# Validity and Reliability of a Measurement of Objective Functional Impairment in Lumbar Degenerative Disc Disease: The Timed Up and Go (TUG) Test

**Oliver P. Gautschi, MD***
**Nicolas R. Smoll, MBBS‡**
**Marco V. Corniola, MD***
**Holger Joswig, MD§**
**Ivan Chau, MD§**
**Gerhard Hildebrandt, MD§**
**Karl Schaller, MD***
**Martin N. Stienen, MD*§**

*Department of Neurosurgery and Faculty of Medicine, University Hospital Geneva, Geneva, Switzerland; ‡School of Medicine and Public Health, University of Newcastle, Newcastle, NSW, Australia; §Department of Neurosurgery, Cantonal Hospital St. Gallen, St. Gallen, Switzerland

**Correspondence:**
Nicolas R. Smoll, MBBS,
School of Medicine and Public Health,
University of Newcastle,
Newcastle, NSW, Australia.
E-mail: nrsmoll@gmail.com

**Received,** August 10, 2015.
**Accepted,** November 10, 2015.



**BACKGROUND:** There are few objective measures of functional impairment to support clinical decision making in lumbar degenerative disc disease (DDD).

**OBJECTIVE:** We present the validation (and reliability measures) of the Timed Up and Go (TUG) test.

**METHODS:** In a prospective, 2-center study, 253 consecutive patients were assessed using the TUG test. A representative cohort of 110 volunteers served as control subjects. The TUG test values were assessed for validity and reliability.

**RESULTS:** The TUG test had excellent intra- (intraclass correlation coefficient: 0.97) and interrater reliability (intraclass correlation coefficient: 0.99), with a standard error of measurement of 0.21 and 0.23 seconds, respectively. The validity of the TUG test was demonstrated by a good correlation with the Visual Analog Scale (VAS) back (Pearson's correlation coefficient [PCC]: 0.25) and VAS (PCC: 0.29) leg pain, functional impairment (Roland-Morris Disability Index [PCC: 0.38] and Oswestry Disability Index [PCC: 0.34]), as well as with health-related quality of life (Short Form-12 Mental Component Summary score [PCC: −0.25], Short Form-12 Physical Component Summary score [PCC: −0.32], and EQ-5D [PCC: −0.28]). The upper limit of "normal" was 11.52 seconds. Mild (lower than the 33rd percentile), moderate (33rd to 66th percentiles), and severe objective functional impairment (higher than the 66th percentile) as determined by the TUG test was <13.4 seconds, 13.4 to 18.4 seconds, and >18.4 seconds, respectively.

**CONCLUSION:** The TUG test is a quick, easy-to-use, valid, and reliable tool to evaluate objective functional impairment in patients with lumbar degenerative disc disease. In the clinical setting, patients scoring a TUG test time of over 12 seconds can be considered to have functional impairment.

**KEY WORDS:** Degenerative disc disease, Minimally clinically detectable change, Objective functional impairment, Objective outcome measurement, Severity index, Timed Up and Go test

*Neurosurgery 0:1–9, 2015* DOI: 10.1227/NEU.0000000000001195 www.neurosurgery-online.com

Lumbar degenerative disc disease (DDD) is a progressive disorder with degenerative changes affecting not only the intervertebral disc, but also the whole spinal motion segment.

**ABBREVIATIONS: BMI,** body mass index; **DDD,** degenerative disc disease; **HRQOL,** health-related quality of life; **ICC,** intraclass correlation; **LDH,** lumbar disc herniation; **LSS,** lumbar spinal stenosis; **ODI,** Oswestry Disability Index; **OFI,** objective functional impairment; **PCC,** Pearson's correlation coefficient; **PCS,** Physical Component Summary; **RMDI,** Roland-Morris Disability Index; **SF,** Short Form; **VAS,** visual analog scale

Progressive lumbar DDD commonly presents with pain and functional impairment and sometimes with neurological deficits in a substantial proportion of patients. Usually, lumbar DDD is more common in elderly patients and generally affects single or multiple segments of the spine. However, lumbar DDD should be distinguished from the normal aging process, which results in radiological evidence of degenerative changes, but with little or no clinical signs of disease.[1-5] Only a fraction of this population will sooner or later become symptomatic with lumbar DDD, depending on intrinsic- (eg, somatic, psychological)

as well as extrinsic factors (eg, morphological, external). Furthermore, the prevalence of lumbar DDD in developed countries is constantly increasing, mainly due to longer life expectancy.[6,7]

A number of patients with acute low back pain related to DDD will experience pain relief and regain functionality after medical treatment. However, the efficacy of noninvasive (or so-called conservative) treatment may only be temporary in some patients. In patients in whom the best medical treatment fails to alleviate symptoms, in patients suffering with symptoms and in those presenting with severe or progressive neurological deficits, a surgical intervention with decompression of the involved nervous structures (with additional stabilization of the spine, whenever necessary) has proved to be highly effective.[8,9]

The indication for surgical treatment of lumbar DDD is primarily based on pain, functional disability, and impaired health-related quality of life (HRQOL) in the presence of corresponding radiological imaging findings. Secondary factors such as age, comorbidities, symptom duration, and the expected natural course of DDD may also play a role in the decision-making process for a surgical procedure.[10] Therefore, to accurately measure pain, functional impairment and HRQOL become a matter of paramount importance when considering surgery as well as afterward when the efficacy of the treatment is evaluated and followed.[11,12] In lumbar DDD, back and leg pain is commonly measured using the visual analog scale (VAS). On the other hand, functional impairment is commonly measured using the Roland-Morris Disability Index (RMDI) or the Oswestry Disability Index (ODI).[13] Several questionnaires have been validated to assess HRQOL for lower back pain and DDD such as the Short Form (SF)-36 and its shorter version SF-12 or the EQ-5D.[11,14]

