Özden *Egypt J Neurol Psychiatry Neurosurg* (2024) 60:25
https://doi.org/10.1186/s41983-024-00805-z

The Egyptian Journal of Neurology, Psychiatry and Neurosurgery

REVIEW

Open Access

The reliability and validity of the Timed Up and Go test in patients ongoing or following lumbar spine surgery: a systematic review and meta-analysis

Fatih Özden[1*]

Abstract

Background No other systematic review examined the measurement properties of the TUG in LSS. The present systematic review and meta-analysis aimed to investigate the measurement properties of the Timed Up and Go (TUG) in patients with Lumbar Spine Surgery (LSS). A literature search yielded 906 studies [PubMed:71, Web of Science (WoS):80, Scopus:214, ScienceDirect:471 and Cochrane Library:70]. Included 10 studies were assessed for risk of bias and quality using the "four-point COSMIN tool" and "COSMIN quality criteria tool". Criterion validity and responsiveness results were pooled with "correlation coefficient" and "Hedges' g" based effect size, respectively.

Results The correlation coefficient pooling between TUG and VAS back and leg pain was 0.26 (moderate) (95% CI 0.19–0.34) and 0.28 (moderate) (95% CI 0.20–0.36). The pooled coefficient of TUG with ODI and RMDI was 0.33 (moderate) (95% CI 0.27–0.39) and 0.33 (moderate) (95% CI 0.24–0.42), respectively. Besides, TUG has correlated with the quality-of-life PROMs with a coefficient of −0.22 to −0.26 (moderate) (EQ5D Index 95% CI −0.35 to −0.16), (SF12-PCS 95% CI −0.33 to −0.15) and (SF12-MCS 95% CI −0.32 to −0.13). The pooled coefficient of TUG with COMI, ZCQ-PF and ZCQ-SS was 0.46 (moderate) (95% CI 0.30–0.59), 0.43 (moderate) (95% CI 0.26–0.56), and 0.38 (moderate) (95% CI 0.21–0.52), respectively. TUG's 3-day and 6-week responsiveness results were 0.14 (low) (95% CI −0.02 to 0.29) and 0.74 (moderate to strong) (95% CI 0.60–0.89), respectively. TUG was responsive at the mid-term (6 weeks) follow-up.

Conclusion In clinical practice, the TUG can be used as a reliable, valid and responsive tool to assess LSS patients' general status, especially in mid-term.

Keywords Decompression surgery, Fusion surgery, Physical function, Psychometrics, TUG

Introduction

Assessment of pain, range of motion, function, quality of life, and psychosocial status before and after lumbar spine surgery (LSS) is essential to monitor the success of surgery and rehabilitation [1, 2]. Function evaluation is mainly evaluated with physical performance tests or patient-reported outcome measures (PROMs) [3]. PROMs are valuable for evaluating subjective patient opinions [4]. In particular, the functional status of patients before and after surgery and the assessment of personal difficulty-ease improvements in activities of daily living can be evaluated practically and cost-effectively with questionnaires [5]. However, physical performance tests are used as a gold standard measurement

*Correspondence:
Fatih Özden
fatihozden@mu.edu.tr
[1] Health Care Department, Köyceğiz Vocational School of Health Services, Muğla Sıtkı Koçman University, 48800 Köyceğiz, Muğla, Turkey

method to observe the objective performance-based functions of individuals [6, 7].

Various physical performance tests containing daily life tasks (gait, sit to stand, turns, steps, stair ascent and descent, straight leg raising, squat) are developed within standardized protocols, and their measurement properties are proven in clinical studies [3, 8]. Since the essence of pain and functional advancements before and after LSS surgery is known, functional improvements of individuals are objectively evaluated with performance tests [9]. One of the most preferred tests in individuals with LSS is Timed Up and Go (TUG). TUG is a practical assessment tool including sit-to-stand, gait, and 180-degree turnaround tasks without requiring expensive equipment [10].

LSS patients have rehabilitated to be independent during the activities of daily living in the post-operative period [11, 12]. Holistic exercise programs, including strengthening, endurance, balance, core stabilization, proprioception and aerobic exercises, provide essential recovery during the post-operative period [13, 14]. Studies demonstrated the improvements in sit-to-stand and gait speed in individuals with LSS regarding lower extremity strength and endurance progress [15, 16]. Patients' somatosensorial parameters, including balance and proprioception, also improve during the turn tasks of walking. Therefore, the TUG test is a significant physical indicator assessment of patients before and after LSS [10, 17].

In 2016, Gautschi and colleagues proved the reliability of TUG in LSS with a high intraclass correlation coefficient (ICC) (0.95–0.97) [10]. Current studies have also extensively addressed the validity of the TUG with a comparison of pain, function and quality of life outcomes [3, 10, 18–21]. Furthermore, TUG was analyzed regarding responsiveness before and after surgery with short, medium and long-term follow-up results [3, 18–20, 22–25]. In addition, studies also proved minimal clinically important difference (MCID), standard error of measurement (SEM), standardized response mean (SMR) and minimal important change (MIC) values with the scope of measurement error of TUG [3, 18–20, 24, 25].

Measurement properties are essential to reveal whether physical performance tests provide accurate measurement responses in the relevant case group [26]. In addition, considering the different types of surgery (fusion, decompression, instrumentation), intervention methods (minimally invasive, conventional methods), patient follow-up duration (immediate, acute, mid-term, chronic) and differences in statistical methods (reliability, validity, responsiveness), it is essential to review whether TUG provides consistent results in individuals with LSS [13, 14, 26]. No other systematic review examined the measurement properties of the TUG in LSS. The present systematic review and meta-analysis aimed to investigate TUG's measurement properties (including criterion validity, responsiveness, measurement error and reliability) in patients with LSS.

Materials and methods

Search strategy and selection criteria

The recommendations and guidelines of the "Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)" [27], the "COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN)" [26], and the "Cochrane recommendations for systematic reviews" were followed in conducting this systematic review and meta-analysis [28]. The literature was searched with the relevant keywords (combination of boolean operators: "AND, OR") ["Lumbar Surgery" AND "Timed Up and Go Test"; "Lumbar Degenerative Disease" AND "Timed Up and Go Test"; "Lumbar Fusion" AND "Timed Up and Go Test"; "Lumbar Decompression" AND "Timed Up and Go Test"; "Lumbar AND Timed Up and Go Test"] between October 2022 and December 2022. A total of 906 studies [PubMed:71, Web of Science (WoS):80, Scopus:214, ScienceDirect:471 and Cochrane Library:70] were obtained. Details of the search are presented in Additional file 1: Appendix S1.

