Annals of Agricultural and Environmental Medicine 2014, Vol 21, No 4, 866–870
www.aaem.pl

ORIGINAL ARTICLE

Validation of the Polish language version of the SF-36 Health Survey in patients suffering from lumbar spinal stenosis

Michał Kłosiński[1,2*], Krzysztof A. Tomaszewski[2,3*], Iwona M. Tomaszewska[4], Piotr Kłosiński[1], Janusz Skrzat[2], Jerzy A. Walocha[2]

[1] *Department of Traumatology and Neuroorthopaedics, Rydygier Specialistic Hospital, Krakow, Poland*
[2] *Department of Anatomy, Jagiellonian University Medical College, Krakow, Poland*
[3] *Department of Orthopaedics and Trauma Surgery, 5th Military Hospital, Krakow, Poland*
[4] *Department of Medical Didactics, Jagiellonian University Medical College, Krakow, Poland*

* – these authors contributed equally to this study

Kłosiński M, Tomaszewski KA, Tomaszewska IM, Kłosiński P, Skrzat J, Walocha JA. Validation of the Polish language version of the SF-36 Health Survey in patients suffering from lumbar spinal stenosis. Ann Agric Environ Med. 2014; 21(4): 866–870. doi: 10.5604/12321966.1129948

Abstract

Introduction and objective. Patient-reported outcome (PRO) questionnaires have become the standard measure for treatment effectiveness after spinal surgery. One of the most widely used generic PROs is the SF-36 Health Survey. The aim of this study was to specifically focus on validating the SF-36 Health Survey to confirm that the tool is an acceptable and psychometrically robust measure to collect HRQoL data in Polish patients with spinal stenosis.

Materials and Methods. Patients were eligible if they were above 18 years of age and had been qualified for spine surgery of the lumbar region due to either discopathy or non-traumatic spinal stenosis. All patients filled-in the Polish version of the SF-36 and a demographic questionnaire. Standard validity and reliability analyses were performed.

Results. 192 patients (83 women – 43.2%) agreed to take part in the study (mean age: 57.5±11.4 years). In 47 patients (24.5%), using MRI, ossification of the ligamenta flava were found. Cronbach's alpha coefficients showed positive internal consistency (0.70–0.92). Interclass correlations for the SF-36 ranged from 0.72 – 0.86 and proved appropriate test-retest reliability. Satisfactory convergent and discriminant validity in multi-trait scaling analyses was seen.

Conclusions. The Polish version of the SF-36 is a reliable and valid tool for measuring HRQoL in patients with spinal stenosis. It can be recommended for use in clinical and epidemiological settings in the Polish population. However, caution is warranted when interpreting the results of the 'role limitations due to physical health problems' and the 'role limitations due to emotional problems' scales because of floor and ceiling effects.

Key words

ligamentum flavum; ossification; pilot-testing; SF-36; spinal stenosis; validation

INTRODUCTION

Patient-reported outcome (PRO) questionnaires have become the standard measure for treatment effectiveness after spinal surgery [1]. The most commonly used PRO questionnaires include pain scales for back and leg pain (visual analog scale/numeric rating scale) [2], Oswestry disability index [2], and the Short Form-36 (SF-36) Health Survey [3].

Broadly defined, there are two types of PROs (sometimes also referred to as 'health-related quality-of-life' questionnaires):

1) generic instruments intended for use both in general population surveys and in studies of patients with diverse health conditions;
2) disease-specific instruments developed for use among specific patient populations (e.g. cancer patients, diabetics, etc.) [4]. The majority of these measures have been developed in English-speaking countries and, until relatively recently, the evidence supporting their validity and reliability has been derived primarily from studies conducted among English-speaking patients [4]. However, the growing interest expressed both by the public (e.g. government health care agencies, patient groups) and private (pharmaceutical industry) sectors facilitates the development of national-level studies that focus on the adaptation and validation of PROs [5].

As mentioned before, one of the most widely-used generic PROs is the SF-36 Health Survey. The SF-36 was developed in the USA in the late 1980s as part of the Medical Outcomes Study, a longitudinal investigation of the self-reported health status of patients with a range of chronic conditions [3].