It is important to acknowledge that recording the patient's subjective perception of his or her disability carries uncertainties insofar as the patient's rating may significantly differ from the impression of the independent examiner. In the daily clinical practice, we often encounter patients reporting severe disability and pain while moving rather unrestrained and appearing less affected than their pain or disability score may suggest. Some patients report difficulties with completing the questionnaires because certain questions may not apply to their situation. There are patients with obvious functional limitations (eg, limping) who report no subjective pain or disability at all. For the reasons outlined above, we have evaluated a simple test, the Timed Up and Go (TUG) test, to aid clinical decision making by adding a more objective dimension to the assessment of functional disability in patients with lumbar DDD.[12,15]

# METHODS

## Study Group

Patients were enrolled between September 2013 and November 2014 at the Departments of Neurosurgery of the University Hospital Geneva and the Cantonal Hospital St. Gallen in Switzerland. We included patients with lumbar DDD scheduled for spine surgery for the following diagnoses: (1) lumbar disc herniation (LDH), (2) lumbar spinal stenosis (LSS), and (3) lumbar DDD with or without instability requiring lumbar fusion (transforaminal lumbar interbody fusion, posterior lumbar interbody fusion, or extreme lateral interbody fusion). Exclusion criteria were age younger than 18 years, pregnancy, severe motor deficits, or other medical reasons interfering with the patients' ability to walk and perform the TUG test (eg, osteoarthritic disease of the lower extremities, Parkinson's disease, heart failure, and hip or knee prosthesis). Patient demographic (age, sex, size, weight, body mass index [BMI], and working status)[16,17] and clinical (British Medical Research Council muscle strength of index muscles of the lower extremity, smoking status, and previous surgery) data were recorded preoperatively. In addition, VAS back and leg pain scores, RMDI and ODI questionnaires, as well as the SF-12 and EQ-5D questionnaires were determined. Immediately after the subjective assessment, patients performed a single TUG test.

## Control Group

Visitors to the hospital and patients' next of kin with no known spinal diseases were recruited as control subjects. With the exception of radiological imaging, the same information was obtained from the control group as for the study group. Control subjects performed the TUG test 3 times, and 2 study investigators measured the performance time independently to determine the test-retest reliability.

## Ethics Approval

The study was approved by the Institutional Review Board of the University of Geneva (14-079) and the Ethics Committee St. Gallen (14/049). Written informed consent was obtained from all study participants (patients and control subjects).

## The TUG Test

All investigators adhered to a common protocol when conducting the TUG test. Patients were asked to sit and lean back on an armchair with their arms resting on the armrests. On command of the examiner, time was recorded (in seconds), and the patients had to get up and walk as fast as possible (without running) to a marked line on the floor at a 3-m distance from the chair. Once they reached the line, they made a 180° turn, returned to the chair, and sat down as quickly as possible (Figure 1). Time was stopped when the participant sat down again. Patients were allowed to wear their regular shoes and use a walking aid, if required.

## Reliability Measures

Inter- and intrarater reliability were determined for the normal population (control group) to verify the independence of the cutoff point for scores indicating no objective functional impairment (OFI) (vs scores indicating OFI) by large standard errors of the mean (SEM). According to McGraw and Wong[18] and Shrout and Fleiss,[19] the intraclass correlation coefficient (ICC) measures consistency and agreement of a measurement. Excellent consistency is achieved when scores obtained from 2 independent examiners consistently differ by the same extent. Excellent agreement implies that the difference between the absolute values of the 2 examiners' scores is very small. The latter is used for the inter- and intrarater SEM calculations.

The SEM is equal to the SD multiplied by the square root of 1 minus the inter- or intrarater ICC-measured agreement, as demonstrated by Stratford and Goldsmith.[20] Here, the SEM is used to measure the gray zone at the cutoff point between patient scores indicating the presence or absence of OFI. This gray zone represents the uncertainty spectrum,

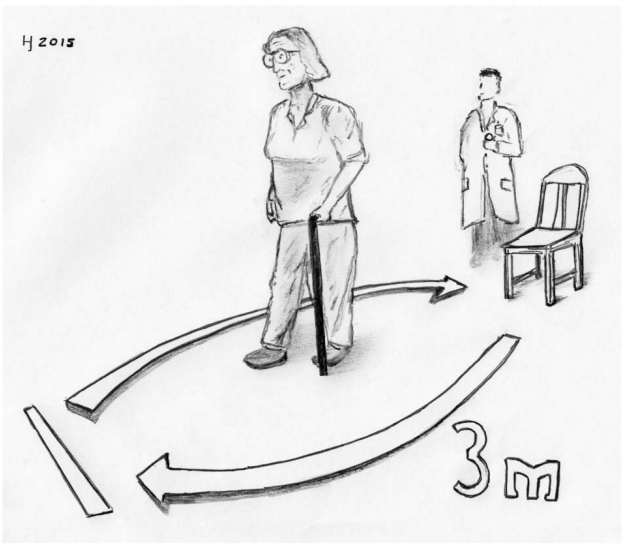


**FIGURE 1.** *Scheme of the Timed Up and Go (TUG) test.*

where clinical judgment must take place to determine whether a patient is considered to have a TUG test result that is normal or pathologic.

### *Z- and T-Scores*

Z- and T-scores express how much a patient's score deviates from normal in a standardized manner. The Z-score is the number of SDs off the mean of the normal population. The T-score is a transformation of the Z-score to a number that is easier to compare with other tests. For the TUG test, we arbitrarily determined the mean of the normal population to be 100, with an SD of 10. Thus, the 99th percentile corresponding to a Z-score of 2.3 (and a T-score of 123) is the upper limit of normal (1-sided test). In other words, any patient with a TUG test result >2.3 SDs above the normal mean plus the SEM is defined as a patient who has a TUG test result indicating OFI.