Eligibility criteria

The inclusion criteria of the review were; (1) studies including patients before or after LSS, (2) studies including the intervention of decompression surgery with or without fusion, (3) cohort or cross-sectional studies to provide an analysis of measurement properties (validity, reliability, measurement error, responsiveness). The exclusion criteria of the review were (1) studies with an external aim than TUG clinometric, (2) studies without primary details of measurement properties of TUG, (3) non-English studies, and (4) studies without full-text available.

Study selection and data extraction

The data files of the obtained studies (906) were transferred to Rayyan (Rayyan Systems Inc., USA) software via endnote (Clarivate Analytics, USA) outputs. Rayyan is a systematic review screening software to detect irrelevant or duplicate studies [29]. During the screening process, two expert academicians independently searched the studies' topic (title, abstract and keywords) and checked the "include, exclude or maybe" options. In cases where consensus could not be reached in the choices of two academicians, the decisive opinion of a third colleague was obtained. As a result of this initial screening, a total of 18 studies were acquired. Eight studies were excluded for the reasons as follows: (5 studies) did not provide measurement properties, (2 studies) had no full-text

available, and (1 study) did not provide specific values of measurement properties. A total of 10 studies were included in the systematic review and meta-analysis (Fig. 1). Descriptive information about the studies (year, study type, study population, follow-up period, number of cases, age, gender, surgery, diagnosis, and outcome measures) is presented in Table 1.

Risk of bias and quality assessment

The "COSMIN" tools were used for risk of bias and quality analysis. Included 10 studies were assessed for risk of bias and quality using the "four-point COSMIN tool" [26]. This tool classifies the studies as "poor, fair, good and excellent" by considering the sample size of the measurement characteristics, statistical method, and methodological deficiencies regarding possible bias. In addition, qualitative analysis of methodological design was classified with the "COSMIN quality criteria tool" [30]. This instrument classified the studies according to their primary methodological features and resulted in positive (+), indeterminate (?), negative (−) scores, and (0) no information categories. Both instruments scored the criterion validity, responsiveness and other measurement characteristics (if any) of the studies. Two independent expert academicians rated the risk of bias and quality of the included studies.

Evidence synthesis

Measurement properties of the studies with heterogenous data were presented by narrative/qualitative synthesis. These studies' results are also presented in Table 2 with the outcomes of the numerical data. Qualitative

