Generic Preference-based Measures for Low Back Pain

Which of Them Should Be Used?

Aureliano Paolo Finch, MSc,* Melina Dritsaki, PhD,† and Claudio Jommi, MSc*†

Study Design. Systematic review.

Objective. This systematic review examines validity and responsiveness of three generic preference-based measures in patients with low back pain (LBP).

Summary of Background Data. LBP is a very common incapacitating disease with a significant impact on health-related quality of life (HRQoL). Health state utility values can be derived from various preference-based HRQoL instruments, and among them the most widely ones are EuroQol 5 dimensions (EQ-5D), Short Form 6 Dimensions (SF-6D), and Health Utilities Index 3 (HUI III). The ability of these instruments to reflect HRQoL has been tested in various contexts, but never for LBP populations.

Methods. A systematic search on electronic literature databases was undertaken to identify studies of patients with LBP where health state utility values were reported. Records were screened using a set of predefined eligibility criteria. Data on validity (correlations and known group methods) and responsiveness (effect sizes, standardized response means, tests of statistical significance) of instruments were extracted using a customized extraction template, and assessed using predefined criteria.

Results. There were substantial variations in the 37 included papers identified in relation to study design and outcome measures used. EQ-5D demonstrated good convergent validity, as it was able to distinguish between known groups. EQ-5D was also able to capture changes of health states as results of different interventions. Evidence for SF-6D and HUI III was limited to allow an appropriate evaluation.

Conclusion. EQ-5D performs well in LBP population and its scores seem to be suitable for economic evaluation of LBP interventions. However, the paucity of information on the other instruments makes it impossible to determine its relative validity and responsiveness compared with them.

Key words: EQ-5D, health economics, health policy, HUI III, low back pain, preference-based measures, psychometric characteristics, responsiveness, SF-6D, validity.

Level of Evidence: 2
Spine 2016;41:E364–E374

Low back pain (LBP) is a common health problem. A review of studies published between 1966 and 1998 reported that LBP lifetime prevalence reaches an 84% peak, whereas point prevalence and 1-year prevalence ranges from 12% to 33% and from 22% to 65%, respectively.

As an incapacitating disease LBP has an important impact on health-related quality of life (HRQoL), making cost-utility analysis (CUA) the preferred economic evaluation for LBP interventions. In CUA, life years gained are weighted for health state utility values (HSUVs), which are commonly derived from three generic preference-based measures: EuroQol Five Dimensions (EQ-5D), Short Form Six Dimensions (SF-6D), and Health Utility Index Three (HUI III). Preference-based HRQoL instruments typically comprise a descriptive system covering core dimensions of health (e.g., mobility, self-care, usual activities, pain, and anxieties) and an attached value set, which is obtained on the basis of population’s relative desire for dimensions of health. These generic measures are claimed to be applicable across all disease areas, therefore representing an important clinical outcome as well as a common currency for health technology assessment.

These instruments psychometric performance in terms of validity (i.e., reaching the objectives it has been developed for) and responsiveness (i.e., ability to detect changes over time and across participants) has been already tested in...
different decision contexts, and more precisely in patients with visual disorders, cardiovascular diseases, cancer, rheumatoid arthritis, musculoskeletal diseases, and multiple sclerosis, but not in LBP populations. This systematic review aims at covering this gap and establishing whether these instruments use is appropriate in LBP populations. As it is common in similar studies, included articles will not be required to have conducted an assessment of validity and responsiveness themselves, but will contain information from which the instruments performance can be analyzed.

MATERIALS AND METHODS
The study design is a systematic literature review.

Literature Search
Medline, Embase, and Web of Science were investigated using a strategy developed around the four main constructs of the research question: EQ-5D, SF-6D, HUI-III, LBP. Terms searched were derived from Brazier et al,10 Brooks,11 Dolan,12 Fenny et al,13 Hayden et al,14 and Lin et al.15 The searching strategy included synonyms and spelling variations and was refined using truncation, wildcards, phrase search and proximity operators, and adjusted for differences in databases. Related terms such as “validity” or “psychometric characteristics” were not used because of this systematic review objective (this would have been useful in a systematic review of studies assessing the validity of preference-based instruments). No publication date limit was set. All studies published in English or for which a translator was available were considered. As an example, the complete search strategy for MEDLINE (Ovid) is provided in Appendix I, http://links.lww.com/BRS/B57.

Study Selection
Relevant records were imported on Refworks and duplicates were removed. Studies were included in the systematic review if they met all the eligibility criteria presented in Table 1.

Data Extraction
A customized extraction template model was used for the collection of relevant data, including study characteristics (e.g., study design), patients characteristics (e.g., age), type and method of validity assessment (e.g., convergent, correlations), method of responsiveness assessment (e.g., standardize response mean), validity and responsiveness data.

Quality Assessment
Quality was assessed using the COmnesus-based Standards for the selection of health Measurement INstrument (COSMIN) checklist,16 a rating tool to evaluate the methodological quality of studies on measurement properties of health status instruments. For the four psychometric characteristics relevant to the current systematic review (“measurement error,” “hypothesis testing,” “cross cultural validity,” and “responsiveness”) 11 to 18 items per characteristic were analyzed. Each item was assigned one of the four possible scores: “excellent,” “good,” “fair,” or “poor.” The item with the lowest score determined the overall score for the property under investigation.

Assessment of Validity and Responsiveness
Construct validity has been defined as the extent to which an instrument measures what it is intended to measure. Construct validity was analyzed when papers reported on convergent validity (correlation between instruments) and known groups differences detected by instruments.