The Z-score of a patient is calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

x = observed test time; $\mu$ = normal population mean test times; $\sigma$ = normal population SD of test times.

The means and SDs of the normal population will be provided. The corresponding T-score is calculated using the following equation:

$$t = 10z + 100$$

t = T-score; z = Z-score.

### Baseline Severity Index

The baseline severity index (BSI) is a simple score that excludes normal TUG test results and stratifies test results that indicate OFI into thirds: lower than the 33rd percentile = mild functional impairment, 33rd to 66th percentiles = moderate functional impairment, and higher than the 66th percentile = severe functional impairment. This nonparametric method was chosen due to the highly skewed TUG test times in the patient population.

### Statistical Analysis

Baseline characteristics between the study group and control subjects were compared using Welch's *t* test for unequal variances for continuous variables and $\chi^2$ tests for categorical variables. The relationship between TUG test times and various subjective outcome measures was measured using Pearson's correlation coefficient (PCC) after a log transformation to ensure normally distributed TUG test times in the patient population. Data were analyzed using Stata version 13.1 (StataCorp, College Station, Texas).

## RESULTS

Two hundred fifty-three consecutive patients with DDD scheduled for lumbar spine surgery (study group) and 110 control subjects (control group) were enrolled in this study. The group of control subjects was younger compared with the study cohort and consisted of more female subjects (Table 1). The HRQOL data for both patients and control subjects closely resembled the range

**TABLE 1. Baseline Characteristics of the Study Group and the Control Group[a,b]**

| | Control Group | | Study Group | | *P* Value |
|---|---|---|---|---|---|
| Age | 52.23 | 15.61 | 58.40 | 16.14 | .0004 |
| Sex | | | | | .008 |
| Male | 47 | 43% | 146 | 57% | |
| Female | 63 | 57% | 107 | 42% | |
| Height, cm | 169.8 | 8.9 | 169.8 | 9.7 | .4842 |
| Weight, kg | 74.2 | 14.6 | 78.2 | 16.5 | .0103 |
| Smoking status | | | | | .78 |
| Nonsmokers | 82 | 75% | 192 | 76% | |
| Smokers | 28 | 25% | 61 | 24% | |
| VAS pain measures | | | | | |
| Back | 0.55 | 1.35 | 3.91 | 2.85 | <.0001 |
| Leg | 0.42 | 1.41 | 4.88 | 2.90 | <.0001 |
| Functional disability | | | | | |
| RMDI | 0.91 | 2.47 | 11.80 | 5.19 | <.0001 |
| ODI | 5.22 | 10.49 | 48.39 | 20.78 | <.0001 |
| HRQOL | | | | | |
| EQ-5D index | 0.87 | 0.18 | 0.46 | 0.23 | <.0001 |
| EQ-5D VAS | 85.39 | 16.35 | 51.13 | 21.62 | <.0001 |
| SF-12 PCS | 51.85 | 7.77 | 30.25 | 8.06 | <.0001 |
| SF-12 MCS | 49.82 | 9.44 | 43.89 | 11.30 | <.0001 |
| TUG test time, s | 6.03 | 2.36 | 11.28 | 5.83 | <.0001 |
| Total | n = 110 | 100% | n = 253 | 100% | |

[a]VAS, visual analog scale; RMDI, Roland-Morris Disability Index; ODI, Oswestry Disability Index; HRQOL, health-related quality of life; SF-12, Short Form-12; PCS, Physical Component Summary; MCS, Mental Component Summary; TUG, Time Up and Go.

[b]Continuous variables are presented as means and SDs, and categorical variables are presented as frequencies and percentages. Differences between the patient group and control group were tested using Welch's T-test for unequal variances in continuous variables and Pearson's $\chi^2$ test for categorical variables.

of previously published values for patients with lumbar DDD and the German-speaking population norm, respectively.[21,22] The mean TUG test time for the study group was 11.3 seconds (SD 5.8) compared with 6.0 seconds (SD 2.3) for the control group ($P < .0001$). A histogram of the TUG test results of both groups is depicted in Figure 2. The TUG test data in the normal population are expected to be a normal distribution, especially if stratified by age and sex.

## Content Validity

The logarithmically transformed TUG test times correlated directly (using PCC) with the VAS scales for back pain (0.29) and leg pain (0.30), functional impairment as determined by the RMDI questionnaire (0.43), and the ODI questionnaire (0.36) (Figures 3 and 4). It equally inversely correlated with HRQOL measured by the 2 subscores of the EQ-5D, the EQ-5D Index (−0.41), and EQ5D VAS (−0.28), as well as measured by the 2 component summary subscores of the SF-12, the SF-12 Physical Component Summary (PCS) (−0.33), and SF-12 Mental Component Summary (−0.27) scores (Figures 5-7).

## Reliability Measurements

In 110 control subjects, the TUG test was performed 3 times, and the time was measured independently by 2 examiners, resulting in a total of 660 measurements. The TUG test showed an intrarater reliability of 97% (SEM = 0.21) and interrater reliability of 99% (SEM = 0.22) (Table 2). The SEM of 0.22 (the larger of the 2 SEMs) indicates that TUG test times that are within ±1.96 SEM (or 0.43 seconds) of the cutoff value could be considered a gray zone for which no clear interpretation of normal or abnormal can be made. This is the 95% confidence interval of a test value.