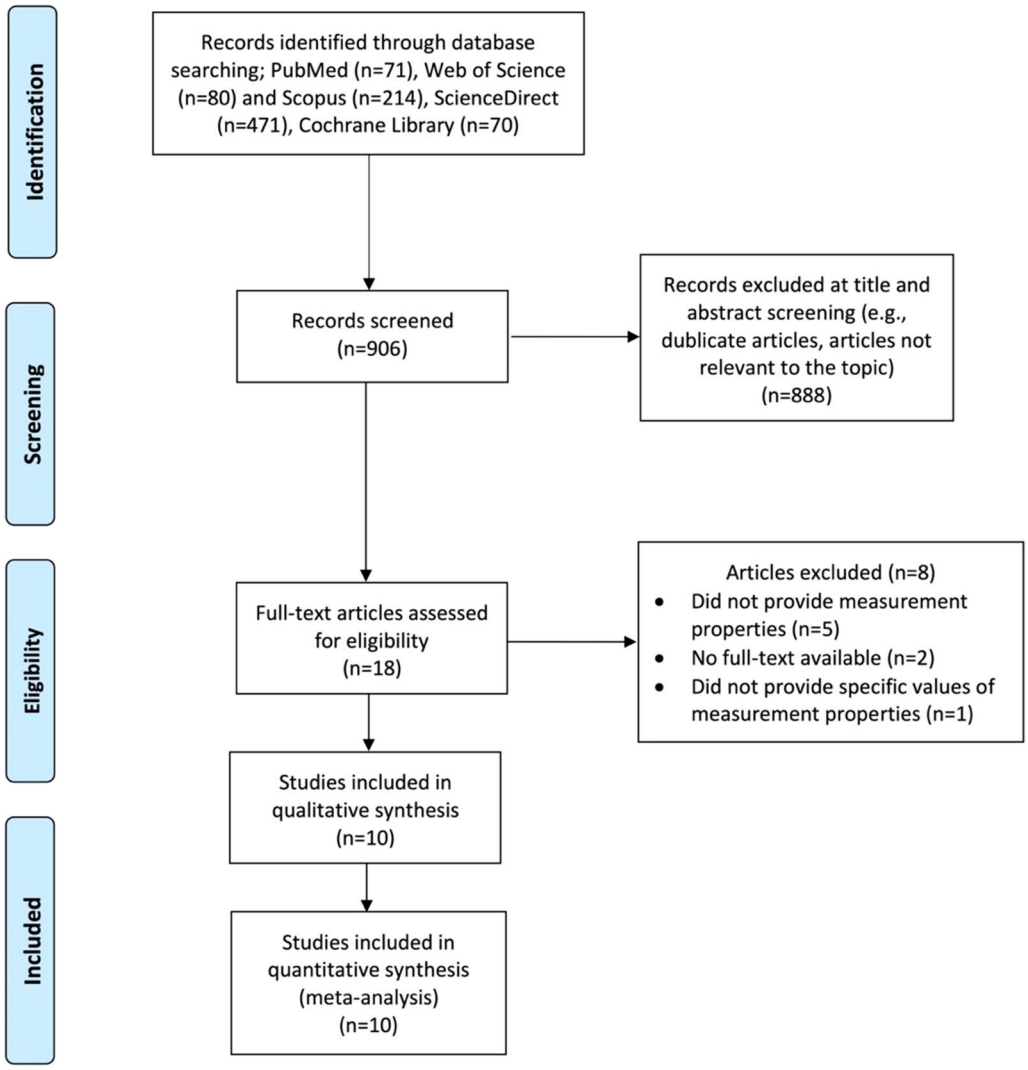

Fig. 1 PRISMA flow diagram of the study

Table 1 The characteristic overview of the studies

Author	Year	Study type	Study population	Follow-up	n	Age (years)	Gender (%)	Surgery/ diagnosis	Validation tests
Gautschi et al.	2015	Cohort	Patients ongoing and following lumbar LSS	6 weeks	30	56.6	43.3% women	Lumbar spinal stenosis Lumbar disc herniation Degenerative disc disease	TUG, VAS, RMDI, ODI, SF-12, EQ5D
Gautschi et al.	2016	Cohort	Patients ongoing lumbar LSS	n/a	253	58.4	42.0% women	Lumbar spinal stenosis Lumbar disc herniation Degenerative disc disease	TUG, VAS, RMDI, ODI, SF-12, EQ5D
Gautschi et al.	2016	Cohort	Patients ongoing and following lumbar LSS	6 weeks	136	57.7	55.9% women	Microdiscectomy Decompression Fusion surgery	TUG, VAS, RMDI, ODI, SF-12, EQ5D
Gautschi et al.	2017	Cohort	Patients ongoing and following lumbar LSS	6 weeks	100	46.2	43% women	Microdiscectomy Decompression Fusion surgery	TUG, VAS, RMDI, ODI, SF-12, EQ5D
Jakobsson et al.	2020	Clinimetric	Patients ongoing and following lumbar LFS	6 months	118	46.5	54.8% women	Fusion surgery	TUG, 5-MWT, 1-MSCT, 50-FTWT, ODI, VAS, HADS, TSK, PCS
Stienen et al.	2019	Cohort	Patients ongoing LDS with or without fusion	6 weeks	64	66.8	50% women	Decompression with or without fusion	TUG, OFI, VAS, RMDI, ODI, SF-12, EQ5D
Master et al.	2020	Trial	Patients ongoing and following lumbar LSS	12 months	248	62.2	50.8% women	Laminectomy with or without fusion	TUG, 5-STS, ODI, BPI
Maldaner et al.	2021	Cohort	Patients ongoing and following lumbar LSS	6 weeks	49	55.5	41% women	LSS with or without fusion	TUG, VAS, ZCQ, COMI
Maldaner et al.	2021	Cohort	Patients ongoing and following lumbar LSS	6 weeks	49	55.5	41% women	LSS with or without fusion	TUG, 6-MWT, VAS, ZCQ, COMI
Stienen et al.	2021	Cohort	Patients ongoing lumbar LSS	n/a	70	55.9	38.6% women	Degenerative disc disease	TUG, VAS, ZCQ, COMI

n number of participants, *n/a* not available, *LSS* lumbar spine surgery, *LFS* lumbar fusion surgery, *LDS* lumbar decompression surgery, *TUG* Timed Up and Go, *VAS* visual analog scale, *ODI* Oswestry Disability Index, *RMDQ* Roland Morris Disability Questionnaire, *SF-12* short form-12, *EQ5D* EuroQoL 5 dimension, *COMI* core outcome measures index, *ZCQ* Zurich Claudication Questionnaire, *6-MWT* 6-meter walk test, *BPI* brief pain inventory, *5-MWT* 5-minute walk test, *1-MSC* 1-minute stair climbing, *50-FTWT* 50-foot walk test, *TSK* Tampa Scale of Kinesiophobia, *HADS* Hospital Anxiety and Depression Scale

synthesis was performed through three steps: "pre-synthesis, exploring the relationships within and between the experiments, and evaluating the synthesis's robustness" [31]. The results of the synthesis are also detailed in "Results" section.

Meta-analysis (quantitative analysis of studies)

Meta-Mar software (Philipps-Universität Marburg, Germany) was used to meta-analyze the included studies [32]. The results of criterion validity and responsiveness of homogeneous data were pooled in the meta-analysis with "correlation coefficient" and "Hedges' g" based effect size, respectively. In correlation pooling, correlation coefficients of TUG with Visual Analog Scale (VAS) based back pain and leg pain, Oswestry Disability Index (ODI), Roland Morris Disability Questionnaire (RMDQ), EuroQoL 5 Dimension (EQ5D) index score, Short Form-12 (SF-12), Core Outcome Measures Index (COMI), and Zurich Claudication Questionnaire (ZCQ) were used. In responsiveness pooling, the mean change, standard deviation (SD) of the changed score, and Standardized Mean Difference (SMD) for sample sizes were calculated for two separate follow-up periods: pre-op to 3 days and pre-op to 6 weeks. The Cochrane handbook guidelines were used to determine the undefined SD of studies. "SMD, confidence interval (CI), weighted mean effect size and p-value of each pooled score" are given. "I^2, Tau^2 and Chi^2" values described the heterogeneity

Table 2 The results and COSMIN scores of the studies

Author	Year	Validity		Responsiveness		Other (if any)	
		Results	COSMIN score	Results	COSMIN score	Results	COSMIN score
Gautschi et al.	2015	n/a	n/a	Baseline: 12.0 ± 6.6 3rd day: 9.4 ± 4.8 6th week: 6.6 ± 5.5	Fair	n/a	n/a
Gautschi et al.	2016	TUG-VAS (back pain): 0.25 TUG-VAS (leg pain): 0.29 TUG-ODI: 0.34 TUG-RMDI: 0.38 TUG-EQD5 (index): − 0.28 TUG-SF12 (PCS): − 0.25 TUG-SF12 (MCS): − 0.32	Good	n/a	n/a	*Reliability* ICC (intrarater): 0.97 ICC (interrater): 0.99 *Measurement error* SEM (intrarater): 0.21 s SEM (interrater): 0.23 s	*Reliability* Fair *Measurement error* Poor
Gautschi et al.	2016	TUG-VAS (back pain): 0.19 TUG-VAS (leg pain): 0.18 TUG-ODI: 0.32 TUG-RMDI: 0.13 TUG-EQD5 (index): − 0.22 TUG-SF12 (PCS): − 0.09 TUG-SF12 (MCS): − 0.17	Good	Baseline: 10.3 ± 6.3 3rd day: 9.5 ± 4.3 6th week: 6.5 ± 2.8	Good	n/a	n/a
Gautschi et al.	2017	n/a	n/a	Baseline: 10.1 ± 5.0 3rd day: 9.4 ± 4.5 6th week: 6.6 ± 3.2	Poor	*Measurement error* MCID: 3.4 s	Good
Stienen et al.	2019	n/a	n/a	Baseline: 10.2 ± 5.5 3rd day: 10.4 ± 5.4 6th week: 7.2 ± 3.9	Good	n/a	n/a
Jakobsson et al.	2020	TUG-VAS (back pain): 0.28 TUG-ODI: 0.41 TUG-5MWT: − 0.58 TUG-1MST: − 0.67 TUG-50FWT: 0.66	Excellent	Baseline: 9.1 ± 4.4 6th month: 5.7 ± 1.1 (n: 31 subgroup)	Good	*Measurement error* MIC (95% CI) − 17.6% (− 20.7 to − 10.2)	Good
Master et al.	2020	*Pre-operative* TUG-BPI (back pain): 0.06 TUG-BPI (leg pain): 0.006 TUG-ODI: 0.29 *12th month* TUG-BPI (back pain): 0.15 TUG-BPI (leg pain): 0.11 TUG-ODI: 0.22	Excellent	Baseline: 15.5 ± 8.1 12th month: 10.6 ± 5.1	Excellent	*Measurement error* MCID: 1.3 s	Excellent
Maldaner et al.	2021	n/a	n/a	Baseline: 10.4 ± 4.3 6th week: 8.4 ± 3.3	Fair	*Measurement error* MCID: 0.9 to 3 s	Fair
Maldaner et al.	2021	TUG-VAS (back pain): 0.35 TUG-VAS (leg pain): 0.36 TUG-COMI: 0.40 TUG-ZCQ (PF): 0.45 TUG-ZCQ (SS): 0.40 TUG-ZCQ (PS): 0.38	Fair	Baseline: 10.4 ± 4.3 6th week: 8.4 ± 3.3	Fair	*Measurement error* SRM: 0.67	Fair
Stienen et al.	2021	TUG-VAS (back pain): 0.37 TUG-VAS (leg pain): 0.37 TUG-COMI: 0.50 TUG-ZCQ (PF): 0.41 TUG-ZCQ (SS): 0.36	Good	n/a	n/a	n/a	n/a

