Spine

SPINE Volume 46, Number 2, pp E118–E125

HEALTH SERVICES RESEARCH

Measurement Properties of the Oswestry Disability Index in Recipients of Lumbar Spine Surgery

Chad E. Cook, PhD, PT, FAPTA,[a] Alessandra N. Garcia, PhD, PT,[a] Alexis Wright, PhD, PT,[b] Christopher Shaffrey, MD,[c] and Oren Gottfried, MD[c]

Study Design. This is an observational study on the measurement properties of the Oswestry Disability Index (ODI) version 1.0.

Objectives. To (1) determine the construct validity of the tool, specifically structural validity; (2) analyze the criterion validity of the tool, specifically concurrent validity against proxy measures of pain, function, and quality of life and predictive validity of each item to proxy measures of disability; and (3) reliability of the tool, specifically internal consistency.

Summary of Background Data. We endeavored to investigate the measurement properties of the ODI on a spine surgery population to test the assumption that a more disabled population may influence the properties of the tool.

Methods. Data were pulled from the Quality Outcomes Database (QOD) Spine Registry. A total of 57,199 participants who underwent primary or revision lumbar spine surgeries were included. Structural validity was assessed by exploratory and confirmatory factor analysis, concurrent validity, predictive validity by odds ratios, and internal consistency by Cronbach alpha. The Visual Analog Scale for back pain, two standard open questions, and the EuroQol 5 Dimension/Visual Analogue Scale were included as proxy measures of pain, function, and quality of life, respectively. Hospital readmission, return to operating room for treatment and revision surgery (all within 30 days) were included as proxy measures of disability to assess the predictive validity of each ODI item.

Results. The ODI demonstrated a two-factor structural solution, which explained 54.9% of the total variance. Fair internal consistency (0.74–0.77), and fair criterion validity (concurrent) and significant findings with predictive validity ($P < 0.01$) substantiated the use of each item of the ODI as well as the summary score and ODI thresholds.

Conclusion. Our study lends value to a burgeoning repository of evidence that suggests the ODI is a useful tool for capturing outcomes in clinical practice. We recommend its continued use in clinical practice.

Key words: measurement properties, patient-reported outcome measures, spine surgery.

Level of Evidence: 4

Spine 2021;46:E118–E125

The Oswestry Disability Index (ODI) is one of the most common patient-reported outcome measures used to evaluate the impact of back pain on patients' activities of daily living.[1] The ODI was developed in 1976[2] and was first published in 1980.[3] Version 2.0 was published 20 years later, after several revisions.[2] The ODI versions (1.0, 2.0, 2.1, and 2.2) are commonly used outcome measures and have been translated into over 80 distinct languages and cultures.[4] Regarding its measurement properties, the ODI is valid, reliable, and easy to administer and score in a clinical setting.[5] Multiple studies have indicated that ODI is a unidimensional tool with good internal consistency.[6–8] Item response theory has shown that the discriminative ability of the ODI is better when the patient population has higher levels of disability.[9,10] This suggests that the population in which the ODI is tested may influence the measurement properties of the tool.

By definition, measurement properties include validity, reliability, and responsiveness.[11] Accordingly to COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN),[12] the measurement property of validity includes construct validity (the degree to which a test measures what it claims, or purports, to be measuring), content validity (the extent to which a measure

From the [a]Duke University Division of Physical Therapy, Duke Department of Orthopaedic Surgery, Duke Clinical Research Institute, Durham, NC; [b]Department of Public Health and Community Medicine, School of Medicine, Tufts University; and [c]Department of Neurosurgery, Duke University Medical Center, Durham, NC.

Acknowledgment date: April 30, 2020. First revision date: June 23, 2020. Acceptance date: August 6, 2020.

The manuscript submitted does not contain information about medical device(s)/drug(s).

No funds were received in support of this work.

Relevant financial activities outside the submitted work: consultancy.

Address correspondence and reprint requests to Chad E. Cook, PhD, PT, FAPTA, Duke University, 311 Trent Drive, Durham, NC 27710; E-mail: chad.cook@duke.edu

DOI: 10.1097/BRS.0000000000003732

represents all facets of a given construct), and criterion validity (the extent to which a measure is related to an outcome). Reliability may include reliability (*e.g.,* test-retest reliability), measurement error, and internal consistency (*i.e.,* correlations between the different items within a scale). Responsiveness is unidimensional and reflects the ability of the tool to measure when change has actually occurred. Prior to or continue use in clinical practice requires careful evaluation of the measurement properties of patient-reported outcomes measures (PROMs).[12] Because examination of each of the measurement properties is complex, most studies examine 1 to 4 of the properties of the given PROM.