There is an ongoing debate whether health-related quality-of-life (HRQoL) can be a valid proxy in patients undergoing spine surgery [1]. Chow et al. [6] have defined patient satisfaction as 'The degree to which patients feel they have received high quality health care'. However, distinct from quality and effectiveness, satisfaction is an entirely subjective measurement, defined differently by different people and related to many factors, including age, gender, education, lifestyle, expectations, psychological status, and individual values [1, 6]. Many authors advocate the importance of collecting patient QoL data [6, 7]. Chow et al. [6] view satisfaction as the 'ultimate end-point to the health-care pathway'. However, other authors see satisfaction as a simple measure of service, and of secondary importance to safety and effectiveness of care [1, 8]. Most physicians across all specialties agree that care delivery that is ineffective or not safe is of low quality, regardless of whether patients are satisfied with their health-care service [1].

Address for correspondence: Krzysztof A. Tomaszewski, Department of Anatomy, Jagiellonian University Medical College, 12 Kopernika street, 31–034 Krakow, Poland
e-mail: krtomaszewski@gmail.com

Received: 28 January 2014; Accepted: 12 April 2014

To add to this discussion the authors of the presented study decided to validate and assess the acceptability of the Polish version of the SF-36 [9] in patients suffering from lumbar spinal stenosis, and who were being prepared to undergo surgery of the spine.

OBJECTIVES

The aim of the study was to specifically focus on validating the SF-36 Health Survey to confirm that the tool is an acceptable and psychometrically robust measure to collect HRQoL data in Polish patients with spinal stenosis. The authors have previous experience in performing this kind of validation studies [10, 11, 12].

MATERIALS AND METHOD

Patients. The patients were recruited prospectively between January 2011 – September 2013 in the Department of Traumatology and Neuroorthopaedics in the Rydygier Specialist Hospital in Krakow, Poland. Patients were eligible if they were above 18 years of age and had been qualified for spine surgery of the lumbar region due to either discopathy or non-traumatic spinal stenosis. Exclusion criteria included lack of consent to participate in the study, inability to understand or complete the questionnaires, and spinal stenosis due to a malignant process or trauma.

The protocol of the study was approved by the Jagiellonian University Medical College Bioethical Committee (Registry No. KBET/176/B/2011). Each patient gave informed consent to participate in the study, which was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

The patients included were classified into groups based on whether or not their ligamenta flava were ossified (ossification of the ligamentum flavum – OLF, determined on the basis of MRI) to additionally validate the SF-36 in this subgroup of patients.

Clinical history and physical examination were registered for all patients from patient files. Each patient qualified for the study had either a CT or MRI performed of the lumbosacral region. MRI was the imaging method of choice, and CT was only performed in the case of patients having metal implants preventing MRI. Imaging allowed assessment of the level of spinal stenosis. This level was always treated as the site from which the OLF was excised. Ligamentum flavum thickening and spinal stenosis were defined as per the definitions given by Sakamaki et al. [13].

Interview procedure. The patients were approached one day before the surgery and informed about the study. They were interviewed only after written informed consent was obtained. Each patient completed the Polish version of the SF-36 and a questionnaire concerning demographic data. The questionnaires were administered by qualified clinical staff – medical doctors.

A subset of randomly chosen (based on a computer generated algorithm) patients completed the questionnaires twice for evaluation of stability (n=30) and responsiveness (n=30). Patients completing the SF-36 for stability were assessed at 14 days pre-operatively and one day pre-operatively. Patients completing the SF-36 for responsiveness were assessed 1 day pre-operatively and 42 days post-operatively. All patients agreed to fill-in the questionnaire a second time.

The SF-36 Health Survey. This survey is composed of 36 questions and standardized response choices, organized into eight multi-item scales:
- physical functioning (PF);
- role limitations due to physical health problems (RP);
- bodily pain (BP), general health perceptions (GH);
- vitality (VT), social functioning (SF);
- role limitations due to emotional problems (RE);
- general mental health (MH).