Responsiveness has been defined as the extent to which an instrument is sensitive to statistically significant changes in health over time or between treatment arms. Responsiveness was analyzed when papers reported on tests of statistical significance (TSS), effect sizes (ES), and/or standardize response mean (SRM).

Instruments validity and responsiveness was assessed against a set of hypotheses derived from the literature,22 (Table 2).

Monotonic correlations were considered very weak between 0 and 0.19; weak between 0.2 and 0.39; moderate between 0.4 and 0.59; strong between 0.6 and 0.79; and very strong between 0.8 and 1. Changes in SRM and ES were considered very weak between 0 and 0.19; weak between 0.2 and 0.49; moderate between 0.5 and 0.79; and strong between 0.8 and 1.

RESULTS
Characteristics of the Included Studies
A total of 739 potentially relevant articles were found. Title and abstract screening excluded 223 and 432 records respectively. After reviewing the articles full text, 37 reports referred to 35 studies were included. The process is described in Figure 1.

The design feature of included studies varied significantly. The majority of them were randomized controlled trials (RCT),25–41 followed by cross-sectional studies,42–47 observational longitudinal48–58 and cohort studies.59–61

Quality of the Included Papers
Quality scores for the three mostly investigated psychometric characteristics (“measurement error,” “hypothesis testing,” “cross cultural validity,” and “responsiveness”) 11 to 18 items per characteristic were analyzed. Each item was assigned one of the four possible scores: “excellent,” “good,” “fair,” or “poor.” The item with the lowest score determined the overall score for the property under investigation.

TABLE 1. Eligibility Criteria

| Inclusion criteria                                                                 | Exclusion criteria                                                                 |
|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| The study population had LBP                                                        | The study focused on a condition other than LBP                                    |
| The study examined at least one of the three general preference based instrument (EQ5D, SF6D, HUI3) | The study examined LBP with comorbidities                                           |
| The study reported an estimate of mean score for the preference based instrument/s examined and for a comparator (e.g., disease specific) | Pharmacodynamic and pharmacokinetic studies                                       |
|                                                                                   | Presentation at conferences and poster presentation                                 |


TABLE 2. Validity and Responsiveness Hypothesis

| Convergent validity |  |
|---------------------|--|
| Hypothesis 1. A positive and moderate-to-very strong correlation (r > 0.4) between generic instruments and disease-specific instruments for those disease-specific instruments measuring improvements through a reduction in the scores a negative correlation is expected. |  |
| Hypothesis 2. A positive and strong-to-very strong correlation between generic instruments (r > 0.6) |  |
| Hypothesis 3. Stronger correlations between generic preference-based instruments and disease-specific instruments than generic preference-based instruments and disease construct-specific instruments |  |

| Known groups |  |
|----------------|---|
| Hypothesis 1. Generic instruments to distinguish between different grades of disability (lower scores at increasing level of disability) |  |
| Hypothesis 2. Generic instruments to distinguish between groups with disability and groups without disability (lower scores in the presence of disability) |  |
| Hypothesis 3. Generic instruments to distinguish between men and women (lower scores for women than for men) |  |
| Hypothesis 4. Generic instruments to distinguish between acute and recurrent LBP (lower scores for acute cases) |  |

| Test of statistical significance |  |
|----------------|---|
| Hypothesis 1. Generic instruments to be able to detect changes because of treatments |  |
| Hypothesis 2. Generic instruments to be able to detect differences between interventions |  |
| Hypothesis 3. Generic instruments to be able to detect changes coherent with those reported by other generic or disease-specific measures |  |

| Standardized response mean and effect sizes |  |
|----------------|---|
| Hypothesis 1. SRM and ES to be moderate to strong (r > 0.5) |  |

Five of them analyzed EQ-5D and ODI correlations and results were generally moderate to strong (in absolute terms). Correlation coefficients were between 0.510 and 0.739 in three studies, 0.48 in one, and between 0.232 and 0.206 in one. In one study data were too sparse to assess correlations. Rather strangely, the direction of the correlation changed across studies.

Three of them assessed convergent validity between EQ-5D and RMDQ. Correlations were moderate to strong (p < 0.02) in all of them.

EQ-5D was also found to moderately correlate with ABPS ($r = -0.44$) in one study and with Specific Sexual Function Questions ($r = -0.51$) and Core Outcome Measure Index (COMI) ($r = -0.54$) in two others.

One study presented results for both EQ-5D and HUI III correlations with ODI and found moderate correlations at 3 and 6 months for both instruments. Correlations between HUI III and ODI were stronger than those between EQ-5D and ODI at 3 months but weaker at 6 months.

Overall, given that only one study did not reflect our prior expectations of moderate-to-very strong correlations, findings support the first hypothesis of convergent validity for the EQ-5D, and the small evidence found sustains the first hypothesis for the HUI III.

**Hypothesis 2**

EQ-5D correlation with other HRQoL instruments was assessed in five studies. EQ-5D and visual analogue scale (VAS) agreement was examined by three studies. Burnstrom et al reported strong correlations between the two instruments ($r = 0.67$). Similarly, in the two papers of Kovacs et al., correlations between EQ-5D and VAS were strong at both 15 and 60 days follow-up. More precisely, the correlation coefficients...
were 0.70 at 15 days investigation point and 0.76 at 60 days investigation time points and 0.67 at 15 days investigation point in all of the cases. Differently from what we expected, correlations were only moderate at baseline ($r = 0.52$ and $r = 0.42$).