The ODI has become an important clinical measure in the medical routine to evaluate disability in patients undergoing spine surgery.[1] Spine surgery is an invasive approach that is performed on individuals who have failed conservative care and have extreme pain and/or deformity.[13,14] Surgery recipients typically exhibit higher levels of disability and greater challenges associated with function. We endeavored to investigate the ODI on a spine surgery population from a nationwide registry to test the assumption that a more disabled population may influence the measurement properties of the tool. We aimed to (1) determine the construct validity of the tool, specifically structural validity (which is a submeasure of construct validity, and is the degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured); (2) analyze the criterion validity of the tool, specifically (a) concurrent validity (how well a test or test item correlates with another construct) against proxy measures of pain, function, and quality of life and (b) predictive validity of each item to measures of future complications (hospital readmission, return to operating room for treatment, and revision surgery, all within 30 days); and (3) reliability of the tool, specifically internal consistency (which is the degree of the interrelatedness among the items). We hypothesized that the ODI would exhibit acceptable measurement properties, even in a population of patients (*i.e.,* spine surgery recipients) who have higher levels of disability than most of the previously reported studies.

METHODS

Study Design and Setting

Data for this observational study on measurement properties of ODI were taken from a subset of lumbar spine surgeries only from the Quality Outcomes Database (QOD) Spine Registry. The dataset included patient participation from 2012 to 2018.[15] Data from the QOD registry is voluntarily pulled from multiple clinical sites across 38 US states. The QOD reports health-related factors and outcomes measures with baseline, 3-month, 1-year, and 2-year follow up. We followed the REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement to guide the study reporting,[16] and the COSMIN Study Design checklist to guide study design.[17] COSMIN has created a taxonomy, which we have adopted the terminology for this study.[18] The Institutional Review Board at Duke University determined the protocol to be exempt (Pro00029554).

Participants

All participants had a diagnosis of degenerative lumbar disorders *(e.g.* disc herniation, lumbar stenosis, spondylolisthesis). We included all individuals who had baseline ODI items, whether their lumbar surgery was a primary or revision surgery (*e.g.,* microdiscectomy and laminectomy with or without fusion). The QOD excludes individuals less than 18 years of age and with severe lumbar disorders (*e.g.,* spine infection, tumor, fracture, traumatic dislocation, and neurologic paralysis).[19] Our subset of the QOD consisted of 57,199 individuals.

Measurement Properties and Comparative Variables

To avoid mode of administration bias, all clinical sites across 38 US states administered the original ODI (version 1.0) using a standardized instructional set. The self-completed ODI contains 10 topics concerning intensity of pain, lifting, ability to care for oneself, ability to walk, ability to sit, sexual function, ability to stand, social life, sleep quality, and ability to travel. Patient were instructed to check the statement, which most closely resembles their situation. Each question is scored on a scale of 0–5 with the first statement being zero and indicating the least amount of disability and the last statement is scored 5 indicating most severe disability. Participants were given the opportunity to "skip" the sex question of the ODI. The scores for all questions answered are summed, then multiplied by two to obtain the index (range 0–100), and modified if an item had a missing value. The originators of the ODI have identified benchmarks of 21 to 40 as moderate disability, 41 to 60 as severe disability, 61 to 80 as "crippled", and 81 and above as either bed bound or exaggerating their symptoms.[2] For this study, we calculated the ODI raw score (0–50), the ODI summative percentile score (0%–100%), and dichotomized the ODI percentile scores to 21 and higher (moderate disability), 41 and higher (severe disability), and 61 and higher ("crippled").

All 10 items of the ODI at baseline were included in the structural validity, internal consistency, concurrent validity, and predictive validity testing. Structural validity refers to "the degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured."[12] Internal consistency refers to "the degree of the interrelatedness among the items"[12] and concurrent validity is defined as "a subtype of criterion validity domain and is defined as a method that involves administering the test of interest and reference standard test at nearly the same time."[20] Table 1 provides details on the variables and measures used in this study. Details on descriptive variables are also reported in Table 1.