All raw scale scores were linearly converted to a 0 – 100 scale, with higher scores indicating higher levels of functioning or well-being. In this study, the pre-translated Polish version of the SF-36 was used [9].

Measures of SF-36 acceptability. The acceptability of the SF-36 was assessed by the response rate, percentage of missing data, assistance and time needed to complete the questionnaire, and details of items considered upsetting, confusing or difficult in the questionnaire [10, 12]. Assessment of whether the patients found any of the SF-36 questions 'upsetting, confusing or difficult' was carried out by asking the patients directly which (if any) of the SF-36 questions were upsetting, confusing or difficult. If a patient answered 'yes' to any of the above, he was asked for additional comments on this subject.

Statistical analysis. Several pre-planned standard psychometric tests were conducted. These approaches can be seen in the EORTC Module Development Guidelines [14]. Scoring of the two measures followed the standard scoring instructions [3, 4, 9]. To analyse the data, descriptive statistics (mean, standard deviation, percentage distribution) were used.

The significance level was set at $p<0.05$. Statistical analysis was conducted using computer software Statistica 10.0 PL by StatSoft Poland, licensed to the Jagiellonian University Medical College In Kraków.

Statistical analysis – Validity. To confirm the hypothesized scale structure of the SF-36, convergent and discriminant validity were used. Convergent validity was assessed by correlating each item with its own scale of the SF-36 [4, 10, 12]. Evidence of item convergent validity was defined as a correlation of 0.40 or greater between an item and its own scale (corrected for overlap). Discriminant validity was assessed by correlating each item with any other scale of the SF-36 [4, 10, 12]. 'Any other scale' means each of the SF-36 scales, apart from the scale from which the relevant item originates. A scaling success for an item was seen when the correlation between an item and its own scale (corrected for overlap) was significantly higher (i.e. two standard errors or greater) than its correlation with other scales [15].

Clinical validity was assessed using the Wilcoxon rank sum nonparametric test. This assesses if the questionnaires were able to discriminate between subgroups of patients differing in clinical status [12]. The known-group used in this study was *a priori* based on the presence or lack of OLF.

Statistical analysis – Reliability. Cronbach's alpha coefficient was calculated to assess the internal consistency of the Polish version of the SF-36. Internal consistency estimates of a magnitude of >0.70 were considered acceptable for group comparisons [15]. Test-retest reliability of the SF-36 was assessed using interclass correlations (ICC) between baseline and retest. A correlation of >0.80 was considered 'good' [15].

Statistical analysis – Responsiveness to change over time. Assessment of responsiveness of the scales to treatment was performed by comparing pre-treatment and post-treatment assessments of patients (n=30). Due to the non-normality of the data, the Mann-Whitney U test was used.

Statistical analysis – Sample size calculation. Study sample size was based on the proposal of Tabachnick and Fidell [16], which states that in order to obtain reliable estimates, the number of observations should be 5–10 times the number of variables in the model. Thus, the required number of patients to conduct this study was between 180 – 360.

RESULTS

Patient characteristics and acceptability. During the 32 month recruitment period a total of 192 patients (83 women – 43.2%) agreed to take part in the study, with a mean age of 57.5±11.4 years. Patients' demographic data are presented in Table 1. In 47 patients (24.5%), using MRI, OLF were found.

Table 1. Patients' demographic data

Variable	Overall n=192
Age (mean±SD)	57.5±11.4
Female (%)	83 (43.2%)
Male (%)	109 (56.8%)
Education (%)	
Elementary	62 (32.3%)
High School or equivalent	101 (52.6%)
University	29 (15.1%)
Current working status (%)	
Employed	103 (53.6%)
Unemployed	17 (8.9%)
Retired/Pensioner	72 (37.5%)
Living (%)	
Alone	23 (12.0%)
With partner or family	162 (84.4%)
With others	7 (3.6%)

During the recruitment period, 243 patients who qualified for the study were approached. Of this number, 192 (79%) agreed to take part in the study. Overall, only 4.3% of item responses were missing.