In one study, EQ-5D and SF-6D correlation was moderate ($r = 0.553$). Similarly, one study found moderate correlations between EQ-5D and SF36 ($r = 0.49$).

Although some papers present data that sustain our prior expectations of positive and strong-to-very strong correlations for the second hypothesis, results are not conclusive given that moderate correlations were also frequently reported.

**Hypothesis 3**

Only one study presented results for correlation between a generic and a disease construct-specific instrument. In detail, Genevay et al\(^{51}\) found that EQ-5D was weakly associated with COMI symptom specific ($r = -0.36$).

This study supports the third hypothesis of convergent validity of weak correlations between generic preference-based instruments and disease construct-specific instruments.

**Known Group Method**

Ten studies allowed an assessment of known groups for EQ-5D.\(^{45,46,49,52,53,56,58-61}\)

**Hypothesis 1**

Five studies (six reports) permitted an assessment of EQ-5D validity after the first hypothesis.\(^{45,46,49,56,58,61}\)

Two reports referred to the same study\(^{45,46}\) showed that EQ-5D was able to detect variations in groups with different severity grades of lumbar spondylitis. Differences were statistically significant. One study\(^{49}\) showed that EQ-5D is able to distinguish between women with non-lumbar-pelvic pain, women with lumbar pain, women with pregnancy-related pelvic girdle pain, and women with combined pain. Differences between groups were statistically significant between women without lumbar-pelvic pain and all the other groups, and between women with lumbar pain and women with combined pain. Differences between lumbar pain and pregnancy-related pelvic girdle pain were not statistically significant. One study\(^{38}\) reported EQ-5D to differentiate between the group of patients for which the treatment was successful and the group of patients who did not respond to it ($P = 0.003$). Parker et al\(^{56}\) presented similar results between patients categorized according to three severity grades: stable; worst and best clinical situation ($P \leq 0.005$). EQ-5D presented the highest values for the best clinical situation and the lowest values for the worst situation. Van der Roer et al\(^{61}\) reported similar results for the same severity groups, although it did not provide results for statistical significance.

Overall, EQ-5D responds well when tested on different severity known groups distinguishing between different grades of disability and therefore sustaining the first hypothesis for known group methods.

**Hypothesis 2**

Only one study permitted an assessment of the second hypothesis for known groups.\(^{45,46}\)

The two reports of Muraki et al\(^{45,46}\) registered a higher mean score ($P < 0.05$) for those patients who declared not to have LBP if compared with those with the symptom.

This sustains the second hypothesis of known group methods, which is the ability of generic preference measure to distinguish between patients and general population.

**Hypothesis 3**

The third hypothesis of known group method has been tested in four studies.\(^{42-44,60}\)

All of them reported women to have significantly lower EQ-5D utility scores than men\(^{42-44,60}\) maintaining constant...
the clinical condition, and this was always statistically significant.

Results support the third hypothesis of known groups assessment (distinguishing between male and female).

**Hypothesis 4**

Only one study permitted to evaluate the fourth hypothesis of known group. This study showed EQ-5D to perform well in differentiating patients with acute or recurrent LBP, presenting higher pain and dysfunction for the acute group.

This confirms the fourth hypothesis of the study, namely the ability of distinguishing between acute and recurrent LBP.

| Table 3: Quality of the Included Papers |
|---------------------------------------|
| Name of the First Author and Year of Publication | Measurement Error | Hypothesis Testing | Cross Cultural Validity | Responsiveness |
|-----------------------------------------------|-------------------|--------------------|------------------------|---------------|
| RCTs                                          |                   |                    |                        |               |
| Bastiaenen et al, 2008                        | Good              | Poor               | n/r                    | Poor          |
| Berg et al, 2009                             | Fair              | Poor               | n/r                    | Fair          |
| Berg et al, 2009                             | Fair              | Poor               | n/r                    | Fair          |
| Carr et al, 2005                             | Fair              | Poor               | n/r                    | Fair          |
| Casserly-Forney et al, 2012                  | Fair              | Poor               | n/r                    | Fair          |
| Choong et al, 2008                           | Fair              | Poor               | n/r                    | Fair          |
| Cox et al, 2010                              | Fair              | Poor               | n/r                    | Fair          |
| Del Pozo-Cruz et al, 2011                    | Good              | Good               | n/r                    | Good          |
| Djas et al, 2005                             | Fair              | Poor               | n/r                    | Fair          |
| Gilbert et al, 2004                          | Good              | Good               | n/r                    | Good          |
| Gilbert et al, 2004                          | Fair              | Good               | n/r                    | Fair          |
| Hellum et al, 2011                           | Good              | Excellent          | n/r                    | Excellent     |
| Hurley et al, 2001                           | Fair              | Fair               | n/r                    | Fair          |
| Kendrick et al, 2001                         | Fair              | Fair               | n/r                    | Fair          |
| Miller et al, 2002                           | Good              | Good               | n/r                    | Fair          |
| Rivero-Arrias et al, 2006                    | Excellent         | Fair               | n/r                    | Poor          |
| Wilkens et al, 2010                          | Good              | Good               | n/r                    | Good          |
| Cross-sectional                              |                   |                    |                        |               |
| Burstrom et al, 2001                         | Poor              | Good               | n/r                    | Poor          |
| Eker et al, 2007                             | Good              | Good               | n/r                    | Poor          |
| Klemenc-Ketis, 2011                          | Poor              | Fair               | n/r                    | Poor          |
| Muraki et al, 2011                           | Poor              | Fair               | n/r                    | Poor          |
| Muraki et al, 2012                           | Poor              | Fair               | n/r                    | Poor          |
| Sogaard et al, 2009                          | Poor              | Good               | n/r                    | Poor          |
| Observational longitudinal                  |                   |                    |                        |               |
| Aghayevo et al, 2010                         | Fair              | Poor               | n/r                    | Poor          |
| Cheshire et al, 2011                         | Fair              | Poor               | n/r                    | Poor          |
| Garratt et al, 2001                          | Fair              | Good               | n/r                    | Good          |
| Genevay et al, 2012                          | Good              | Good               | Good                   | Good          |
| Gutke et al, 2011                            | Good              | Good               | n/r                    | Good          |
| Klemenc-Ketis, 2011                          | Poor              | Fair               | Fair                   | Poor          |
| Kovacs et al, 2005                           | Good              | Good               | n/r                    | Good          |
| Kovacs et al, 2004                           | Good              | Good               | n/r                    | Good          |
| Parker et al, 2016                           | Good              | Good               | n/r                    | Good          |
| Schuelmann                                  | Fair              | Poor               | n/r                    | Poor          |
| Suarez-Almazor et al, 2000                   | Fair              | Fair               | n/r                    | Fair          |
| Cohort studies                              |                   |                    |                        |               |
| Gutke et al, 2006                            | Poor              | Good               | n/r                    | Poor          |
| Jannson et al, 2009                          | Poor              | Good               | n/r                    | Poor          |
| Van der Roer et al, 2006                     | Fair              | Poor               | n/r                    | Good          |