Missing Data and Cleaning

Less than 0.2% of ODI items were missing, with the exception of the item associated with sex life, which was

TABLE 1. Variables and Measures Used in this Study

Measure	Description	Scoring
ODI[2]	The Oswestry Disability Index (ODI) has questions associated with pain, personal care, lifting, walking sitting, standing, sleeping, sex, social life, and traveling	For raw scoring of the ODI, each of the 10 ODI items, there are six possible Likert-type selections, unique to each item, scored from ''0'' (low disability) to ''5'' (high disability). Each of the 10 items is summated for a total summary score of ''0'' (low disability) to ''50'' (high disability). For summary scoring of the ODI, the raw scored was doubled. The index was calculated by dividing the summed score by the total possible score, which is then multiplied by 100 and expressed as a percentage. Thus, for every question not answered, the denominator is reduced by.
Proxy measures		
Pain (VAS)[25]	Visual Analog Scale (VAS) is a 0-10 scale that measures pain intensity	''0'' represents ''no pain'' and ''10'' represents the ''worst possible pain.''
Function (standard baseline questions)	These questions involve the ''ability to perform usual activities'' and ''ability to perform self-care activities''.	There were three responses: ''no problems'', ''some problems'', and ''inability to either do usual activities or perform self-care''.
Quality of life (EQ-5D VAS)	Quality of life was evaluated using the EuroQol 5 Dimension, 5-Level (EQ-5D-5L) Visual Analogue Scale (EQ-5D VAS).[26]	Values are scored from 0 to 100, with higher values reflecting better-reported levels of quality of life.
Proxy measures of disability (predictive validity of each ODI item)		
30-day hospital readmission	As a measure of significance, 30-day hospital readmission has been well-defined as a measure of importance.[28]	Dichotomized as ''yes'' or ''no''
30-day return to operating room for treatment	A return to an operating room generally is caused by a major complication and entails a significant amount of risk.[29]	Dichotomized as ''yes'' or ''no''
Revision surgery within 30 days	Spinal revision surgeries may occur because of instrumentation failure, radicular/myelopathic changes, adjacent joint disease, scarring, and/or progression of deformity.[30,31]	Dichotomized as ''yes'' or ''no''
Descriptive variables		
Age	Mean (SD)	Continuous variable
Gender	Female or male	Dichotomized as ''yes'' or ''no''
Race	Caucasian percentage	Dichotomized as ''yes'' or ''no''
Hispanic ethnicity	Hispanic or others	Dichotomized as ''yes'' or ''no''
Patient education	Less than high school, High school diploma, 2-year college degree, 4-year college degree, Postcollege	Dichotomized as ''yes'' or ''no''
Height and weight	Inches and pounds	Continuous variable
Insurance type	Uninsured, medicare, medicaid Veterans Affairs/Government or Private	Nominal variable
Employment status	Employed and working, employed but not working, or unemployed	Nominal variable
Baseline ODI raw item score	The ODI consists of 10 questions involving the following concepts: personal care, lifting, walking, sex social activities, pain intensity, sleep, standing, traveling, and sitting	For each of the 10 ODI items, there are six possible Likert-type selections, unique to each item, scored from ''0'' (low disability) to ''5'' (high disability). A total raw summary score of ''0'' (low disability) to ''50'' (high disability)
Baseline ODI summary score	The ODI consists of 10 questions involving the following concepts: personal care, lifting, walking, sex social activities, pain intensity, sleep, standing, traveling, and sitting	For summary scoring of the ODI, the raw scored was doubled. The index was calculated by dividing the summed score by the total possible score, which is then multiplied by 100 and expressed as a percentage. Thus, for every question not answered, the denominator is reduced by.

TABLE 1 (*Continued*)

Measure	Description	Scoring
American Society of Anesthesiologists (ASA) grade	Reflects the general health of an individual and is categorized into six groups: ASA (1) a healthy individual, ASA (2) a mild systemic disease, ASA (3) a severe systemic disease, ASA (4) a patient with severe systemic disease that is a constant threat to life, ASA (5) A moribund patient who is not expected to survive the operation, and ASA (6) A declared brain-dead patient whose organs are being removed for donor purposes.[32]	Nominal variable
Medical diagnoses	Diabetes, coronary artery disease, peripheral vascular disease, anxiety, depression, osteoarthritis, renal disease, chronic obstructive pulmonary disease, osteoporosis, and multiple sclerosis.	Dichotomized as ''yes'' or ''no''

missing in 42.7% of cases (participants were allowed to ''skip'' this question as part of the standardized instructional set). There were no missing values for EQ-5D, usual activities, and self-care activities. Thirty-day hospital readmission, 30-day return to operating room, and revision missing value percentages were 2.4%, 2.0%, and 23.7%, respectively. Because the data were primarily ordinal or nominal, we elected to use Listwise deletion, a method for handling missing data, in which an entire record is excluded from analysis if any single value is missing.[21]

Statistical Analysis

All analyses were conducted in IBM Statistical Package for the Social Sciences IBM SPSS Statistics for Windows, version 26.0 (IBM Corp., Armonk, N.Y., USA). Descriptive statistics were tabulated into mean and standard deviations (using linear mixed-effects modeling) and frequencies and percentages (using Pearson Chi-square test) for the whole group. Statistical significance was set at $P < 0.05$.