Fifty-two interviewees (27.1%) required assistance completing the questionnaires, mostly in order to read the items and mark the answers. The total time for completion of the questionnaires and interview was 13.8±2.1 minutes without assistance and 24.5±3.7 with assistance.

Table 2 presents means, standard deviations, and percentage of floor and ceiling for SF-36 scales.

Table 2. Means, standard deviations, and percentage of floor and ceiling for SF-36 scales

SF-36 scales	Whole group (n=192)		
	Mean (SD)	Floor (%)	Ceiling (%)
PF	71.1 (20.3)	1	10.9
RP	54.0 (18.6)	10.4	16.1
BP	63.9 (27.2)	2.1	14.6
GH	64.5 (18.0)	1.6	3.7
VT	68.7 (20.6)	1	1.6
SF	84.1 (16.2)	1.6	24.5
RE	69.9 (31.4)	14.1	21.4
MH	71.8 (18.3)	0.5	2.6

SD – standard deviation; PF – physical functioning; RP – role limitations due to physical health problems; BP – bodily pain; GH – general health perceptions; VT – vitality; SF – social functioning; RE – role limitations due to emotional problems; MH – general mental health.

Reliability and validity. Results of multi-trait scaling analyses are presented in Table 3.

Table 3. SF-36 multi-trait scaling analyses

SF-36 scales	Whole group (n=192)		
	Convergent validity[1]	*Discriminant validity*[2]	*Cronbach's alpha*
PF	0.54–0.69	0.14–0.39	0.92
RP	0.50–0.77	0.21–0.32	0.85
BP	0.49	0.01–0.21	0.72
GH	0.61–0.72	0.09–0.41	0.78
VT	0.44–0.65	0.11–0.18	0.71
SF	0.62	0.06–0.43	0.70
RE	0.48–0.67	0.03–0.22	0.74
MH	0.42–0.51	0.14–0.29	0.73

SD – standard deviation; [1] – Item-own scale correlation, Spearman correlation coefficient, corrected for overlap; [2] – Item-other scale correlation, absolute values displayed, Spearman correlation coefficient.
PF – physical functioning; RP – role limitations due to physical health problems; BP – bodily pain; GH – general health perceptions; VT – vitality; SF – social functioning; RE – role limitations due to emotional problems; MH – general mental health.

Taking into account the SF-36, its own-scale correlations were considered good. All item correlations within their own scales exceeded the 0.40 criterion, and were correlated higher with their own scale than with the other scales. All presented Cronbach alpha values exceeded the 0.7 criterion.

For test-retest assessment ICC was used. The ICC's for the SF-36 ranged from 0.72 – 0.86, showing good repeatability of the scales.

Clinical validity assessment by known-group comparison showed that the SF-36 was not able to discriminate between patients with and without OLF (p>0.05)

Responsiveness to treatment. Differences between pre-surgery and on-surgery assessments were evaluated for the scales of the SF-36. The scales that displayed significant differences between the two assessments were: PF (p<0.001), BP (p<0.001), RP (p=0.01), GH (p=0.01) and MH (p=0.03). The VT (p=0.54), SF (p=0.61) and RE (p=0.39) scales failed to display treatment-associated differences.

DISCUSSION

The presented manuscript reports on the validation of the SF-36 Health Survey to confirm that this tool is an acceptable and psychometrically robust measure to collect HRQoL data in Polish patients with spinal stenosis. To the best of the authors' knowledge, this is the first study to validate the SF-36 in Polish patients suffering from orthopaedic/neurosurgical problems, in this case, spinal stenosis.

As new treatment options arise, it is imperative to remember that HRQoL should always accompany the surgical outcome. Generic HRQoL measures may help to assess the overall HRQoL of a patient, and thus highlight important, treatment-related issues.

There is no doubt that the use of PROs represents an important step towards patient-centered care, and can help drive the demand for a specific health-care entity in a consumer-driven market [1]. However, the authors agree that patient satisfaction scores alone should not be used to represent the overall quality of spine care. Patient-centered measures of safety and effectiveness of care should remain the most important measures of quality. Compromised safety and effectiveness of care in the setting of high patient satisfaction undermine the aims of the quality movement. As this study shows, PROs can be helpful in quality improvement; however, it is agreed that they should not be used as a proxy for overall quality of care in surgical spine care.