n/a not available, *TUG* Timed Up and Go, *VAS* Visual Analog Scale, *ODI* Oswestry Disability Index, *RMDQ* Roland Morris Disability Questionnaire, *SF-12 (PCS)* short form-12 (physical components summary), *SF-12 (MCS)* short form-12 (mental components summary), *EQ5D* EuroQoL 5 dimension, *COMI* core outcome measures index, *ZCQ (PF)* Zurich Claudication Questionnaire (physical function), *ZCQ (SS)* Zurich Claudication Questionnaire (symptom severity), *ZCQ (PS)* Zurich Claudication Questionnaire (patient satisfaction), *ZCQ (NIS)* Zurich Claudication Questionnaire, *6-MWT* 6-meter walk test, *BPI* brief pain inventory, *5-MWT* 5-minute walk test, *1-MSC* 1-minute stair climbing, *50-FTWT* 50-foot walk test, *ICC* intraclass correlation coefficient, *MCID* minimal clinically important difference, *SEM* standard error of measurement, *SMR* standardized response mean, *MIC* minimal important change

of the calculations. Forest plots of the results were also provided. The interpretation of effect sizes, as stated by Cohen, was considered for the correlation coefficient (r); 0.10: small, 0.30, medium and 0.50: large; for the coefficient in the responsiveness analysis (d); 020: small, 0.50: medium and 0.80: large [33].

Results

Study characteristics

The median age of the 1117 individuals in the ten studies included in the systematic review (Fig. 1) was 56.25 years (25th–75th percentile: 53.25–59.35) [3, 10, 18–25]. Eight studies had cohort design [10, 18, 19, 21–25], the other two were clinometric [20] and the secondary results of a randomized controlled trial [3]. The studies were conducted between 2015 and 2021 [3, 10, 18–25]. In 8 studies, patients were evaluated in the pre-op and post-op periods [3, 18–20, 22–25]; in 2 studies, degenerative disc disease patients were evaluated only in the pre-op period [10, 21]. The follow-up periods of the patients were a minimum of three days (immediate-term follow-up) and a maximum of 12 months (long-term follow-up) [3, 10, 18–25]. In 6 studies, male cases were more prevalent [10, 18, 21, 22, 24, 25]. Studies applied LSS intervention (laminectomy, microdiscectomy) with or without lumbar fusion surgery (instrumentation) [3, 10, 18–25]. In addition to the TUG assessment, VAS (9 studies), ODI (7 studies), RMDI (5 studies), SF-12 (5 studies), EQ5D (5 studies), COMI (3 studies), ZCQ (3 studies), and one each of 6-Meter Walk Test (6-MWT), Brief Pain Inventory (BPI), 5-Minute Walk Test (5-MWT), 1-Minute Stair Climbing (1-MSC) climbing, 50-FTWT, Tampa Scale of Kinesiophobia (TSK) and Hospital Anxiety and Depression Scale (HADS) assessments were used to evaluate the patients) [3, 10, 18–25] (Table 1).

Quality assessment and evidence level

Within the scope of criterion validity, three studies had "good" [10, 19, 21], two studies had "excellent" [3, 20], and 1 study had "fair" [18] quality. Within the scope of responsiveness, three studies had "good" [19, 20, 23], other three studies had "fair" [18, 22, 24], one study had "excellent" [3] and the other one had "poor" [25] class quality. Regarding measurement error, 2 of the six studies were classified as "fair" [18, 24], two were "good" [20, 25], one was "excellent" [3], and the other one was "poor" [10]. Regarding reliability, there was only one "fair" quality study [10] (Table 2).

Quantitative quality assessment results

Most studies (6 studies) rated the "(–) negative" [3, 10, 18–21] class for criterion validity. Four studies did not address validity [22–25]. Four studies were categorized in "(0) no information" for responsiveness [3, 19, 23, 25]. Three studies were categorized as "(?) indeterminate" [18, 22, 24], and two studies did not address responsiveness [10, 21]. Of the five studies that measured measurement error, two were "(?) indeterminate" [18, 24], the other two were "(0) no information" [3, 25], and 1 had a "(+) positive" rating [20]. Only 1 study analyzed reliability and received a "(+) positive" rating [10] (Table 3).

Table 3 Evidence level of the studies

Author	Year	Criteron validity	Responsiveness	Other (if any)
Gautschi et al.	2015	n/a	(?)	n/a
Gautschi et al.	2016	(–)	n/a	Reliability: (+)
Gautschi et al.	2016	(–)	(0)	n/a
Gautschi et al.	2017	n/a	(0)	Measurement error: (0)
Stienen et al.	2019	n/a	(0)	n/a
Jakobsson et al.	2020	(–)	(+)	Measurement error: (+)
Master et al.	2020	(–)	(0)	Measurement error: (0)
Maldaner et al.	2021	n/a	(?)	Measurement error: (?)
Maldaner et al.	2021	(–)	(?)	Measurement error: (?)
Stienen et al.	2021	(–)	n/a	n/a

(+) positive rating, (?) indeterminate, (0) no information, (-) negative rating

Criterion validity and responsiveness

The correlation coefficient pooling between TUG and VAS back and leg pain was 0.26 (moderate) (95% CI 0.19 to 0.34) and 0.28 (moderate) (95% CI 0.20 to 0.36) [10, 19–21, 24]. The pooled coefficient of TUG with ODI [3, 10, 19, 20] and RMDI [10, 19] was 0.33 (moderate) (95% CI 0.27 to 0.39) and 0.33 (moderate) (95% CI 0.24 to 0.42), respectively. Besides, TUG has correlated with the quality-of-life PROMs with a coefficient of −0.22 to −0.26 (moderate) (EQ5D Index 95% CI −0.35 to −0.16) [10, 19], (SF12-PCS 95% CI −0.33 to −0.15) [10, 19] and (SF12-MCS 95% CI −0.32 to −0.13) [10, 19]. The pooled coefficient of TUG with COMI, ZCQ-PF and ZCQ-SS was 0.46 (moderate) (95% CI 0.30 to 0.59), 0.43 (moderate) (95% CI 0.26 to 0.56), and 0.38 (moderate) (95% CI 0.21 to 0.52), respectively [18, 21]. Correlation coefficients based on heterogeneous data (each only in one study) were TUG-5MWT: −0.58, TUG-1MST: −0.67, TUG-50FWT: 0.66, TUG-BPI (back pain): 0.06, TUG-BPI (leg pain): 0.006, ZCQ (PS): 0.38, ZCQ (SS): 0.27 [3, 18, 20, 21] (Figs. 2, 3, 4, 5).