Structural Validity

Construct validity (structural validity) was determined by using exploratory and confirmatory factor analysis. Exploratory factor analysis (EFA) is a statistical method used to uncover the underlying structure of a relatively large set of variables.[17] Confirmatory factor analysis (CFA) is a multivariate statistical procedure that is used to test how well the measured variables represent the number of constructs defined in the EFA. Knowing that factor analyses includes both quantitative and qualitative judgments, we a priori determined thresholds on what we identified as appropriate representative dimensions.

For both factor analyses, Eigenvalues greater than 1.0 were considered representative of a unique dimension and were confirmed by parallel analysis and the visual Scree test. Eigenvalues are a special set of characteristic values associated with a linear system of equations. The Scree test is a graphic test for determining the number of factors to be retained and is based on the identification of important factors versus more trivial factors that should not be retained. To determine whether the given data were appropriate for both factor analyses, a Kaiser-Meyer-Olkin (KMO) statistic was used to determine if the sum of partial correlations was greater than the sum of correlations. We also analyzed the Bartlett test for sphericity, which is a test for normalization of the sample and the strength of the relationship among variables. Variables were considered as defining part of a factor when factor structure loading coefficients were ≥ 0.4.

Internal Consistency

Reliability (internal consistency) for each factor construct was analyzed using Cronbach alpha. Internal consistency is an assessment of how reliably survey or test items that are designed to measure the same construct actually do so. Cronbach alpha values are defined as:0.00 to 0.69 = Poor; 0.70 to 0.79 = Fair; 0.80 to 0.89 = Good; and 0.90 to 0.99 = Excellent/Strong.[17]

Concurrent Validity

Concurrent validity was analyzed by comparing the ODI items' raw scores, the factor loadings, the summative ODI percentile scores, and three ODI scores dichotomized to include score 21 or higher, 41 or higher, and 61 or higher against the proxy measures of pain, function (usual activities and self-care activities), and quality of life. Ordinal to nominal measures (*e.g.*, each ODI item versus usual activities) were calculated using a rank biserial correlation. Ordinal to continuous (*e.g.*, each ODI item versus pain) measures were calculated using a Kendall coefficient rank tau-sub-b. Nominal to nominal values (*e.g.*, ODI of 21 and above versus activities) were calculated using a Phi coefficient. Proposed boundaries for correlations used in medicine are as follows: 0 = none, 0.01 to 0.29 = poor, 0.30 to 0.59 = fair, 0.60 to 0.79 = moderate, 0.80 to.99 = very strong, and 1.0 = perfect.[22]

Predictive Validity

Predictive capacity of each ODI item, the factor loadings, the ODI summative percentile score, and three ODI scores dichotomized to include score 21 or higher, 41 or higher, and 61 or higher to future complications, specifically 30-day

hospital readmission, 30-day return to the operating room for treatment, and revision surgery within 30 days were measured using logistic regression analysis. Odds ratio (ORs) and 95% confidence intervals (CIs) were used to express the strength of association between variables. A P-value of < 0.05 was considered significant.

RESULTS

Descriptive Statistics

Our sample consisted of 57,199 spine surgery recipients. Over 80% of the recorded cases involved primary spine surgeries. The mean age was 58.8 (SD = 14.3), mostly male (52.9%), White (88.7%), and non-Hispanic (94%), and 50.1% had completed some college or more. The mean height was 67.4 (SD = 4.2) inches and the mean weight was 198.6 (SD = 46.6) pounds. Over 53% reported they were unemployed at the time of surgery. Thirty-eight percent (38%) of individuals had Medicare insurance or private insurance (52.8%). The baseline ODI raw score was 23.1 out of 50 (SD = 8.1). The baseline summative percentage score was 47.2 (SD = 16.4-range 0–100) suggesting 'severe' disability.[2] Many of the surgical recipients had numerous comorbidities, the most common being diabetes (19.6%), coronary artery disease (11.1%), anxiety disorders (18.5%), depression (21.6%), and arthritis (19.2%). Notably, 42.5% of participants scored an ASA grade of three or higher (Table 1).