The results of the presented study indicate that the Polish version of the SF-36 demonstrates good agreement with the original questionnaire and other language versions [4, 17, 18]. The findings described in this study show and confirm that the SF-36 has adequate levels of cross-cultural validity, and might also be applicable to other languages and cultures. However, it has to be borne in mind that such cultural aspects may influence the relationship between overall HRQoL and its sub-dimensions [19].

The SF-36 proved to be acceptable to the tested sample of patients. This is further enhanced by the low number of missing item responses. In the original study [20], as well as other validation studies [4, 17, 18], construct analysis of the EORTC QLQ-OV28 confirmed the presence of eight distinct scales, in which items within each scale were highly correlated with one another, compared with items from other scales. The analysis also showed appropriate Cronbach's alpha values for all of the SF-36 scales. Even though some scales had a border Cronbach's alpha of 0.7, it was recognized that these are only guidelines, rather than simple cut-off or threshold scores. Test-retest values were considered good, as was responsiveness to change, which showed that most of the SF-36 scales respond to patient's change in HRQoL following surgical treatment.

The results of known-group comparison demonstrated that the SF-36 is not able to discriminate between patient subgroups differing in clinical status. This is most probably caused by the fact that OLF formation is usually asymptomatic [21]. Even when OLF becomes symptomatic, they are hard or impossible for the patients and clinicians to distinguish from other causes of spinal stenosis. This is only possible with the use of appropriate imaging modalities. However, due to the results of this study, it is now known that the SF-36 is a valid HRQoL measure also in patients with spinal stenosis caused by OLF formation.

Problems were detected with two SF-36 scales – RP and RE – both of which had significant floor and ceiling effects. This may have considerable implications for the interpretation of treatment effects because floor and ceiling effects represent cohorts of people whose scale scores may not be accurate measurements of their true level of functioning [22]. An alternative explanation is that the floor-ceiling effects on the two scales represent a cohort of people whose functioning could not be improved [22].

These findings do not preclude the use of the SF-36 in spine surgery; they simply underline the importance of using a scale that matches the spectrum of health covered by the study sample. Even generic measures have some degree of specificity, and it should be recognized that the term 'generic' is relative and does not indicate universal applicability [22]. Following the study of Baron et al. [22], it is suggested that the RP and RE SF-36 scales should not be used to measure changes in health status in either routine clinical practice, or in clinical trials, because floor and ceiling effects are likely to lead to an underestimate of treatment effectiveness.

This study has one limitation: the fact that responsiveness of the scales to treatment was performed on a fairly small group of patients. This warrants caution when interpreting the results of this part of the statistical analysis.

The findings of this study have demonstrated that SF-36 scale scores are valid, but have certain limitations that somewhat restrict their usefulness in the evaluation of HRQoL in patients undergoing spine surgery.

CONCLUSIONS

In conclusion, the Polish version of the SF-36 is a reliable and valid tool for measuring HRQoL in patients with spinal stenosis. It can be recommended for use in clinical and epidemiological settings in the Polish population. However, caution is warranted when interpreting the results of the 'role limitations due to physical health problems' and 'role limitations due to emotional problems' scales because of floor and ceiling effects.

Acknowledgements

Krzysztof A. Tomaszewski received a scholarship to prepare his PhD thesis from the National Science Center – Poland under award number DEC-2013/08/T/NZ5/00020.

REFERENCES

1. Godil SS, Parker SL, Zuckerman SL, Mendenhall SK, Devin CJ, Asher AL, et al. Determining the quality and effectiveness of surgical spine care: patient satisfaction is not a valid proxy. Spine J. 2013; 13(9): 1006–1012.
2. Grönblad M, Hupli M, Wennerstrand P, Järvinen E, Lukinmaa A, Kouri JP, et al. Intercorrelation and test-retest reliability of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ) and their correlation with pain intensity in low back pain patients. Clin J Pain. 1993; 9(3): 189–195.
3. Ware JE Jr. SF-36 health survey update. Spine (Phila Pa 1976). 2000; 25(24): 3130–3139.
4. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol. 1998; 51(11): 1055–1068.
5. American Society of Clinical Oncology. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. J Clin Oncol. 1996; 14(2): 671–679.