## TUG Test: Upper Limit of Normal

For the group of control subjects, the upper limit of normal was calculated as 6.03 (mean) + (2.33 × 2.36 [SD]) = 11.52 seconds. Upper limits of normal, stratified by age and sex, are presented in Table 3. In this cohort, 33% of patients were found to have TUG test results indicating some degree of OFI using the overall cutoff. When stratified for age and sex, however, 39.5% of the TUG test results revealed some degree of OFI. The gray zone for the overall upper limit of normal is between 11.30 and 11.74, indicating that values in this bracket cannot clearly define patients with or without OFI. In this gray zone, the physician should look at sex- and age-specific cutoff values for more information. For clinical practice, values >11 and <12 seconds can reasonably be viewed as inconclusive, and the patient should either be retested at a later stage or other tests should be used for the measurement of functional disability.

## BSI Stratification

Because the TUG test values are right skewed, nonparametric cutoff values in the BSI are superior- to normal-based (parametric) cutoff values. Mild (lower than the 33rd percentile), moderate (33rd to 66th percentiles), and severe OFI (higher than the 66th percentile), as determined by the raw TUG test values, were determined to be

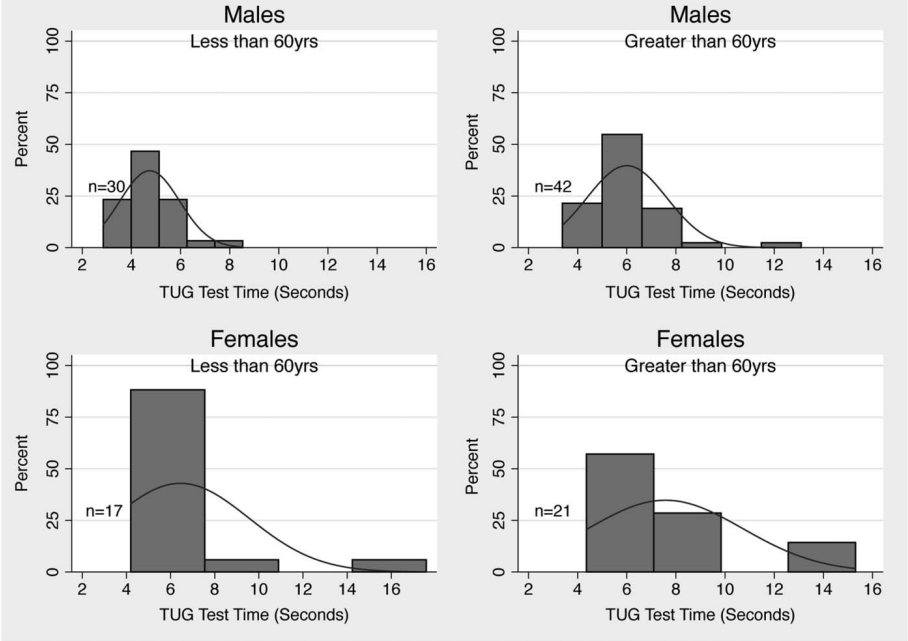


**FIGURE 2.** *Histogram of normal Timed Up and Go (TUG) test results by sex and age category.*

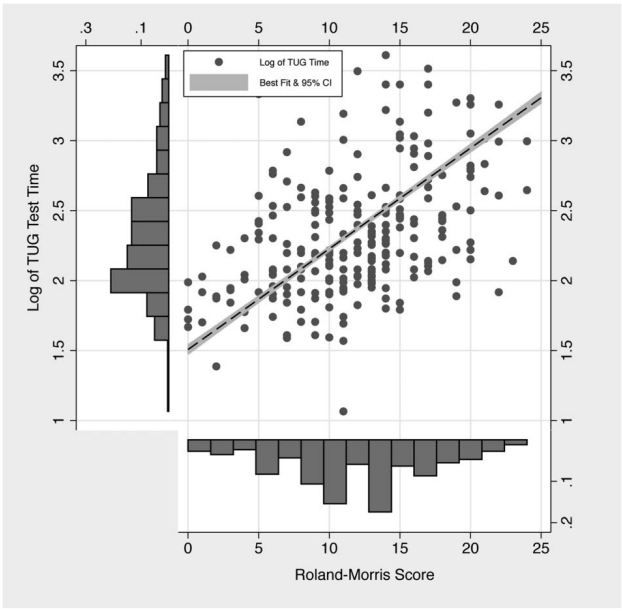


**FIGURE 3.** *Relationship between the logarithm of the Timed Up and Go (TUG) test results and the Roland-Morris Disability Index. Pearson's correlation coefficient (r) was 0.43 ($r^2$ = 0.18). CI, confidence interval.*

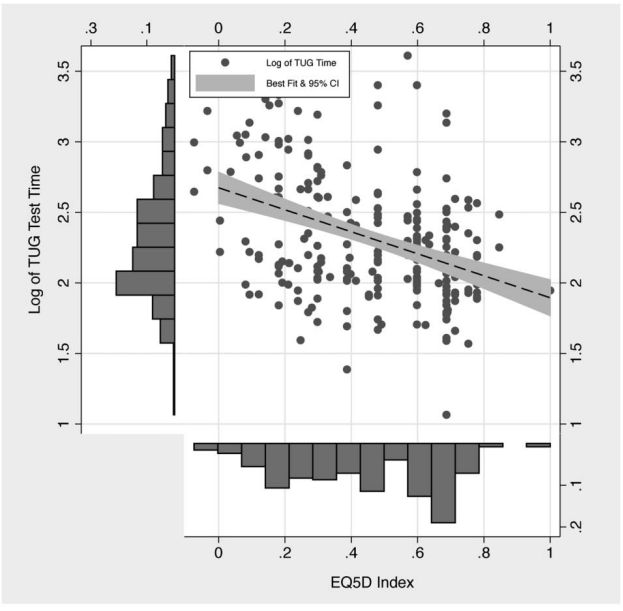


**FIGURE 5.** *Relationship between Timed Up and Go (TUG) test results and the EQ-5D index. Pearson's correlation coefficient (r) was −0.41 ($r^2$ = 0.17). CI, confidence interval.*

