VALUE IN HEALTH ▮ (2014) ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Comparison of Contemporaneous EQ-5D and SF-6D Responses Using Scoring Algorithms Derived from Similar Valuation Exercises

David G.T. Whitehurst, PhD, MSc, LLB[1,2,*], *Richard Norman, PhD, MSc, BA*[3], *John E. Brazier, PhD, MSc, BA*[4], *Rosalie Viney, PhD, MEc, BEc*[3]

[1]*Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada;* [2]*Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute, Vancouver, BC, Canada;* [3]*Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, Australia;* [4]*Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Sheffield, UK*

ABSTRACT

Objectives: Poor agreement between preference-based health-related quality-of-life instruments has been widely reported across patient and community-based samples. This study compares index scores generated from contemporaneous EQ-5D (3-level version) and SF-6D (SF-36 version) responses using scoring algorithms derived from independently-conducted Australian population-representative discrete choice experiments (DCEs), providing the first comparative analysis of health state valuations using the same method of valuation across the full value sets. **Methods:** EQ-5D and SF-6D responses from seven patient data sets were transformed into health state valuations using published DCE-derived scoring algorithms. The empirical comparative evaluation consisted of graphical illustration of the location and spread of index scores, reporting of basic descriptive statistics, exploration of between-measure differences in mean index scores, and analysis of agreement. **Results:** Compared with previously published findings regarding the comparability of "conventional" EQ-5D and SF-6D index scores, health state valuations from the DCE-derived scoring procedures showed that agreement between scores remained "fair" (intraclass correlation coefficient values across the seven data sets ranged from 0.375 to 0.615). Mean SF-6D scores were significantly lower than the respective mean EQ-5D score across all patient groups (mean difference for the whole sample = 0.253). **Conclusions:** The magnitude of disagreement previously reported between EQ-5D and SF-6D index scores is not ameliorated through the application of DCE-derived value sets; sizeable discrepancies remain. These findings suggest that differences between EQ-5D and SF-6D index scores persist because of their respective descriptive systems. Further research is required to explore the implications of variations in the descriptive systems of preference-based instruments.

Keywords: comparative analysis, discrete choice experiment, EQ-5D, SF-6D, standard gamble, time trade-off.

Introduction

Preference-based measurement of health-related quality of life (HRQOL) has become an important area of research over recent years, due largely to the increasing role of cost-utility analysis and the quality-adjusted life-year (QALY) metric within reimbursement mechanisms across publicly funded health care systems [1–6]. Although preferences can be elicited as a means to directly reflect an individual's own valuation of his or her health state—using techniques such as standard gamble (SG) and time trade-off (TTO)—the current economic evaluation framework that is practiced in many jurisdictions uses "off-the-shelf" questionnaires to capture health state valuations based on public preferences. While there are advocates for the use of individual (e.g., patient) preferences, the justification for using public preferences is based on theoretical, normative, and pragmatic arguments for incorporating general population values into health care decision making [7–10].

Preference-based HRQOL measures are made up of two components—a descriptive system and a valuation system—and there are considerable variations within these components across different questionnaires. The descriptive system defines respondents' HRQOL as one of a finite number of health states; the dimensions and associated response options that permit respondents to describe their current health state are fixed. The valuation component is a procedure for scoring each health state defined by the questionnaire. This procedure, often referred to as a scoring algorithm, is typically based on community-derived preferences and provides a single index score. Irrespective of the measure used to generate the index score, the value is interpreted on a scale on which one indicates full health and zero represents a health state considered equivalent to being dead.

* *Address correspondence to:* David G.T. Whitehurst, Faculty of Health Sciences, Simon Fraser University, Blusson Hall, Room 10504, 8888 University Drive, Burnaby, BC, Canada V5A 1S6.

E-mail: david_whitehurst@sfu.ca.

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2014.03.1720

The purpose of the single index score is to represent the relative value that society places on living in different health states. Negative index scores can be generated, which represent health states considered to be worse than death.

One of the major advantages of using standardized outcome measures, in any area of health research, is that they provide a common yardstick for interpreting results across studies. Currently, a number of different preference-based HRQOL measures are used to estimate patient benefit within economic evaluations, such as 15D [11], Assessment of Quality of Life (AQoL-4D, AQoL-6D, and the AQoL-8D) [12,13], EQ-5D (three-level and five-level versions) [14,15], Health Utilities Index (Mark 2 and Mark 3) [16], Quality of Well Being Self-Administered scale [17], and SF-6D (SF-36 and SF-12 versions) [18,19]. There has also been a substantial growth in the number of condition-specific preference-based measures [20–22]. The availability of multiple instruments that purport to measure the same underlying construct raises a concern about the cross-study comparability of cost-effectiveness evidence if different measures have been used to generate QALYs [23–26].