6. Chow A, Mayer EK, Darzi AW, Athanasiou T. Patient-reported outcome measures: the importance of patient satisfaction in surgery. Surgery. 2009; 146(3): 435–443.
7. Korolija D, Wood-Dauphinee S, Pointner R. Patient-reported outcomes. How important are they? Surg Endosc. 2007; 21(4): 503–507.
8. Wright JG. Outcomes research: what to measure. World J Surg. 1999; 23(12): 1224–1226.
9. Tylka J, Piotrowicz R. Quality of life SF-36 questionnaire – the Polish version. Kardiol Pol. 2009; 67(10): 1166–1169 (in Polish).
10. Püsküllüoğlu M, Tomaszewski KA, Bottomley A, Holden L, Tomaszewska IM, Głowacki R, et al. Validation of the Polish version of the EORTC QLQ-BM22 module for the assessment of health-related quality of life in patients with bone metastases. Qual Life Res. 2013. doi: 10.1007/s11136-013-0486-6.
11. Chmielowska K, Tomaszewski KA, Pogrzebielski A, Brandberg Y, Romanowska-Dixon B. Translation and validation of the Polish version of the EORTC QLQ-OPT30 module for the assessment of health-related quality of life in patients with uveal melanoma. Eur J Cancer Care (Engl). 2013; 22(1): 88–96.
12. Paradowska D, Tomaszewski KA, Bałajewicz-Nowak M, Bereza K, Tomaszewska IM, Paradowski J, et al. Validation of the Polish version of the EORTC QLQ-CX24 module for the assessment of health-related quality of life in women with cervical cancer. Eur J Cancer Care (Engl). 2013. doi: 10.1111/ecc.12103.
13. Sakamaki T, Sairyo K, Sakai T, Tamura T, Okada Y, Mikami H. Measurements of ligamentum flavum thickening at lumbar spine using MRI. Arch Orthop Trauma Surg 2009; 129(10): 1415–1419.
14. Johnson CD, Aaronson N, Blazeby JM, Bottomley A, Fayers P, Koller M, et al. Guidelines for developing Quality of Life Questionnaires. 4th ed. Brussels (EORTC Publications), 2011.
15. Fayers P, Machin D. Quality of Life: The Assessment Analysis and Interpretation of Patient Reported Outcomes. Chichester (Wiley), 2007.
16. Tabachnick BG. Fidell LS. Using Multivariate Statistics. 4th ed. London (HarperCollins), 2011.
17. Pappa E, Kontodimopoulos N, Niakas D. Validating and norming of the Greek SF-36 Health Survey. Qual Life Res. 2005; 14(5): 1433–1438.
18. Montazeri A, Goshtasebi A, Vahdaninia M, Gandek B. The Short Form Health Survey (SF-36): translation and validation study of the Iranian version. Qual Life Res. 2005; 14(3): 875–882.
19. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. The relationship between overall quality of life and its subdimensions was influenced by culture: analysis of an international database. J Clin Epidemiol. 2008; 61(8): 788–795.
20. Gandek B, Ware JE Jr, Aaronson NK, Alonso J, Apolone G, Bjorner J, et al. Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol. 1998; 51(11): 1149–1158.
21. Kang KC, Lee CS, Shin SK, Park SJ, Chung CH, Chung SS. Ossification of the ligamentum flavum of the thoracic spine in the Korean population. J Neurosurg Spine. 2011; 14(4): 513–519.
22. Baron R, Elashaal A, Germon T, Hobart J. Measuring outcomes in cervical spine surgery: think twice before using the SF-36. Spine (Phila Pa 1976). 2006; 31(22): 2575–2584.