TUG's 3-day [19, 22, 23, 25] and 6-week [18, 19, 22, 23, 25] pooled responsiveness results were 0.14 (low) (95% CI −0.02 to 0.29) and 0.74 (moderate to strong) (95% CI 0.60 to 0.89), respectively. Among the studies based on heterogeneous data, Jakobsson and colleagues presented TUG's pre-op and post-op values as 9.1 ± 4.4 and 5.7 ± 1.1 in a subgroup of 31 patients ($p < 0.05$) [20]. On the other hand, Master and colleagues reported a TUG score of 15.5 ± 8.1 pre-op and 10.6 ± 5.1 postoperative 12th months ($p < 0.001$) [3] (Table 2; Fig. 6).

Other psychometric properties

The reliability results analyzed in only one study were excellent, with 0.97 for intra-rater ICC and 0.99 for inter-rater ICC. Gautschi et al. [10] also provided the SEM value of TUG. The SEM intrarater and interrater values were 0.21 s and 0.23 s, respectively. In the three studies, the MCID was between 0.9 and 3.4 s [3, 24, 25]. Only one study calculated the MIC value as (95% CI) −17.6% (−20.7 to −10.2%) [20] (Table 2).

Discussion

TUG test is one of the most commonly used physical performance assessment tools for ongoing and following LSS [10, 22]. The present systematic review and meta-analysis aimed to investigate the measurement properties of the TUG in patients with LSS. According to the results, TUG was agreeably responsive (moderate to strong) at the mid-term (6 weeks) follow-up. TUG was primarily associated with COMI (moderate), evaluating pain, function, symptom-specific well-being, quality of life, and disability. TUG was also moderately related to physical function, pain and quality of life, respectively. In clinical practice, the TUG can be used as a reliable, valid and responsive tool to assess LSS patients' general status, especially in the mid-term.

Lumbar decompression surgery (with or without fusion) is a safe surgical procedure that has been performed for years to reduce pain, loss of function and improve patients' independence in daily living [13, 14]. It is crucial to evaluate the physical performance of individuals before these surgeries with measurement tests that