Comparative evaluations of preference-based HRQOL measures have addressed many different combinations of instruments; moderate agreement, at best, has been reported between instruments, across numerous patient and community-based samples [27–33]. The EQ-5D and SF-6D are the two most frequently compared measures due to having scoring algorithms derived from nationally representative, community-based samples from the same region—the United Kingdom [25]. Using these UK-derived scoring procedures, the EQ-5D has a considerably larger scoring range (−0.594 to 1.000) than the SF-6D (0.301 to 1.000 for the SF-6D [SF-36]; 0.345 to 1.000 for the SF-6D [SF-12]) (further details of the EQ-5D and SF-6D are provided in the Methods section). As a mere reflection of the range of possible scores, individuals in more severe health states tend to report lower EQ-5D scores than SF-6D scores. This led to concerns that the SF-6D suffers from floor effects [28]. Conversely, evidence of ceiling effects has been widely reported for the EQ-5D [28,32,33]. What is unclear is whether differences in index scores are a result of variation in the respective descriptive systems or methods of valuation, or both. Previous attempts to look at this suggest that differences in the valuation technique may account for some of the differences [34,35]. Furthermore, a comparative evaluation of multiple generic instruments using item response theory demonstrated that a simple linear function will transform one index score into the other for poorer health states [36]. Authors of the item response theory study suggest that the "which instrument is better?" question (for poorer health states only) is not one to be answered using psychometric criteria; rather, it is an issue concerning which utility scale correctly represents the values that the analysis wishes to reflect. What would help to better understand EQ-5D and SF-6D differences would be to compare index scores that have been valued using the same methods.

The commonly applied UK-specific EQ-5D and SF-6D health state valuations were generated from TTO and SG exercises, respectively [18,37]. These conventional elicitation techniques are grounded in utility theory but impose strong assumptions about the form of the utility function, suggesting that the resultant index scores reflect preferences for health states under a relatively narrow set of restrictions [38,39]. Concerns have also been raised about the cognitive difficulty of completing TTO and SG tasks for certain populations [40]. The use of discrete choice experiments (DCEs) to obtain health state valuations has been proposed, offering an alternative preference elicitation technique that has been claimed to allow for the investigation of more flexible model specifications and simpler valuation exercises for respondents [40–42].

The aim of this study was to assess the comparability of EQ-5D and SF-6D index scores derived from scoring algorithms constructed from two recent Australian population-representative DCEs [43,44]. In addition to reporting descriptive statistics for new scoring algorithms across a range of patient samples, the analysis provides the first comparative analysis of DCE-derived health state valuations for preference-based HRQOL instruments.

Methods

A summary of the EQ-5D (three-level version) and SF-6D (SF-36 version) descriptive systems is provided in Table 1, along with details of the valuation studies and associated scoring ranges for the DCE-derived algorithms used in the analysis. The same information is reported for the conventional EQ-5D (TTO) and SF-6D (SG) algorithms to provide further context. Details of the full descriptive systems of the two measures (dimensions and levels), illustrating the differences in wording, are provided in Appendix 1 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2014.03.1720.

Data Set

The comparative assessment reported in this study uses the same data set that compared TTO-derived EQ-5D index scores and SG-derived SF-6D index scores across seven patient groups, reported by Brazier et al. [28] in 2004. The seven patient groups are low back pain, chronic obstructive pulmonary disease, irritable bowel syndrome, leg ulcer, menopausal women, knee osteoarthritis, and a group of healthy older women (75+ years). Further details of the patient samples are reported elsewhere [28].

DCE-Derived Index Scores

The algorithms used to value EQ-5D and SF-6D health state descriptions were derived from two independently conducted, Australian population-representative DCEs. Brief details of these DCE studies and resultant algorithms are provided in Appendix 2 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2014.03.1720; extensive details are reported elsewhere [43,44]. Within the DCE valuation studies, a number of model specifications were estimated, investigating, among other issues, the inclusion of higher-order interactions, nonlinear preferences with respect to time, and the forcing of the algorithm to reflect any intended monotonicity within the instrument. To address the relevant policy consideration regarding the comparability of index scores generated from the DCE-derived algorithms, the primary comparative analysis reported in the current study focuses on the recommended model specifications for the EQ-5D and SF-6D algorithms; namely, the "Two-Factor Interaction" algorithm for the EQ-5D and the "Main Effects" algorithm for the SF-6D [43,44].

Statistical Analysis and Sensitivity Analysis

The empirical comparative evaluation of DCE-derived index scores consisted of: 1) graphical illustration of the location and spread of index scores (scatter graph and box-plot); 2) reporting basic descriptive statistics by instrument and by patient group; 3) exploring between-measure differences in mean index scores (floor and ceiling variation, and paired t tests); and 4) analysis of agreement.

Within this study, consideration of floor and ceiling effects focuses on the respective index scores only. The distribution of EQ-5D and SF-6D responses across dimensions has been reported previously and, obviously, these responses have not changed; it is