<13.37 seconds, 13.37 to 18.38 seconds, and >18.38 seconds for the complete cohort, respectively (see Table 4 for the BSI by age and sex). For patients with mild functional impairment, this translates to a T-score >123 (upper limit of normal) and <131.1. For patients with moderate functional impairment, this translates to a T-score of 131.1 to 152.0, and severe functional impairment for patients with

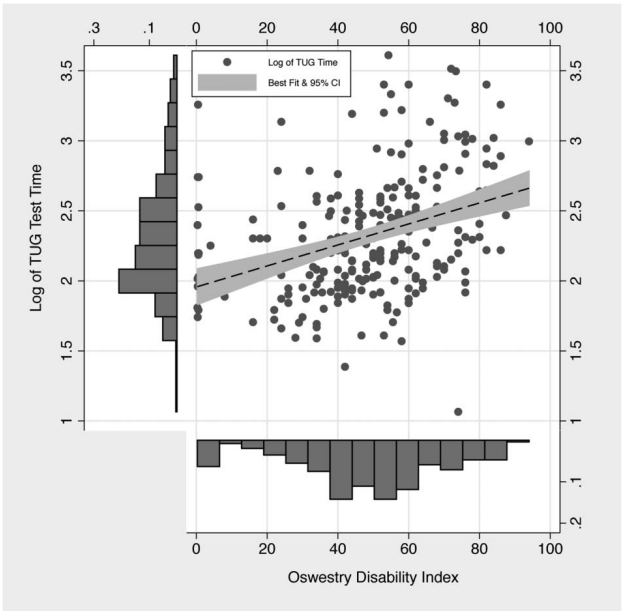


**FIGURE 4.** *Relationship between Timed Up and Go (TUG) test results and the Oswestry Disability Index. Pearson's correlation coefficient (r) was 0.36 ($r^2$ = 0.13). CI, confidence interval.*

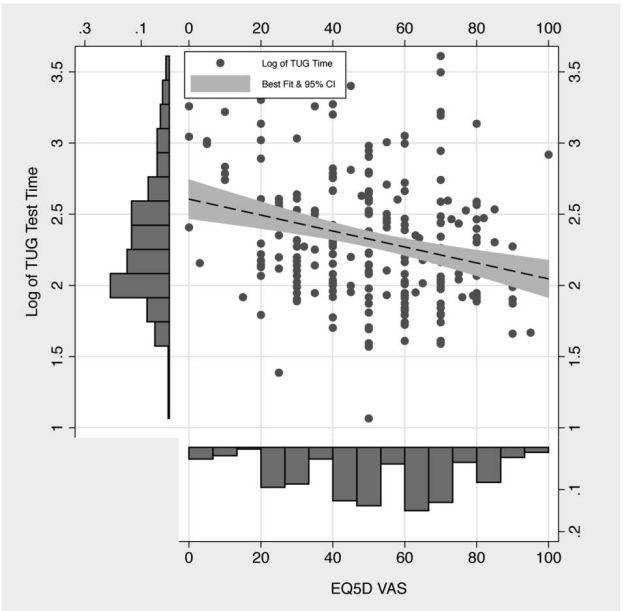


**FIGURE 6.** *Relationship between Timed Up and Go (TUG) test results and the EQ-5D Visual Analog Scale (VAS). Pearson's correlation coefficient (r) was −0.28 ($r^2$ = 0.08). CI, confidence interval.*

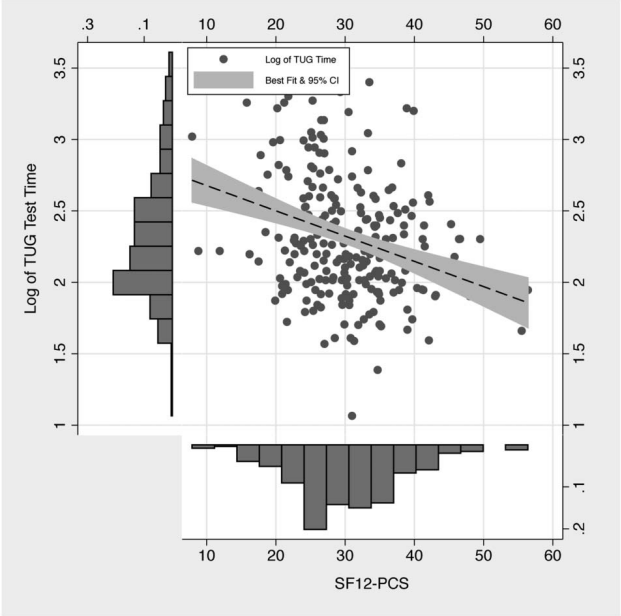


**FIGURE 7.** *Relationship between Timed Up and Go (TUG) test results and the Short Form-12 (SF-12) Physical Component Summary (PCS). Pearson's correlation coefficient (r) was −0.33 ($r^2$ = 0.11). CI, confidence interval.*

T-scores >152.0. In Figure 8, the distribution of patients with and without OFI according to the T-scores is demonstrated. In our sample, 39.5% of patients with lumbar DDD had TUG test results suggesting some degree of OFI.

