it was the HAQ-DI and utility coefficients that were used. The model incorporates a linear equation to model the utility change as mapped from the HAQ-DI. Although newer models now use a quadratic equation to describe the relationship between utility and HAQ-DI, in this case, there was no significant statistical difference between the quadratic model and the linear model [39]. The model

---

**Table 1 - Baseline demographics.**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>RA (n = 345), mean ± SD (range)*</th>
<th>PsA (n = 159), mean ± SD (range)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female sex, no. (%)</td>
<td>245 (71)</td>
<td>82 (52)</td>
</tr>
<tr>
<td>Age at inclusion in years</td>
<td>54 ± 12.9 (17, 85)</td>
<td>45 ± 12.8 (15, 77)</td>
</tr>
<tr>
<td>Duration of disease in years</td>
<td>12 ± 9.4 (0, 42)</td>
<td>11 ± 10.1 (0, 45)</td>
</tr>
<tr>
<td>ESR</td>
<td>35 ± 25.8 (2, 140)</td>
<td>22 ± 21.1 (1, 120)</td>
</tr>
<tr>
<td>CRP</td>
<td>29 ± 29.5 (2, 158)</td>
<td>18 ± 22.7 (0, 149)</td>
</tr>
<tr>
<td>DAS28: C-reactive protein</td>
<td>5.39 ± 1.18 (1, 9)</td>
<td>4.91 ± 1.0 (1, 7)</td>
</tr>
<tr>
<td>Patient Global Assessment (10 cm VAS)</td>
<td>6 ± 2.3 (0, 10)</td>
<td>5.5 ± 2.3 (0, 10)</td>
</tr>
<tr>
<td>Pain (10 cm VAS)</td>
<td>6 ± 2.3 (0, 10)</td>
<td>5 ± 2.3 (0, 10)</td>
</tr>
<tr>
<td>Tender joint count (range 0–28)</td>
<td>5 ± 2.3 (0, 28)</td>
<td>8 ± 6 (0, 28)</td>
</tr>
<tr>
<td>Fatigue (10 cm VAS)</td>
<td>6 ± 2.4 (0, 10)</td>
<td>6 ± 2.6 (0, 10)</td>
</tr>
<tr>
<td>Swollen joint count (range 0–28)</td>
<td>10 ± 6.6 (0, 25)</td>
<td>7 ± 6 (0, 28)</td>
</tr>
<tr>
<td>Tender joint count (range 0–66)</td>
<td>10 ± 6.0 (1, 10)</td>
<td>12 ± 9 (0, 43)</td>
</tr>
<tr>
<td>Concomitant methotrexate, (n, %)</td>
<td>220 (64)</td>
<td>56 (35)</td>
</tr>
<tr>
<td>Previous DMARDS, (n)</td>
<td>292</td>
<td>118</td>
</tr>
<tr>
<td>HAQ-DI (0–3)</td>
<td>1.3 ± 0.7 (0, 3)</td>
<td>0.96 ± 0.7 (0, 2.5)</td>
</tr>
<tr>
<td>SF-36 PCS (0–100)</td>
<td>50 ± 15.8 (12, 57)</td>
<td>34 ± 9.5 (13, 58)</td>
</tr>
<tr>
<td>SF-36 MCS (0–100)</td>
<td>45 ± 10.4 (17, 72)</td>
<td>46 ± 12.2 (20, 66)</td>
</tr>
<tr>
<td>SF-6D utility</td>
<td>0.54 ± 0.09 (0.3, 0.7)</td>
<td>0.57 ± 0.12 (0.25, 0.80)</td>
</tr>
<tr>
<td>EQ-5D UK TTO utility</td>
<td>0.43 ± 0.32 (–0.43, 1)</td>
<td>0.53 ± 0.32 (–0.24, 1)</td>
</tr>
<tr>
<td>Revised EQ-5D UK TTO utility</td>
<td>0.576 ± 0.22 (–0.14, 0.9954)</td>
<td>0.638 ± 0.19 (0.046, 0.9954)</td>
</tr>
</tbody>
</table>

CRP, C-reactive protein; DAS28, Disease Activity Score (28 joint); DMARDS, disease-modifying antirheumatic drugs; ESR, erythrocyte sedimentation rate; HAQ, Health Assessment Questionnaire; HAQ-DI, Health Assessment Questionnaire Disease Index; MCS, mental component summary; PCS, physical component summary; PsA, psoriatic arthritis; RA, rheumatoid arthritis; VAS, visual analogue scale.

* Unless otherwise indicated.
† Missing data (n = 18 patients).
‡ Missing data (n = 4).
was populated using Irish cost data. Our objective was to examine the change in the ICER when the utility estimates are changed.