Construct Validity-Exploratory Factor Analysis (EFA) and Reliability-Internal Consistency

The KMO was.899 suggesting strong sample adequacy. The Bartlett test for sphericity was also significant. The anti-image correlation found no values below 0.500, suggesting no need to remove items. We ran the factor analysis with an Oblimin rotation initially, to determine the level of component correlation. The component correlation matrix demonstrated low levels of association (< 0.40), which allowed us to use an orthogonal rotation (Varimax). The initial exploratory factor analysis with Varimax rotation delineated a two-factor structure, which explained 54.9% of the total variance. The first eigenvalue was 4.34 (43.4% of variance explained) and the second was 1.15 (11.5% of the variance explained) suggesting two factors met our a priori guidelines.

Construct Validity-Confirmatory Factor Analysis (CFA) and Internal Consistency

The CFA confirmed a two-factor structure (Table 2), with the same explanations of the total variance. Both factors consisted of five items. Factor 1, which explained the majority of the variance consisted of items 3 (lifting), 4 (walking), 6 (standing), 8 (sex life), and 9 (social life), whereas Factor 2 consisted of items 1 (pain), 2 (personal care), 5 (sitting), 7 (sleeping), and 10 (traveling). Component matrices ranged from.497 to.825. Internal reliability demonstrated "fair" associations. Factor 1 had a Cronbach alpha of 0.74, whereas Factor 2 had a Cronbach alpha of 0.77.

Criterion Validity-Concurrent Validity

Every item of the ODI, both factors, the ODI summary score, and the dichotomized ODI values (>21, >41, >61) were statistically significantly associated ($P < 0.01$) with our proxy measures of self-care, usual activities, the EQ-5D, and the pain visual analog scale. None of the associations approached 1.0 or −1.0, and most were in the range of 0.15 to 0.40. All ranged from poor to fair in designation. The strongest associations were identified between the ODI summative percentile score, and self-care and usual activities. Table 3 provides the correlational values for concurrent validity measures.

Criterion Validity-Predictive Validity

Predictive validity was evaluated for (1) hospital readmission at 30 days, (2) operating room revisit by 30 days, and (3) revision surgery by 30 days. A majority of significant relationships were found in each analysis. Table 4 represents the associations between each ODI item, factor 1 and factor 2 coefficients created from the CFA, and the total summary score and hospital readmission within 30 days, operating room visitation within 30 days, and revision surgery within 30 days. Higher scores on all items increased odds of 30-day hospitalization, with the exception of item 5, "sitting" ($P = 0.12$). Higher scores on all items increased odds of 30-day operating room visitation, with the exception of item 5, "sitting" ($P = 0.43$) and the factor 1 coefficient

TABLE 2. Factor Loading, Structural Validity Results

Item	One-Factor	Two-Factor
1 - Pain intensity		0.583
2 - Personal care (washing, dressing, *etc.*)		0.549
3 - Lifting	0.534	
4 - Walking	0.825	
5 - Sitting		0.801
6 - Standing	0.803	
7 - Sleeping		0.692
8 - Sex life (if applicable)	0.497	
9 - Social life	0.568	
10 - Traveling		0.703

TABLE 3. Correlation Matrix of ODI Items and Concurrent Measures of Quality of Life, Pain, and Function

	Function		Quality of life (EQ-5D VAS)	Pain VAS
	Self-care activities	Usual activities		
Item 1 - Pain intensity	0.236**	0.235**	−0.186**	0.361**
Item 2 - Personal care (washing, dressing etc.)	0.329**	0.244**	−0.153**	0.164**
Item 3 - Lifting	0.239**	0.263**	−0.159**	0.184**
Item 4 - Walking	0.259**	0.293**	−0.195**	0.199**
Item 5 - Sitting	0.198**	0.177**	−0.128**	0.167**
Item 6 - Standing	0.202**	0.238**	−0.156**	0.160**
Item 7 - Sleeping	0.236**	0.205**	−0.167**	0.213**
Item 8 - Sex life (if applicable)	0.258**	0.265**	−0.189**	0.212**
Item 9 - Social life	0.314**	0.370**	−0.246**	0.222**
Item 10 - Traveling	0.268**	0.281**	−0.189**	0.203**
Factor 1	0.317**	0.274**	−0.199**	0.282**
Factor 2	0.262**	0.308**	−0.202**	0.192**
ODI summary score	0.389**	0.388**	−0.276**	0.320**
ODI% of 21 or more (moderate disability)	0.210**	0.210**	−0.156**	0.193**
ODI% of 41 or more (severe disability)	0.383**	0.343**	−0.261**	0.301**
ODI% of 61 or more (''crippled'')	0.331**	0.370**	−0.230**	0.267**

*P value < 0.05.
**P value < 0.01.
VAS, Visual Analog Scale; EQ-5D VAS: EuroQol 5 Dimension, 5-Level, Visual Analogue Scale

($P = 0.17$). Higher scores on the ODI items increased odds of 30-day revision surgery on all items. For each outcome variable (1) hospital readmission at 30 days, (2) operating room revisit by 30 days, and (3) revision surgery by 30 days), the predictors of ODI scores of 21 and greater and 41 and greater exhibited the strongest odds ratios.