n/r indicates not relevant.

**Responsiveness**

Twenty-four studies allowed for an assessment of responsiveness. Twenty-one of them reported TSS, three of them ES, and one of them SRM. Twenty of the studies (19 reports) permitted an assessment of the first hypothesis of responsiveness.

**Test of Statistical Significance Method**

**Hypothesis 1**

Eighteen studies (19 reports) permitted an assessment of the first hypothesis of responsiveness. Hellum et al managed to detect statistically significant improvements in patients treated with surgery with disc prosthesis and patients treated with rehabilitation therapy.
Schluessmann et al\textsuperscript{57} presented significant changes in patients receiving total disc arthroplasty, with an EQ-5D mean score of 0.32 at baseline, and improvements to 0.72 at 3 months and 0.73 at 1 year. Parker et al\textsuperscript{56} registered significant improvement of EQ-5D after patients had undergone lumbar fusion, which were statistically significant. Also Berg et al\textsuperscript{56, 26} Chown et al\textsuperscript{56} Aghayev et al\textsuperscript{48} and Cheshire et al\textsuperscript{49} reported similar results, which were statistically significant.

In studies conducted by Bastiaenen et al\textsuperscript{25} Carr et al\textsuperscript{27} Casserley-Feeney et al\textsuperscript{49} Djas and Kalim,\textsuperscript{33} Gilbert et al\textsuperscript{34, 35} Hurley et al\textsuperscript{57} Jansson et al\textsuperscript{60} and Wilkens et al\textsuperscript{41} EQ-5D values appeared responsive to improvements because of the treatment of LBP, although these were not statistically significant.

According to Kovacs et al\textsuperscript{54} Rivero-Arrias et al\textsuperscript{40} and Van der Roer et al\textsuperscript{41} the EQ-5D is responsive to variations in the health status because of treatment.