**Results**

**Patient demographics**

At baseline, the mean age at inclusion was 54 years for the RA cohort and 45 years for the PsA group (Table 1). The average disease duration was similar (RA = 12 years, PsA = 11 years). The mean DAS28 was significantly higher in the RA group (5.39 [95% CI 5.16–5.43]) than in the PsA group (4.91 [95% CI 4.65–5.05]), as was the mean HAQ-DI (RA: 1.3 [95% CI 1.26–1.46] vs. PsA: 0.96 [95% CI 0.81–1.08]).

The mean utility scores and SDs are provided for each of the three methods (Table 2). In the RA group, the overall change was less for the revised EQ-5D scoring than the original EQ-5D (TTO) but greater than the SF-6D (Table 2). The change was greater in the PsA group across all three methods, and a trend between the scoring methods was seen similar to that in the RA group; the greatest change was produced when using the EQ-5D, less so with the revised method, and considerably less so with the SF-6D.

To describe the relationship between the measures, we fitted regression lines between each of the QOL instruments and the HAQ-DI. The equations for the mapping, including the coefficients of the regression, are presented in Appendix 2 found at doi:10.1016/j.val.2011.03.002. We calculated both a linear equation and a quadratic equation for each of the methods and HAQ-DI. Because there was no statistical difference between the models, we used the more parsimonious linear model. We plotted the regression lines for the measures to investigate the relationship between them (Fig. 1).

The revised scoring for the EQ-5D lessens the gap between the SF-6D and the original EQ-5D (Fig. 1). The slope produced by the relationship between the HAQ-DI and the revised scoring is less steep than that produced by the HAQ-DI with the original scoring.

The distribution of the utility score produced by the two methods of scoring for the EQ-5D differs. The marginal distribution of
Impact on ICER

The ICER for a biological agent is presented for each of the utility measures. The original EQ-5D mapping produced the lowest ICER, €42,402 (95% CI €36,837–€52,061). The SF-6D mapping produced the highest ICER, €111,788, and lies outside the acceptable willingness to pay range for most decision makers. The 95% CI was €105,154 to €141,665. The EQ-5D mapping using the revised scoring method fell between these two measures, €57,747, with a 95% CI of €52,032 to €72,845.

Discussion

There is a substantial burden of evidence highlighting the problems associated with the EQ-5D; this evidence is primarily referring not to the instrument itself but to the preference-based scoring method that was used to assign population-weighted values to the raw TTO scores [28,39]. The area of most concern with the original scoring method is how the WTD states were handled. Recent articles propose an alternative method to handle WTD states, but, to date it has not been demonstrated how this method could in practice alter the utility estimates and ultimately the results of an economic model [26,36].

In this article, we present the utility estimates from a large cohort of rheumatology patients scored using this alternative scoring method for the EQ-5D (TTO). In doing so, we provide a practical application of this revised method in an observational cohort of patients and demonstrate that the mean utility estimates produced by each of these methods differ considerably, which in turn influence the estimates of the economic model (Table 2).

The revised method used here handles the raw scores differently, and as a result, the distribution of scores observed when using the revised EQ-5D scoring method is narrower than the original method [26]. The lowest score in this cohort with the revised method is −0.143 and with the original method is −0.43.

The range of ICERs estimated using three different methods of utility measurement highlights the impact that utility has on the overall result in the case presented here. Although a pragmatic approach may be to recommend that one utility measure (either directly measured using questionnaires or via mapping) is used for all economic evaluations, this may restrict our ability to explore uncertainty associated with this parameter. Probabilistic sensitivity analysis explores uncertainty within the limits of the instrument measured. To examine the heterogeneity between utility measures, it may be useful to refit a cost-effectiveness model using multiple metrics and produce a range of ICER estimates.
Conclusion and recommendations

This article presents the results from a large cohort, using an alternative method for scoring the EQ-5D, and examines the relationship between both the revised and original generic measures (EQ-5D and SF-6D) and disease measures (HAQ-DI and DAS28) in inflammatory arthritis.

In the context of inflammatory arthritis, this article demonstrates that choice of scoring method of TTO utility measure may have a significant effect on the ICER, which may therefore have an impact on the reimbursement decision. Furthermore, in choosing just one QOL measure to produce a single ICER estimate, we may be restricting our ability to fully explore the uncertainty within the final estimate of a cost-effectiveness analysis.

Acknowledgments

The authors thank the reviewers for their helpful comments. One of the authors of this paper, Prof. Barry Bresnihan has moved on to a different place. BB, as he was known to many of us who were privileged to work with him had an ability to infuse enthusiasm for learning into all and especially to those tentatively venturing into the world of research. A true scholar who is greatly missed.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at doi:10.1016/j.val.2011.03.002, or if hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

value in health 14 (2011) 921–927