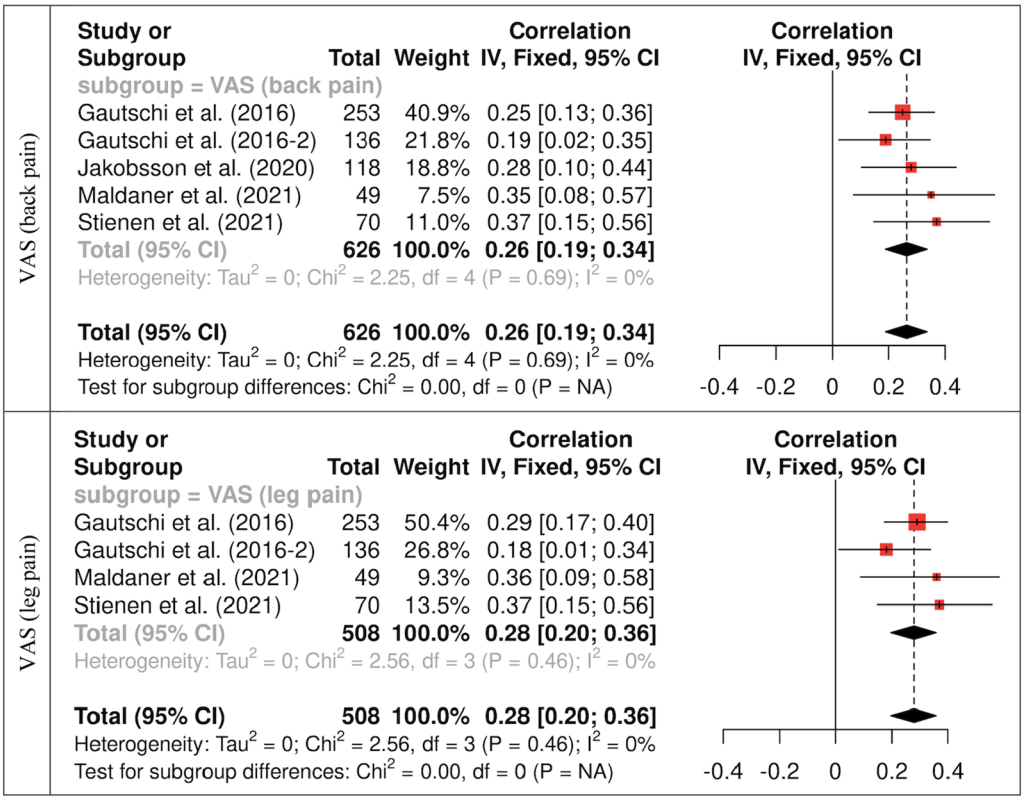

Fig. 2 Pooling results of the correlation coefficient between TUG and VAS

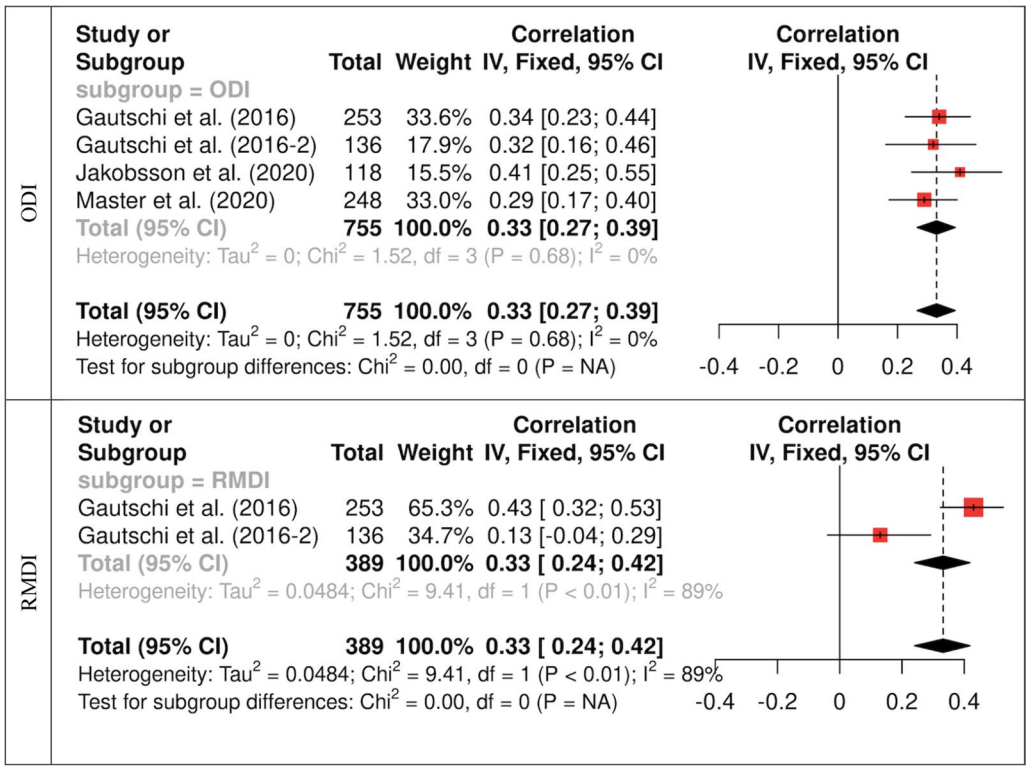

Fig. 3 Pooling results of the correlation coefficient between TUG with ODI and RMDI

include standardized protocols in order to evaluate the patient's actual clinical condition objectively and quantitatively [3, 8]. To our knowledge, no other study has examined the measurement properties of TUG, perhaps the most important of the tests used in clinical practice, in individuals before and after LSS.

The mean age of the sample of the included studies ranged between 46 and 66 years [3, 10, 18–25]. A vast majority of the studies include middle-aged individuals. Hence, some studies enrolled older adults. However, since most of the studies included middle-aged individuals (median 56.25), the decline in physical function observed due to the physiology of aging can be disregarded. The patients were followed during immediate, acute and chronic periods. Responsiveness of TUG during these several follow-up periods provided essential data to clinical practice [18, 20]. In addition, although there were more male subjects in most studies, approximately 40% of female subjects displayed a homogeneous gender distribution.

The most notable result of the quality analysis was a negative (−) and "fair to good" score in most studies for criterion validity. The main reason for this issue was the < 100 sample size and correlation coefficient values less than 0.70 in COSMIN scoring [26, 30]. In the responsiveness analysis, studies ranked "fair to good", "(0) no information", and "(?) indeterminate" scores as a result of insufficient data in sample size and statistical analysis. In addition, only 1 of the studies provided measurement and statistical data on reliability. On the other hand, due to lacking statistical analysis and a small sample size on "measurement error", the results of the studies had lower quality. In this context, future studies can address TUG's test–retest or inter-rater reliability more comprehensively with specific ICC Shrout Fleiss models [34]. In addition, responsiveness results should also address the ROC and AUC curve with longer-term follow-up to provide more apparent measurement characteristics of TUG in individuals with LSS [35]. Within the scope of criterion validity, TUG needed to be adequately compared with gold-standard performance tests such as the Five Times Sit to Stand Test, Stair Test, 6MWT, and 30 s Chair Sit to Stand Test. The correlation of these tests with each other may provide coefficients above 0.70, which might improve validity inferences' quality at a higher evidence level [26, 30].

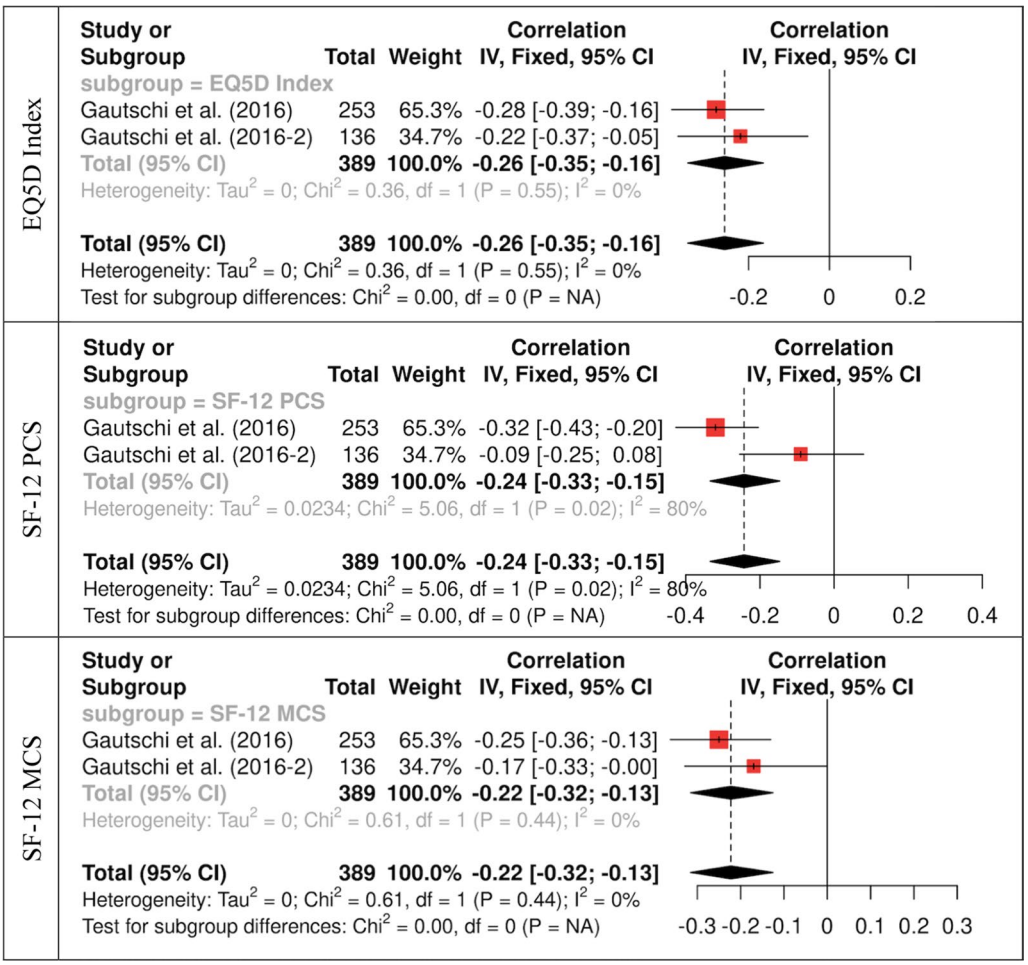

Fig. 4 Pooling results of the correlation coefficient between TUG with EQ5D and SF-12