## DISCUSSION

We could demonstrate that the TUG test is an objective tool with excellent intra- and interrater reliability for the assessment of OFI in patients with lumbar DDD. The upper limit of normal was determined to be 11.5 seconds, thus serving as a cutoff to discriminate between subjects who are functionally impaired. Overall, patients performing the test in <13.4 seconds, between 13.4 and 18.4 seconds, or >18.4 seconds can be considered as having none/mild, moderate, or severe OFI, respectively. Furthermore, TUG test scores were stratified according to age and sex (Table 4). Given that the SEM is 0.22 seconds, a TUG test time of 11.5 seconds has a 95% confidence interval of 11.07, and 11.93 seconds can be considered to be normal or abnormal. This is a gray zone, a zone where an overly strict interpretation of the TUG test in the *clinical* setting may result in the inappropriate classification of a patient. Here, either clinical judgment (with or without retesting using the TUG test) or other tests may be more appropriate to classify the patient accurately. *Clinically* (using the single cutoff score for all patients), patients younger than 60 years of age are expected to have TUG test scores well below the cutoff score and can be considered functionally impaired in the 11- to 12- second range. On the other hand, for patients older than 65 years of age, the range can be considered normal for that patient (especially if you go back to the age- and sex-specific cutoffs). Any patient with lumbar DDD and a TUG test score ≥22 seconds can be considered to have severe OFI with great certainty, irrespective of the patient's age or sex (Table 4).

Why does it make sense to use the TUG test for clinical patient evaluation? The TUG test reproduces a simple task that is performed by every ambulant human being on a daily basis: stand up, walk, change direction, walk again, and sit down. These are essential functions enabling patients to resume activities of daily living after spine surgery.[15] The TUG test serves as a quick and easy-to-perform tool to measure OFI in clinical daily practice. It can be performed as needed by the physician in outpatient clinics to better appreciate the magnitude of OFI of a newly presenting patient. Because the TUG test is sensitive to changes in OFI, it has great potential to become an established standardized inpatient test, which can also be performed on a regular basis by physiotherapists or nurses.[23] The TUG test may prove to be very useful in documenting clinical progress and give guidance for questions pertaining to the need for postsurgery rehabilitation or take up employment. It is language independent and can thus be used in nonnative speakers or illiterate patients who would have difficulties filling out standardized questionnaires, for example.

The 99th percentile was used to discriminate between TUG test results of control subjects and patients in the study group. Using this approach, an overall cutoff value of 11.52 seconds was calculated, and we noticed that our patient cohort comprised a significant proportion of patients without or with only mild OFI (better discrimination can be obtained by using age- and

**TABLE 2. Intra- and Interrater Measurements of Reliability and Measurement Error**

| | Intrarater | Interrater |
|---|---|---|
| Consistency | 0.97 | 0.99 |
| Absolute | 0.97 | 0.99 |
| SEM | 0.21 | 0.23 |

**TABLE 3. Normal Population Time Up and Go Test Time Results: Summary of Means, SDs, and Upper Limits of Normal, Stratified by Age and Sex**[a]

| | Male | | | Female | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | UL | Mean | SD | UL | Mean | SD | UL |
| ≤60 y | 4.75 | 1.22 | 7.58 | 6.00 | 1.63 | 9.80 | 5.48 | 1.59 | 9.18 |
| >60 y | 6.45 | 3.11 | 13.69 | 7.57 | 3.14 | 14.89 | 7.07 | 3.14 | 14.38 |
| Overall | 5.36 | 2.23 | 10.56 | 6.53 | 2.34 | 11.99 | 6.03 | 2.36 | 11.52 |

[a]UL, upper limits of normal.

**TABLE 4. Baseline Severity Stratification of Objective Functional Impairment: Classifies Groups of Patients Into Mild (Lower Than the 33rd Percentile), Moderate (Between the 33rd and 66th Percentiles), and Severe (Higher Than the 66th Percentile)[a]**

| | Male | | Female | | Overall | |
|---|---|---|---|---|---|---|
| | (33%) | (66%) | (33%) | (66%) | (33%) | (66%) |
| ≤60 y | 8.97 (135) | 11.92 (159) | 13.30 (145) | 17.22 (169) | 11.74 (139) | 15.15 (161) |
| >60 y | 15.56 (129) | 21.95 (150) | 19.00 (136) | 25.22 (156) | 18.79 (137) | 24.61 (156) |
| Overall | 12.18 (131) | 15.50 (146) | 14.32 (133) | 19.88 (157) | 13.37 (131) | 18.38 (152) |

[a]Timed Up and Go test times in seconds and T-scores in parentheses.

sex-specific cutoff values). As a matter of fact, ~60% of our patients were considered without OFI using the cutoff values shown in Table 4. The constellation of none or mild impairment in patients scheduled for a surgical procedure did not relate exclusively to the TUG test, but also to the subjective measures. For example, 89 of the patients (35%) had a VAS back pain intensity score <3, and 22 (9%) and 17 (7%) had an RMDI <5 and an ODI <10. Ten of the patients (4%) rated their health status higher than 85% when using the EQ-5D VAS and 36 (14%) rated their HRQOL, as measured by the SF-12 PCS, within 1 SD of the German normal population values (n = 2805; PCS: 49.0 ± 9.4).[21] Still, our cohort exclusively consisted of surgical candidates with a long pain history after failed trials of nonsurgical treatment.