TABLE 4. Main Outcome Measures Reported by the Included

| Author, Year | Descriptive System | Rating Scale | Other Instruments Used (Generic non Preference Based, Clinical, Condition specific) |
|--------------|--------------------|--------------|----------------------------------------------------------------------------------|
| Aghayev et al, 2010\textsuperscript{19} | ✔ | ✔ | ✔ |
| Bastiaenen et al, 2008\textsuperscript{24} | ✔ | ✔ | ✔ |
| Berg et al, 2009\textsuperscript{25} | ✔ | ✔ | ✔ |
| Berg et al, 2009\textsuperscript{26} | ✔ | ✔ | ✔ |
| Burstrom et al, 2001\textsuperscript{41} | ✔ | ✔ | ✔ |
| Carr et al, 2005\textsuperscript{27} | ✔ | ✔ | ✔ |
| Casserley-Feeney et al, 2012\textsuperscript{28} | ✔ | ✔ | ✔ |
| Cheshire et al, 2011\textsuperscript{30} | ✔ | ✔ | ✔ |
| Chown et al, 2008\textsuperscript{29} | ✔ | ✔ | ✔ |
| Cox et al, 2010\textsuperscript{30} | ✔ | ✔ | ✔ |
| Del Pozo-Cruz et al, 2011\textsuperscript{11} | ✔ | ✔ | ✔ |
| Djais et al, 2005\textsuperscript{22} | ✔ | ✔ | ✔ |
| Eker et al, 2007\textsuperscript{41} | ✔ | ✔ | ✔ |
| Garratt et al, 2001\textsuperscript{52} | ✔ | ✔ | ✔ |
| Genevay et al, 2012\textsuperscript{23} | ✔ | ✔ | ✔ |
| Gilbert et al, 2004\textsuperscript{34} | ✔ | ✔ | ✔ |
| Gilbert et al, 2004\textsuperscript{34} | ✔ | ✔ | ✔ |
| Gutke et al, 2011\textsuperscript{32} | ✔ | ✔ | ✔ |
| Gutke et al, 2006\textsuperscript{43} | ✔ | ✔ | ✔ |
| Hellum et al, 2011\textsuperscript{13} | ✔ | ✔ | ✔ |
| Hurley et al, 2001\textsuperscript{36} | ✔ | ✔ | ✔ |
| Jansson et al, 2009\textsuperscript{16} | ✔ | ✔ | ✔ |
| Kendrick et al, 2001\textsuperscript{17} | ✔ | ✔ | ✔ |
| Klemenc-Ketis, 2011\textsuperscript{42} | ✔ | ✔ | ✔ |
| Klemenc-Ketis, 2011\textsuperscript{57} | ✔ | ✔ | ✔ |
| Kovacs et al, 2005\textsuperscript{38} | ✔ | ✔ | ✔ |
| Kovacs et al, 2004\textsuperscript{29} | ✔ | ✔ | ✔ |
| Miller et al, 2002\textsuperscript{38} | ✔ | ✔ | ✔ |
| Muraki et al, 2011\textsuperscript{43} | ✔ | ✔ | ✔ |
| Muraki et al, 2010\textsuperscript{44} | ✔ | ✔ | ✔ |
| Parker et al, 2012\textsuperscript{45} | ✔ | ✔ | ✔ |
| Rivero-Arrias et al, 2006\textsuperscript{39} | ✔ | ✔ | ✔ |
| Schluessman et al, 2009\textsuperscript{46} | ✔ | ✔ | ✔ |
| Sogaard et al, 2009\textsuperscript{49} | ✔ | ✔ | ✔ |
| Suarez-Almazor et al, 2000\textsuperscript{48} | ✔ | ✔ | ✔ |
| Van der Roer et al, 2006\textsuperscript{60} | ✔ | ✔ | ✔ |
| Wilkens et al, 2010\textsuperscript{40} | ✔ | ✔ | ✔ |

ABPS indicates Aberdeen Back Pain Scale; EQ-5D, EuroQol 5 Dimensions; HUI 3, Health Utility Index Mark 3; NASS, Lumbar Spine Outcome Assessment Instrument; ODI, Oswestry Disability Index; RDQ, Roland Morris Disability Questionnaire; SF-12, Short Form 12 Dimensions; SF-36, Short Form 36 Dimensions; SF-6D, Short Form 6 Dimensions; VAS, Visual Analogue Scale.
Overall, the first hypothesis for TSS holds given that preference-based measures are able to detect changes because of treatment.

**Hypothesis 2**

Twelve studies permitted to test for the second hypothesis of responsiveness. 25,27–31,33–35,41,48,60

In Chown et al 20 all patients assigned to the exercise, physical therapy, or osteopathy groups improved, but patients in the osteopathy group reported significantly higher EQ-5D values if compared with patients in the group exercise (P < 0.01). Similarly, Berg et al 27 registered a different increase in mean EQ-5D values from baseline to 1 year for patients assigned to the total disc replacement group compared with patients assigned to the fusion group, with the total disc replacement being more effective (P < 0.05). Aghayev et al 48 found that EQ-5D was able to distinguish between patients receiving Dynardi total disc arthroplasty and patients receiving total disc replacement, with the differences between the two groups being statistically significant at P < 0.001. Gilbert et al 34,35 found that EQ-5D differentiated between magnetic resonance imaging and delayed magnetic resonance imaging at 8 and 24 months, and that differences were statistically significant in this latter follow-up.

Other seven studies presented data that supported the second hypothesis, although results were not statistically significant. 25,28,29,31,33,36,41 Carr et al, 25 for instance, registered an increase in EQ-5D mean values from baseline to 3 months of 0.028 and from baseline to 12 months of 0.045 for the individual physiotherapy group, whereas improvements for the group exercise were milder. Similarly, Casserley-Feeney et al 30 reported EQ-5D to differ between public physiotherapy and private physiotherapy patients, Djas and Kalim 13 for the instrument to be sensitive to differences between patients undergoing radiography and patients not undergoing radiography and Wilkens et al 34 for the measurement to recognize patients administered with glucosaming and patients administered with placebo. Bastiaenen et al 25, 34,35 Hellum et al 16 and Cox et al 29 reported similar results.

One study 60 managed to differentiate between patients treated with macrodecompression, microscopic decompression, decompression and fusion, and fusion alone.

These results confirm the ability of the EQ-5D to distinguish between different interventions outcomes.

**Hypothesis 3**

Fifteen studies (16 reports) permitted an assessment of the third hypothesis of responsiveness. 25,27–30,33–39,41,54,56,61

Twelve of them reported an EQ-5D behavior that was coherent with the scores registered by other measures. 27–30,33–39,41,54,56 For example, Berg et al 27 registered an increase in EQ-5D values for the total disc replacement group at 1 year, and a reduction of the mean value at 2 years, and similar trends were reported for ODI and VAS. Also Parker et al 56 results of EQ-5D and ODI were coherent. Similarly, Carr et al 25 Chown et al, 30 Djas and Kalim, 33 Hurley et al 37 and Kovacs et al 54 presented improvements that were well detected by both EQ-5D and RMDQ, Vand er Roer et al 61 by EQ-5D and Quebec Pain Disability Scale and Gilbert 34,35 and Hellum et al 16 by EQ-5D and ABPS.