"Validity" is an analysis to indicate the degree of accuracy of the test for an intended parameter [36]. Validity results showed that TUG was primarily related to COMI. Since it is comprehended that COMI represents the general condition, such as function, pain, symptoms, and quality of life, owing to its holistic structure, it can be argued that TUG provides a comprehensive evaluation in cases with LSS [37]. TUG was secondarily associated with ZC-PF, ZCQ-SS, ODI and RMDI. This concordance suggests that TUG secondarily indicates the function of the patients, as expected. It should be noted that TUG represents general condition rather than function. Thirdly, the relationship between pain and TUG was noteworthy. Since it is known that the increase in the pain level of individuals would increase the loss of function, the moderate pooled coefficient correlation with low back and leg pain was not surprising [9]. Among the correlation coefficient pooling, TUG was least associated with quality-of-life scores. Since the correlational analysis of individuals in the pre-op period is usually presented, the correlation of TUG with SF-12 and EQ5D after surgical and rehabilitation interventions may present higher validation coefficients. Also, since the quality of life is more perceptible in the chronic period after the health service is provided, it would be vital to examine the criterion validity after long-term follow-up in future studies [13, 14, 38].

Responsiveness analysis investigated whether the TUG provides a clinical improvement response following the treatment at different follow-up times. While the TUG was low responsive at a 3-day follow-up, it revealed a more responsive clinical improvement at a 6-week mid-term follow-up. This outcome suggests that postoperative functional gains usually occur in a moderate-term period, as rehabilitation effectiveness usually occurs after 1 month in LSS. It would be essential to prove the

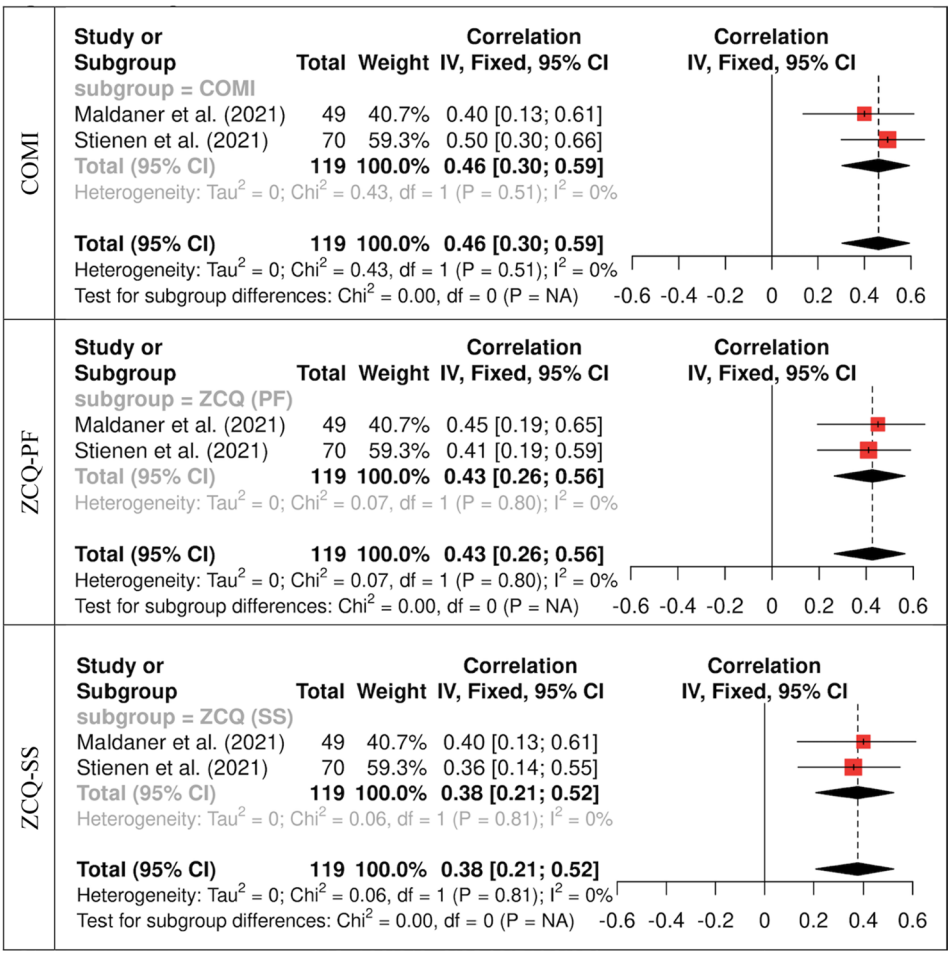

Fig. 5 Pooling results of the correlation coefficient between TUG with COMI and ZCQ

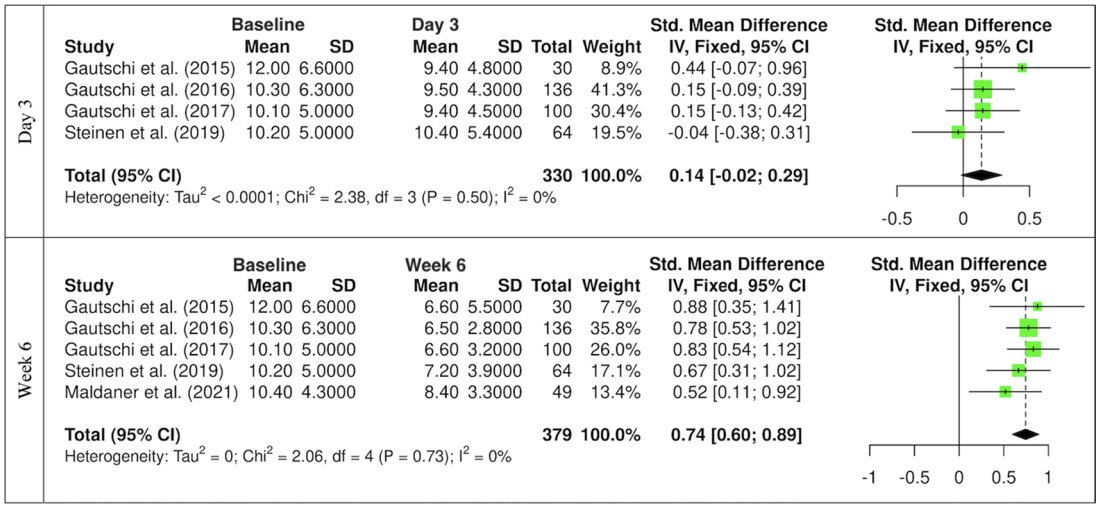

Fig. 6 Pooling results of TUG in terms of responsiveness

further responsiveness of TUG in terms of long-term monitorization of individuals. As a matter of fact, Jakobsson and colleagues and Master and colleagues, which we could not include in the meta-analysis, confirmed that TUG was responsive in individuals after LSS at 6 and 12 months, respectively [3, 20]. Considering the data within the scope of effect size with additional studies may provide pooling results at a high level of evidence.