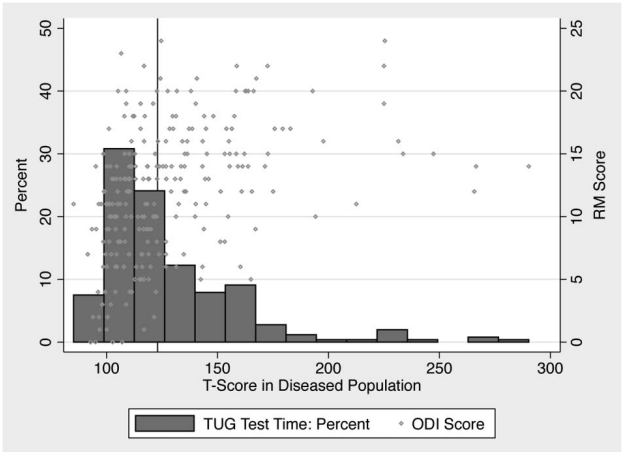


**FIGURE 8.** *Histogram and scatterplot of T-scores on the Timed Up and Go (TUG) test against the Oswestry Disability Index (ODI). Cutoff values to discriminate patients with and without objective functional impairment (OFI) were T-scores >123.3 (**vertical line**). Patients with TUG test times indicating OFI comprised 39.5% of the diseased sample. The T-score used here is equal to 100 + (10 × Z-score). The Z-score is the number of SDs off the mean or (TUG test time of interest − mean TUG test time in the control group)/(SD of TUG test times in the control group). Note that a T-score of 100 in a trial implies that the TUG test time was identical to the mean in the control group. RM, Roland-Morris.*

This brings up 2 important concepts. First, there is no single test or gold standard objective measurement of function, and each patient has an individual set of functions that is important to him/her. The TUG test merely gives a good estimate of some of the basic functions that are vital to every patient (eg, standing up, walking, and sitting down again). As such, it fulfills the purpose of distinguishing patients with some degree of OFI (>11.5 seconds) from a normal population without functional impairment that can be used in the clinical setting. Second, no single test taken alone is sufficiently predictive of the need for surgery. With a single test, the complex constellation of pain, functional disability, and reduced HRQOL in a patient seeking surgical therapy can never be fully appreciated. It is important to understand that not every patient who requires surgical therapy for DDD automatically has an OFI as determined by TUG test results >11.5 seconds. Determining the necessity for and outcome prognosis of surgery is a holistic approach taking into account clinical symptoms and a multitude of different tests together with imaging findings.

The use of T-scores facilitates the comparison of results for different populations (eg, sex, age). For example, a female 55-year-old patient (patient A) with sciatica due to an LDH at L5-S1 with a slight motor deficit performs the TUG test in 15.8 seconds. When interpreting her TUG test result in view of the age- and sex-matched cutoff values (Table 4), her OFI can be considered moderate (between the 33rd and 66th percentiles). Based on the absolute test values only, it is difficult to compare patient A with another patient B because there are different normal cutoffs for patients of different ages and sexes. Patient B is a 90-year-old male patient with LSS without motor deficit and a TUG test result of 16.2 seconds. Patient A's absolute TUG test result corresponds to a Z-score of (15.8 − 6.0 seconds)/1.63 = 6.01 and a T-score of 10 × 6.01 + 100 = 160.1 (formulas given in the Methods section). In contrast to this, patient B's TUG test result of 16.2 seconds (Z-score, 3.13; T-score, 131.3) has a similar raw value as that of patient A, but because of differing sex and age categories, the deviation of his test result from normal is less than that of patient A, hence, the lower T-score. The advantage of using the T-score thus lies in its standardized comparison of results for different patient populations.

The TUG test showed a mild to moderate correlation with established subjective measures of pain intensity, functional

disability, and HRQOL. In every case, the correlation was in the correct direction (eg, as the HRQOL decreased, the TUG test times increased). It is important to note, however, that none of the correlations exceeded 0.43 using the PCC. We interpret this as a demonstration of how the TUG test measures a different dimension of functional impairment. The TUG test results relate to all 3 dimensions (pain, functional impairment, and HRQOL), yet seem to be distinct from them. When comparing the TUG test's correlation with the 3 aspects of well-being, the correlation was higher with functional (RMDI, 0.43; ODI, 0.36) than with pain (VAS back pain score, 0.29; VAS leg pain score, 0.30) or HRQOL measures (EQ-5D Index, −0.28; SF-12 PCS, −0.33; SF-12 MCS, −0.27). The correlation coefficients indicate that the TUG test should not replace subjective measures, but rather be used to complement them and add a new dimension (objective measurement as distinct from subjective measurement) to the patient evaluation to aid in clinical decision making. Using the proposed cutoff values for OFI here, ongoing projects will aim at estimating the minimum clinically important difference of the TUG test that corresponds to clinically meaningful improvement of function.

### Limitations

Other factors may influence the performance on the TUG test that have not been specifically reported here, eg, severe obesity may lead to higher TUG test times in patients without DDD. In this cohort, the mean BMI was 27.1 ± 4.5 kg/m$^2$ and 61 patients in total were obese. TUG test times increased in those classified as grade I (48 patients with a BMI of 30-35 kg/m$^2$ and a mean TUG test time of 11.75 ± 6.4 seconds), grade II (11 patients with a BMI of 35-40 kg/m$^2$ and a mean TUG test time of 12.20 ± 6.1 seconds), and grade III according to the WHO classification (2 patients with a BMI >40 kg/m$^2$ and a mean TUG test time of 16.20 ± 5.6 seconds). The same is true for the British Medical Research Council paresis grade: it is conceivable that a greater motor deficit interferes with the performance on the TUG test. In fact, patients with a grade 3 (15 patients; mean TUG test time, 13.94 ± 9.3 seconds) or grade 4 paresis (66 patients; mean TUG test time, 12.51 ± 6.7 seconds) performed more poorly on the TUG test compared with those without a motor deficit (mean TUG test time, 10.71 ± 5.0 seconds). In addition, patients included in this study had DDD to various degrees. Although most patients underwent surgery for LDH, usually without severe degeneration of the motion segment, we also included patients with LSS, spondylolisthesis, or more complex patterns of DDD. The definition of OFI for a given disease (eg, LDH or LSS) could be more accurate than for DDD in general, as proposed here. Given the short distance required to ambulate, it has been reported previously that patients with LSS may have less difficulty performing the TUG test than patients with sciatica.[23] For the current work, however, we chose only age and sex to detail on cutoff values and abstained from making the analysis more complex or splitting the patient sample into smaller disease-specific cohorts. Future work will focus on the mentioned limitations and give more accurate estimates for certain patient groups (eg, young vs old; male vs female) or conditions (eg, obesity, paresis grades, the specific underlying disease). Overall, the current data are adequate to estimate OFI in most patients with DDD that clinicians face in clinical daily routine. We encourage other groups to investigate and revalidate the cutoffs and the BSI described here.