Although also for Casserley-Feeney et al 29 EQ-5D and RMDQ presented similar results, this latter study evidenced that RMDQ is more sensitive than EQ-5D to small differences at low levels of disability. This lack of sensitivity to change in health states seems confirmed also by other studies. For example, in Miller et al 39 RMDQ is able to detect a small change in patients’ status at 3 months that passed undetected by EQ-5D and in Bastiaenen et al 25 a similar problem occurs with EQ-5D and RMDQ at 6 months. In Kendrick et al, 38 median EQ-5D scores remained stable from baseline to 9 months, whereas RMDQ scores detected a small improvement in patients. Also, Wilkens et al 31 found an extremely small improvement registered by RMDQ at 1 year follow-up not registered by the EQ-5D.

Overall, the evidence collected supports the third hypothesis of responsiveness which is the ability of reporting changes coherent to those reported by other generic or diseases-specific measures.

**Effect Size and Standardize Response Mean**

**Hypothesis 1**

Three studies permitted to test ES 32,53,58 and one study SRM. 30

EQ-5D ES were moderate and statistically significant in two studies. 32,53 The third study 38 reported ES for both EQ-5D and HUI III, and found HUI III to be more discriminative than EQ-5D at 3 months, with effect sizes similar to ODI ones. At 6 months, both EQ-5D and HUI III were highly discriminative.

One study presented EQ-5D SRM and found a moderate responsiveness of the instrument. 30

ES and SRM were moderate to strong, therefore supporting the hypothesis of responsiveness.

EQ-5D validity and responsiveness results are summarized in Table 5.

**DISCUSSION**

The 35 studies (37 reports) included in this systematic review show that LBP decreases HRQoL and that EQ-5D is generally able to detect improvements and deteriorations in health states because of health interventions or disease progression.

Comparing our results with those of similar researches it emerges that EQ-5D performs well in LBP populations. In a review of Tosh et al 3 EQ-5D correlation with visual acuity, a disease-specific instrument for visual disorders, was often poor or nonsignificant for patients with age-related macular degeneration and cataracts. Similarly, a review of Papaioannou et al 62 found generally modest and mostly weak correlations between EQ-5D and disease-specific instruments such as brief psychiatry rating scale and quality-of-life scale.
In light of this, the commonly moderate-to-strong correlations between EQ-5D and disease-specific instruments found in our study show a good performance of the instrument. Differently from what was hypothesized, EQ-5D correlation with other generic instruments was strong at follow-ups, but only moderate at baseline. Weaker correlations for baseline data might be because of EQ-5D being more sensitive to the lower end of the utility scale, having more distributed frequencies among spine patients compared with other generic instruments (the effect of which is lower mean values for patients in worst health states), or EQ-5D measuring constructs that are relevant for greater disability levels than other generic instruments. Nevertheless, moderate correlations between general preference-based instruments have already been seen in other studies, thus this behavior cannot be considered proper evidence against the instrument validity.

EQ-5D known group assessment showed statistically significant differences between different disease severities. Table 5 shows the summary of results for EQ-5D.

### Table 5. EQ-5D Summary of Results

| Author, Year       | Convergent Validity | Validity—Known Groups | Responsiveness TSS | Responsiveness ES |
|--------------------|---------------------|------------------------|-------------------|------------------|
|                    | H1                  | H2                     | H3                | H1               | H2 | H3 | H4 | H1 | H2 | H3 | H4 | H1 | H2 | H3 | H4 |
| Aghayev et al, 2010 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Bastiaenen et al, 2008 | ✓                 | ✓                      | ✓                 | ✓               | ✓   | ✓  |
| Berg et al, 2009   | ✓                   | ✓                      | ✓                 | ✓               | ✓   | ✓  |
| Berg et al, 2009   | ✓                   | ✓                      | ✓                 | ✓               | ✓   | ✓  |
| Burstrom et al, 2001 | ✓                  | ✓                      | ✓                 | ±               | ±   | X  |
| Carr et al, 2005   | ✓                   | ✓                      | ✓                 | ±               | ±   | ✓  |
| Casserley-Feeney et al, 2012 | ✓          | ✓                      | ✓                 | ±               | ±   | ✓  |
| Cheshire et al, 2011 | ✓                 | ✓                      | ✓                 | ✓               | ✓   | ✓  |
| Chown et al, 2009  | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Cox et al, 2010    | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Del Pozo-Cruz et al, 2011 | ✓             | ✓                      | ✓                 | ±               | ±   | X  |
| Dijas et al, 2005  | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Eker et al, 2007   | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Garrett et al, 2001 | ✓                 | ✓                      | ✓                 | ±               | ±   | ✓  |
| Genevay et al, 2012 | ✓                 | ✓                      | ✓                 | ±               | ±   | X  |
| Gilbert et al, 2004 | ✓                 | ✓                      | ✓                 | ±               | ±   | X  |
| Gutke et al, 2011  | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Gutke et al, 2006  | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Hellum et al, 2011 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Hurley et al, 2001 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Jansson et al, 2009 | ✓                 | ✓                      | ✓                 | ±               | ±   | X  |
| Kendrick et al, 2001 | X                 | X                      | X                 | ±               | ±   | X  |
| Klemenc-Ketis, 2011 | ✓                 | ✓                      | ✓                 | ±               | ±   | X  |
| Klemenc-Ketis, 2011 | ✓                 | ✓                      | ✓                 | ±               | ±   | X  |
| Kovacs et al, 2005 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Kovacs et al, 2004 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Miller et al, 2002 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Muraki et al, 2011 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Muraki et al, 2010 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Parker et al, 2012 | ✓                   | ✓                      | ✓                 | ±               | ±   | X  |
| Rivero-Arrias et al, 2006 | X | X                      | X                 | ±               | ±   | X  |
| Schluessman et al, 2009 | X           | X                      | X                 | ±               | ±   | X  |
| Sogaard et al, 2009 | ✓                 | ✓                      | ✓                 | ±               | ±   | X  |
| Suarez-Almazor et al, 2006 | ✓ | ✓                      | ✓                 | ±               | ±   | X  |
| Van der Roer et al, 2006 | ✓            | ✓                      | ✓                 | ±               | ±   | X  |
| Wilkens et al, 2010 | ✓                 | ✓                      | ✓                 | ±               | ±   | X  |