DISCUSSION

The ODI is a commonly used outcome measure that addresses functional disposition in many studies involving low-back pain. Past work has suggested that it is likely that the measurement properties of the ODI are different in populations with higher levels of disability.[9] This prompted us to investigate the construct validity (structural validity) reliability (internal consistency), and criterion validity (concurrent and predictive) of the ODI in a population of spine surgery recipients with a high degree of baseline disability. Indeed, 61.8% of population exhibited baseline disabilities over 41 suggesting ''severe'' disability. These values are substantially higher than those we seen who have also investigated measurement properties of the ODI. Our findings are interesting and contrast somewhat to what has been reported previously.

Unlike past studies[6–8,23] that have primarily reported a one-factor solution, we identified a two-factor solution in our EFA and CFA modeling. Our two-factor model explained 54.9% of the total variance. Factor 1 items, which explained the majority of the variance, included lifting, walking, standing, sex, and social life, whereas Factor 2 items included pain, personal care, sitting, sleeping, and traveling. Others have also reported a two-factor model with slight differences in items that made up each factor: factor 1 (personal care, lifting, walking, sex, and social) and factor 2 (pain, sleep, standing, traveling, and sitting).[24,25] The differences in findings among studies cannot be explained as clearly by the patient populations as there were similarities between the single factor and two factors groups.

Our two-factor solution provided an internal consistency of ''fair'' associations (factor 1 = 0.74, factor 2=0.77). These values fall within the ranges of what has been reported previously in single and two-factor solutions.[5] Because the internal consistency was fair, this suggests that there are potential items within the defined factors that are not correlating as strongly with each other. Nonetheless, Cronbach reliabilities exceeding 0.70 are acceptable for group comparisons, as we have completed in our study.[26]

Concurrent validity, a form of criterion validity, measures how well one test compares to other, established tests or proxy measures. The ODI is a commonly adopted measure, but we decided to measure each item, the two factors, and its summative scores, and thresholds (*e.g.*, >21, >41, >61) against four proxy measures that are captured as part of the QOD. In all our analyses, correlations were low and ranged from poor to fair. This suggests one of three things. One, the ODI items are capturing something unique to our proxy measures; hence the relatively low correlation. The significant P values are reflective of the large sample size and should not be interpreted as an effect measure. Two, our single item proxy measures of self-care and usual activities may lack validity to appropriately compare to the ODI items.