## CONCLUSION

We found the TUG test to be a valid and reliable measure of OFI. In clinical practice, the upper limit of normal, which can be used as the overall population cutoff value, is 11.5 seconds, distinguishing patients from those without OFI. TUG test results greater than 13.4 and 18.4 seconds imply that the patient has moderate or severe OFI, respectively. The use of age- and sex-specific standardized cutoff values, as presented in Table 4, allow for a more accurate evaluation of OFI.

### Disclosure

No funding was received for this study. Expenditures (questionnaires, study nurse) were paid for by the Department of Neurosurgery, Cantonal Hospital St. Gallen. The authors have no personal, financial, or institutional interest in any of the drugs, materials, or devices described in this article.

## REFERENCES

1. Prescher A. Anatomy and pathology of the aging spine. *Eur J Radiol.* 1998;27(3):181-195.
2. Sether LA, Yu S, Haughton VM, Fischer ME. Intervertebral disk: normal age-related changes in MR signal intensity. *Radiology*. 1990;177(2):385-388.
3. Berlemann U, Gries NC, Moore RJ. The relationship between height, shape and histological changes in early degeneration of the lower lumbar discs. *Eur Spine J.* 1998;7(3):212-217.
4. Pearce RH, Grimmer BJ, Adams ME. Degeneration and the chemical composition of the human lumbar intervertebral disc. *J Orthop Res.* 1987;5(2):198-205.
5. Pritzker KP. Aging and degeneration in the lumbar intervertebral disc. *Orthop Clin North Am.* 1977;8(1):66-77.
6. Taylor JA, Bussieres A. Diagnostic imaging for spinal disorders in the elderly: a narrative review. *Chiropr Man Therap.* 2012;20(1):16.
7. Hartvigsen J, Frederiksen H, Christensen K. Back and neck pain in seniors-prevalence and impact. *Eur Spine J.* 2006;15(6):802-806.
8. Weinstein JN, Tosteson TD, Lurie JD, et al. Surgical versus nonsurgical therapy for lumbar spinal stenosis. *N Engl J Med.* 2008;358(8):794-810.
9. Weinstein JN, Tosteson TD, Lurie JD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. *JAMA*. 2006;296(20):2441-2450.
10. Martin BI, Lurie JD, Tosteson AN, et al. Indications for spine surgery: validation of an administrative coding algorithm to classify degenerative diagnoses. *Spine*. 2014;39(9):769-779.
11. DeVine J, Norvell DC, Ecker E, et al. Evaluating the correlation and responsiveness of patient-reported pain with function and quality-of-life outcomes after spine surgery. *Spine*. 2011;36(21 suppl):S69-S74.
12. Stienen MN, Smoll NR, Hildebrandt G, Schaller K, Gautschi OP. Influence of smoking status at time of surgery for herniated lumbar disk on postoperative pain and health-related quality of life. *Clin Neurol Neurosurg.* 2014;122:12-19.
13. Roland M, Fairbank J. The Roland-Morris disability questionnaire and the Oswestry Disability Questionnaire. *Spine*. 2000;25(24):3115-3124.
14. Zanoli G. Outcome assessment in lumbar spine surgery. *Acta Orthop Suppl.* 2005;76(318):5-47.

15. Gautschi OP, Corniola MV, Schaller K, Smoll NR, Stienen MN. The need for an objective outcome measurement in spine surgery-the timed-up-and-go test. *Spine J.* 2014;14(10):2521-2522.
16. Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. *Radiology*. 1988;166(1 pt 1):193-199.
17. Pfirrmann CW, Metzdorf A, Zanetti M, Hodler J, Boos N. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine*. 2001;26(17): 1873-1878.
18. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods.* 1996;1:30-46.
19. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420-428.
20. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther.* 1997;77 (7):745-750.
21. Morfeld M, Kirchberger I, Bullinger M. *SF-36. Fragebogen zum Gesundheitszustand. Deutsche Version des Short Form-36 Health Survey. 2. Ergänzte und Überarbeitete Version*. Göttingen, Bern: Hogrefe Verlag; 2011.
22. Stienen MN, Smoll NR, Hildebrandt G, Schaller K, Gautschi OP. Early surgical education of residents is safe for microscopic lumbar disc surgery. *Acta Neurochir (Wien).* 2014;156(6):1205-1214.
23. Gautschi OP, Corniola MV, Joswig H, et al. The timed up and go test for lumbar degenerative disc disease. *J Clin Neurosci.* 2015;22(12):1943-1948.

## Acknowledgments

The authors thank Cornelia Lüthi (a study nurse in the Department of Surgery, Cantonal Hospital St. Gallen) and Dario Jucker (a medical student at Zurich University) for their important contribution in the data collection. They also thank all of the patients and control subjects who agreed to take part in this study.