**Keys:** ✓ Meeting prior expectations; ± trend meeting prior expectation but not statistically significant; - trend nonmeeting prior expectations; ? mixed/not possible to assess.

When two keys for the same item are used, it is because more than one result was found. ES indicates effect size: H1, hypothesis 1; H2, hypothesis 2; H3, hypothesis 3; H4, hypothesis 4. TSS, test of statistical significance.
patients with/without LBP and respondents/nonrespondents to treatments. There was also strong and statistically significant evidence that EQ-5D can distinguish between women and men perception of health, with the HRQoL values for the former being lower than the latter. These results sustain our prior hypothesis and are in line with those of other systematic reviews on EQ-5D validity in other population (e.g., Peasgood et al).

EQ-5D appears to be a responsive instrument, although it seems to be less responsive than disease-specific ones. This is not surprising. Disease-specific and general preference-based instruments are not perfect substitutes. Disease-specific instruments only contain items or health dimensions that are relevant for the specific condition examined, whereas generic instruments assess all domains of HRQoL. By contrast, general preference-based instruments are meant to be perfect substitutes, at least in theory. The current systematic review presents paucity of data as regards between generic instruments comparison. One study found HUI III to be more responsive than EQ-5D at 3 months and equally responsive at 6 months. Another study presented only moderate correlation between EQ-5D and SF-6D. These results seem to suggest that the three preference-based instruments are not equivalent measures of HRQoL and that they assess different domains. However, results cannot be considered conclusive and a study estimating direct correlations between generic instruments might be useful.

This systematic review has some limitations. First, some of the included studies present small sample sizes. This might be one of the reasons for the lack of statistical significance registered in some reports. Second, there is not enough reference to missing data caused by nonrespondents and how these have been accounted for. Finally, some of the included studies did not control for age, sex, social status, and other variables that can influence LBP evaluation.

Nevertheless, our systematic review represents an important effort. It suggests that EQ-5D performs well in LBP population and that its scores are suitable for economic evaluation of LBP interventions, whereas it recommends the use of EQ-5D in combination with disease-specific instruments for clinical evaluation, given its lack of sensitivity to change in health state compared with them. Results for SF-6D and HUI III are too scarce to draw any conclusion.

Acknowledgments

The authors would like to thank and extend their heartfelt gratitude to the following persons who have made the completion of this work possible: Matthew Glover, Louise Longworth, Joanne Lord, Patrizio Armeni and Francesco Costa. Special thanks go to Yaling Yang, who acted as an advisor to the study and provided helpful comments on the review and manuscript.

Supplemental digital content is available for this article.

Direct URL citations appearing in the printed text are provided in the HTML and PDF version of this article on the journal’s Web site (www.spinejournal.com).

References

1. Walker BF. The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. J Spinal Disord 2000;13: 205–17.
2. Brazier J, Ratcliffe J, Salomon JA, et al. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford: Oxford University Press; 2007.
3. Tosh J, Brazier J, Evans P, et al. Review of generic preference-based measures of health-related quality of life in visual disorders. Value Health 2012;5:118–27.
4. Dyer MTD, Goldsmith KA, Sharples LS, et al. Review of health utilities using the EQ-5D in studies of cardiovascular disease. Health Qual Life Outcomes 2010;28:8–13.
5. Pickard AS, Wilke CT, Lin HW, et al. Health utilities using the EQ-5D in studies of cancer. PharmacoEconomics 2007;25:365–84.
6. Marra CA, Woolcott JC, Kopeck JA, et al. A comparison of generic, indirect utility measures (the HU2, HU3, SF-6D and EQ-5D) and disease specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60:1571–82.
7. Salaffi F, DeAngelis R, Stancati A, et al. Health-related quality of life in multiple musculoskeletal conditions: a cross-sectional population based epidemiological study II. The Mapping Study. Clín Exp Rheumatol 2005;23:829–39.
8. Fisk JD, Brown MG, Sketris IS, et al. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. J Neurol Neurosurg Psychiatry 2005;76:58–63.
9. Chapman JR, Norell DC, Hermens JH, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine 2011;36:554–8.
10. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21: 105–19.
11. Brooks R. EuroQol: the current state of play. Health Policy 1996;37:53–72.
12. Dolan P. Modelling valuations for EuroQol health states. Med Care 1997;35:949–59.
13. Fenny D, Furlong W, Torrance GW, et al. Multi-attribute and single-attribute utility functions for the Health Utility Index Mark 3 System. Med Care 2002;40:113–28.
14. Hayden J, van Tulder MW, Malimaara A, et al. Exercise therapy for treatment of non-specific low back pain (review). Cochrane Database Syst Rev 2011:2.
15. Lin CC, Haas M, Maher C, et al. Cost-effectiveness of guideline-endorsed treatments for low back pain: a systematic review. Eur Spine J 2011;20:1024–38.
16. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651–7.
17. Walters SJ. Quality of Life Outcomes in Clinical Trials and Healthcare Evaluation: A Practical Guide to Analysis and Interpretation. Chichester: Wiley; 2009.
18. Terwee CB, Dekker FW, Wiersinga WM, et al. On assessing responsiveness of health related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003;12:349–62.