TABLE 4. Predictive Validity of 30 day Hospital Readmission

Predictor	OR (95% Confidence Interval)	*P* value	OR (95% Confidence Interval)	*P* value	OR (95% Confidence Interval)	*P* value
	Hospital Readmission in 30 days		Operating Room Visit in 30 days		Revision Surgery in 30 days	
Item 1 - Pain intensity	1.14 (1.09, 1.17)	<0.01	1.10 (1.05, 1.16)	<0.01	1.14 (1.08, 1.21)	<0.01
Item 2 - Personal care (washing, dressing etc.)	1.15 (1.11, 1.19)	<0.01	1.10 (1.05, 1.17)	<0.01	1.17 (1.10, 1.24)	<0.01
Item 3 - Lifting	1.12 (1.07, 1.16)	<0.01	1.09 (1.05, 1.14)	<0.01	1.09 (1.04, 1.15)	<0.01
Item 4 -Walking	1.24 (1.21, 1.29)	<0.01	1.21 (1.16, 1.27)	<0.01	1.12 (1.06, 1.19)	<0.01
Item 5 - Sitting	1.02 (0.99, 1.05)	0.12	1.02 (0.97, 1.06)	0.43	1.10 (1.05, 1.19)	<0.01
Item 6 - Standing	1.15 (1.12, 1.19)	<0.01	1.13 (1.08, 1.18)	<0.01	1.09 (1.03, 1.14)	<0.01
Item 7 -Sleeping	1.07 (1.04, 1.10)	<0.01	1.07 (1.02, 1.11)	<0.01	1.16 (1.11, 1.22)	<0.01
Item 8 - Sex life (if applicable)	1.09 (1.06, 1.13)	<0.01	1.09 (1.05, 1.14)	<0.01	1.09 (1.04, 1.14)	<0.01
Item 9 - Social life	1.14 (1.11, 1.18)	<0.01	1.12 (1.07, 1.17)	<0.01	1.15 (1.09, 1.21)	<0.01
Item 10 - Traveling	1.08 91.06, 1.11)	<0.01	1.05 (1.01, 1.09)	0.01	1.09 (1.04, 1.14)	<0.01
One-factor coefficient	1.06 (1.01, 1.12)	0.01	1.05 (0.97, 1.13)	0.17	1.16 (1.07, 1.27)	<0.01
Two-factor 2 coefficient	1.26 (1.19, 1.33)	<0.01	1.24 (1.14, 1.32)	<0.01	1.13 (1.04, 1.23)	<0.01
ODI summary score	1.01 (1.01, 1.02)	<0.01	1.01 (1.01, 1.01)	<0.01	1.02 (1.01, 1.02)	<0.01
ODI% of 21 or more (moderate disability)	1.54 (1.24, 1.89)	<0.01	1.41 (1.05, 1.89)	0.02	1.65 (1.14, 2.40)	<0.01
ODI% of 41 or more (severe disability)	1.49 (1.37, 1.63)	<0.01	1.49 (1.32, 1.69)	<0.01	1.67 (1.45, 1.95)	<0.01
ODI% of 61 or more (''crippled'')	1.42 (1.30, 1.54)	<0.01	1.28 (1.13, 1.46)	<0.01	1.41 (1.22, 1.64)	<0.01

This is unlikely since the correlations were smaller in the EQ-5D VAS and the VAS for back pain, despite these being well known and validated tools.[27,28] Third, the items, the ODI summative score, the thresholds, and the proxy measures, may all represent unique constructs and we should expect a lower correlation. Indeed, different constructs are recommended as a core set of low back outcomes and this may be the best explanation of the lower correlations.[29]

A unique element of our study was that we looked at each ODI item and compared that single item to a future negative event. Interestingly, a majority of the ODI items had predictive validity toward future negative consequences associated with hospital readmission, OR visitation, or revision surgery. Higher scores on each ODI item were associated with increased odds of future negative consequences. Odds ratios were low but were likely associated with the way the ODI items are coded (0–5). With ordered values (0–5) each incremental increase (*e.g.*, from 2 to 3 or 3 to 4) is associated with a small increase in odds ratio. Dichotomizing the items would have likely led to higher odds ratios but the sensitivity of the finding would have been affected and our confidence intervals would be less precise. Further, we are unaware of any boundaries in creating a threshold to dichotomize each item for the ODI. Our interpretation is that the majority of the items of the ODI are predictive (by themselves) for future negative events, which further supports the criterion validity of the tool.

Limitations

The retrospective use of large repository data introduces limitations that have been outlined previously.[30] We instructed our statistical software to skip missing values, which we felt was the most appropriate action but is a limitation nonetheless. We did elect to use proxy measures that were not validated. Whether the instructional methods used to capture these items was consistent across all capture sites is unknown.

CONCLUSION

The ODI demonstrated a two-factor structural solution, fair internal consistency, and good criterion validity (both concurrent and predictive validity). The individual items of the ODI exhibited predictive and concurrent validity as well, suggesting the individual items reflect an overall construct. Our study lends value to a burgeoning repository of evidence that suggests the ODI is a useful tool for capturing outcomes in clinical practice. We recommend its continued use in clinical practice.

➢ Key Points

- ❑ The ODI has become an important clinical measure in the medical routine to evaluate disability in patients undergoing spine surgery.
- ❑ Past studies suggest that the population in which the ODI is tested may influence the measurement properties of the tool.
- ❑ The ODI demonstrated a two-factor structural solution, which explained 54.9% of the total variance.
- ❑ The ODI revealed fair internal consistency (0.74 to 0.77), fair criterion validity (concurrent), and significant findings with predictive validity ($P < 0.01$).