Key Points

- EQ-5D showed good validity and responsiveness in patients with low back pain.
- EQ-5D can be used for economic evaluation of interventions targeting low back pain.
- EQ-5D appears unable to detect changes in health status at lower levels of severity.
- Assessment for SF-6D and HUI III was not possible because of lack of evidence.
19. Hurst NP, Kind P, Runa, et al. Measuring health related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of Euroqol (EQ-5D). Br J Rheumatol 1997;36:551–9.
20. McHorney CA, Ware JE, Raczek AE. The MOS-36 Item-Short Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247–63.
21. Kuntz MC, Harries WJ, Levitzky K, et al. Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. Med Care 1992;30:MS240–52.
22. Oliwares PR, Gusi N, Prieto J, et al. Fitness and health related quality of life dimensions in community dwelling middle aged and older adults. Health Qual Life Outcomes 2011;9:117.
23. Evans JD. Straightforward Statistics for the Behavioral Science. Pacific Grove: Brooks/Cole Publishing; 1996.
24. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale: Lawrence Erlbaum Associates; 1988.
25. Bastaen CH, De Bie RA, Vlaeyen JW, et al. Long-term effectiveness and cost-effectiveness of a brief self-management intervention in women with pregnancy-related low back pain after delivery. BMC Pregnancy Childbirth 2008;8:1–14.
26. Berg S, Fritzell P, Troppe H. Sex life and sexual function in men and women before and after total disc replacement compared with cervical fusion. Spine J 2009;9:887–94.
27. Berg S, Tullberg T, Branth B, et al. Total disc replacement compared to lumbar fusion: a randomised controlled trial with 2-year follow-up. Eur Spine J 2009;18:1512–9.
28. Carr JL, Klaber Moffett JA, Howarth E, et al. A randomized trial comparing a group exercise programme for back pain patients with individual physiotherapy in a severely deprived area. Disabil Rehabil 2005;27:929–37.
29. Caserley-Feeney SN, Daly L, Hurley DA. The access randomized clinical trial of public versus private physiotherapy for low back pain. Spine 2012;37:85–96.
30. Chown M, Whittamore L, Rush M, et al. A prospective study of patients with chronic back pain randomised to group exercise, physiotherapy or osteopathy. Physiother Theory Pract 2008;24:91–21.
31. Cox H, Tilbrook H, Aplin J, et al. A randomised controlled trial of yoga for the treatment of chronic low back pain: results of a pilot study. Complement Ther Clin Pract 2010;16:87–93.
32. Del Pozo-Cruz B, Hernandez Mocholi MA, Adua JC, et al. Effects of the whole body vibration therapy on main outcome measures for chronic non-specific low back pain: a single-blind randomize controlled trial. J Rehabil Med 2011;43:689–94.
33. Djais N, Kalim H. The role of lumbar spine radiography in the outcomes of patients with simple acute low back pain. APLAR J Rheum 2005;8:45–50.
34. Gilbert FJ, Grant AM, Gillan MG, et al. Does early imaging influence management and improve outcome in patients with low back pain? A pragmatic randomised controlled trial. Health Technol Assess 2004;8:1–131.
35. Gilbert FJ, Grant AM, Gillan MG, et al. Low back pain: influence of early MR imaging or CT on treatment and outcome—multi-centre randomized trial. Radiology 2004;231:434–51.
36. Hellum C, Johnsen LG, Storheim K, et al. Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: a two year follow-up of randomised study. BMJ 2011;342:2786.
37. Hurley DA, Minder PM, McDonough SM, et al. Interferential therapy electrode placement technique in acute low back pain: a preliminary investigation. Arch Phys Med Rehabil 2001;82:485–90.
38. Kendrick D, Fielding K, Bentley E, et al. Radiography of the lumbar spine in primary care patients with low back pain: randomised controlled trial. Br Med J 2001;322:400–5.
39. Miller P, Kendrick D, Bentley E, et al. Cost-effectiveness of lumbar spine radiography in primary care patients with low back pain. Lancet 2002;359:1217–20.
40. Rivero-Arias O, Gray A, Frost H, et al. Cost-utilty analysis of physiotherapy treatment compared with psychotherapy advice in low back pain. Spine 2006;31:1381–7.
61. Van der Roer N, Ostelo RW, Bekkering GE, et al. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. *Spine* 2006;31:578–82.

62. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ5D and SF-36, in Schizophrenia? A systematic review. *Value Health* 2011;14: 907–20.

63. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Econ* 2003;12:1061–7.

64. McDonough CM, Grove MR. Comparison of EQ5D, HUI III and SF36 derived societal health states values among spine patient outcomes research trial participants. *Qual Life Res* 2005;14: 1321–32.

65. Peasgood T, Brazier J, Papaioannou D. A systematic review on the validity and responsiveness of EQ5D and SF6D for depression and anxiety. HEDS Discussion Paper 12/15 [serial online.]. Available from: http://eprints.whiterose.ac.uk/74659/1/12.15.pdf.