References

1. Stokes OM, Cole AA, Breakwell LM, et al. Do we have the right PROMs for measuring outcomes in lumbar spinal surgery?. *Eur Spine J* 2017;26:816–24.
2. Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine (Phila Pa 1976)* 2000;25:2940–52.
3. Fairbank JC, Couper J, Davies JB, et al. The Oswestry low back pain disability questionnaire. *Physiotherapy* 1980;66:271–3.
4. Fairbank J. Oswestry Disability Index (ODI) 2020. Available at: https://eprovide.mapi-trust.org/instruments/oswestry-disability-index. Accessed October 6, 2020.
5. Vianin M. Psychometric properties and clinical usefulness of the Oswestry Disability Index. *J Chiropr Med* 2008;7:161–3.
6. Gabel CP, Cuesta-Vargas A, Qian M, et al. The Oswestry Disability Index, confirmatory factor analysis in a sample of 35,263 verifies a one-factor structure but practicality issues remain. *Eur Spine J* 2017;26:2007–13.
7. Hagg O, Fritzell P, Nordwall A, et al. The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur Spine J* 2003;12:12–20.
8. van Hooff ML, Spruit M, Fairbank JC, et al. The Oswestry Disability Index (version 2.1a): validation of a Dutch language version. *Spine (Phila Pa 1976)* 2015;40:E83–90.
9. Saltychev M, Mattie R, McCormick Z, Bärlund E, Laimi K. Psychometric properties of the Oswestry Disability Index. *Int J Rehabil Res* 2017;40:202–8.
10. Sheahan PJ, Nelson-Wong EJ, Fischer SL. A review of culturally adapted versions of the Oswestry Disability Index: the adaptation process, construct validity, test-retest reliability and internal consistency. *Disabil Rehabil* 2015;37:2367–74.
11. CDCP. Measurement Properties: Validity, Reliability, and Responsiveness 2020 Available at: https://www.cdc.gov/hrqol/measurement.htm. Accessed October 6, 2020.
12. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
13. Jimenez-Avila JMA, Sanchez-Garcia O, Gonzalez-Cisneros AC. Guidelines in the decision of surgical management in spine surgery. *Cir Cir* 2019;87:299–307.
14. Willems P. Decision making in surgical treatment of chronic low back pain: the performance of prognostic tests to select patients for lumbar spinal fusion. *Acta Orthop Suppl* 2013;84:1–35.
15. Asher AL, Speroff T, Dittus RS, et al. The National Neurosurgery Quality and Outcomes Database (N2QOD): a collaborative North American outcomes registry to advance value-based spine care. *Spine (Phila Pa 1976)* 2014;39:S106–16.
16. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;12:e1001885.
17. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1147–57.
18. COSMIN. COSMIN Taxonomy of Measurement Properties 2020. Available at: https://www.cosmin.nl/tools/cosmin-taxonomy-measurement-properties/. Accessed October 6, 2020.
19. McGirt MJ, Speroff T, Dittus RS, Harrell FE Jr, Asher AL. The National Neurosurgery Quality and Outcomes Database (N2QOD): general overview and pilot-year project description. *Neurosurg Focus* 2013;34:E6; doi: 10.3171/2012.10.FOCUS12297.
20. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res* 2018;27:1159–70.
21. Jiang ZW, Li CJ, Wang L, et al. Prevention and handling of missing data in clinical trials. *Yao Xue Xue Bao* 2015;50:1402–7.
22. Chan Y. Biostatistics 104: correlational analysis. *Singapore Med J* 2003;44:614–9.
23. Eranki V, Koul K, Fagan A. Rationalization of outcome scores for low back pain: the Oswestry disability index and the low back outcome score. *ANZ J Surg* 2013;83:871–7.
24. Guermazi M, Mezghani M, Ghroubi S, et al. The Oswestry index for low back pain translated into Arabic and validated in a Arab population. *Ann Readapt Med Phys* 2005;48:1–10.
25. Tan K, Zheng M, Yang BX, et al. Validating the Oswestry Disability Index in patients with low back pain in Sichuan. *Sichuan Da Xue Xue Bao Yi Xue Ban* 2009;40:559–61.
26. Nunnally JC, Bernstein IC. *Psychometric Theory*, 3rd ed New York: McGraw-Hill; 1994.
27. Bijur PE, Silver W, Gallagher EJ. Reliability of the visual analog scale for measurement of acute pain. *Acad Emerg Med* 2001;8:1153–7.
28. Mueller B, Carreon LY, Glassman SD. Comparison of the EuroQOL-5D with the Oswestry Disability Index, back and leg pain scores in patients with degenerative lumbar spine pathology. *Spine (Phila Pa 1976)* 2013;38:757–61.
29. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain* 2018;159:481–95.
30. Pryor DB, Lee KL. Methods for the analysis and assessment of clinical databases: the clinician's perspective. *Stat Med* 1991;10:617–28.