Only 1 study demonstrated test–retest and inter-rater reliability. Reliability indicates whether the questionnaire can consistently capture the clinical condition of the same individual under identical clinical conditions [26, 39]. The TUG provided highly reliable results in individuals with LSS. In future studies, presenting the reliability with Bland Altman agreement analysis could reveal the reliability of TUG in individuals with LSS more comprehensively. MCID revealed the smallest clinically significant change in "seconds". Among these studies, MCID was found to be 3.4 s in the study with a mean age of 46 years and 1.3 s in the study with a mean age of 62 years. In another study with an average age of 49 years, results ranging between 0.9 and 3 s were noteworthy. It was observed that advancements in smaller units were more clinically significant in aging (with greater age) individuals. These data may provide reference outcomes on treatment improvements in clinical practice.

Limitations

All databases were not searched in the present systematic review. Some databases (CINAHL) were inaccessible regarding public sources. Secondly, the surgical procedures in the studies were not homogenous. Since it is comprehended that the outcomes and rehabilitation responses of individuals with "minimally invasive or conventional surgical" methods or "decompression or fusion" techniques differ [13, 14], a more homogeneous pooling should be considered for future studies. Last but not least, the study was not registered in a "systematic review database" (International Prospective Register of Systematic Reviews-PROSPERO). Protocol registration of reviews is essential for the integrity of the methodology.

Conclusions

In conclusion, TUG was agreeably responsive (moderate to strong) at the mid-term (6 weeks) follow-up. TUG was primarily associated with COMI (moderate), evaluating pain, function, symptom-specific well-being, quality of life, and disability. TUG was also moderately related to physical function, pain and quality of life, respectively. In clinical practice, the TUG can be used as a reliable, valid and responsive tool to assess LSS patients' general status, especially in the mid-term.

Abbreviations

PROMs	Patient-reported outcome measures
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
COSMIN	COnsensus-based Standards for the selection of health Measurement Instruments
WoS	Web of Science
LSS	Lumbar Spine Surgery
LFS	Lumbar Fusion Surgery
LDS	Lumbar Decompression Surgery
TUG	Timed Up and Go
VAS	Visual Analog Scale
ODI	Oswestry Disability Index
RMDQ	Roland Morris Disability Questionnaire
SF-12	Short Form-12
EQ5D	EuroQoL 5 Dimension
COMI	Core Outcome Measures Index
ZCQ	Zurich Claudication Questionnaire
6-MWT	6-Meter Walk Test
BPI	Brief Pain Inventory
5-MWT	5-Minute Walk Test
1-MSC	1-Minute Stair Climbing
50-FTWT	50-Foot Walk Test
TSK	Tampa Scale of Kinesiophobia
HADS	Hospital Anxiety and Depression Scale
ICC	Intraclass correlation coefficient
MCID	Minimal clinically important difference
SEM	Standard error of measurement
SMR	Standardized response means
MIC	Minimal important change

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41983-024-00805-z.

Additional file 1: Appendix S1. Search strategies.

Acknowledgements

Thanks to İsmet Tümtürk, PT, MSc for his contributions to the screening and searching procedures of this systematic review.

Author contributions

FÖ and İT (please see "Acknowledgement") researched literature and conceived the study. FÖ was involved in protocol development and writing. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 7 December 2023 Accepted: 6 February 2024
Published online: 19 February 2024

References

1. Rao PJ, Phan K, Maharaj MM, Pelletier MH, Walsh WR, Mobbs RJ, et al. Accelerometers for objective evaluation of physical activity following spine surgery. J Clin Neurosci. 2016;26:14–8.
2. Herrera IH, de la Presa RM, Gutiérrez RG, Ruiz EB, Benassi JG. Evaluation of the postoperative lumbar spine. Radiologia. 2013;55(1):12–23.
3. Master H, Pennings JS, Coronado RA, Henry AL, O'Brien MT, Haug CM, et al. Physical performance tests provide distinct information in both predicting and assessing patient-reported outcomes following lumbar spine surgery. Spine. 2020;45(23):1556–63.
4. Maldaner N, Stienen MN. Subjective and objective measures of symptoms, function, and outcome in patients with degenerative spine disease. Arthritis Care Res. 2020;72:183–99.
5. Gray DR, Rongve I. Role for PROMs data to support quality improvement across the healthcare system: an informed exchange with senior health system leaders. Healthc Pap. 2012;11(4):34.
6. Voglis S, Ziga M, Zeitlberger AM, Sosnova M, Bozinov O, Regli L, et al. Smartphone-based real-life activity data for physical performance outcome in comparison to conventional subjective and objective outcome measures after degenerative lumbar spine surgery. Brain Spine. 2022;2: 100881.
7. Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine. 1998;23(22):2412–21.
8. Dobson F, Hinman RS, Roos EM, Abbott JH, Stratford P, Davis AM, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthr Cartil. 2013;21(8):1042–52.
9. Corniola M-V, Stienen M, Joswig H, Smoll N, Schaller K, Hildebrandt G, et al. Correlation of pain, functional impairment, and health-related quality of life with radiological grading scales of lumbar degenerative disc disease. Acta Neurochir. 2016;158:499–505.
10. Gautschi OP, Smoll NR, Corniola MV, Joswig H, Chau I, Hildebrandt G, et al. Validity and reliability of a measurement of objective functional impairment in lumbar degenerative disc disease: the timed up and go (TUG) test. Neurosurgery. 2016;79(2):270–8.
11. Low M, Burgess LC, Wainwright TW. A critical analysis of the exercise prescription and return to activity advice that is provided in patient information leaflets following lumbar spine surgery. Medicina. 2019;55(7):347.
12. Mannion AF, Denzler R, Dvorak J, Müntener M, Grob D. A randomised controlled trial of post-operative rehabilitation after surgical decompression of the lumbar spine. Eur Spine J. 2007;16:1101–17.
13. Özden F. The effectiveness of physical exercise after lumbar fusion surgery: a systematic review and meta-analysis. World Neurosurg. 2022;163:396–412.
14. Özden F. The effect of exercise interventions after lumbar decompression surgery: a systematic review and meta-analysis. World Neurosurg. 2022;167:1878–8750.
15. Ghent F, Mobbs RJ, Mobbs RR, Sy L, Betteridge C, Choy WJ. Assessment and post-intervention recovery after surgery for lumbar disk herniation based on objective gait metrics from wearable devices using the gait posture index. World Neurosurg. 2020;142:111–6.
16. Janssens L, Brumagne S, Claeys K, Pijnenburg M, Goossens N, Rummens S, et al. Proprioceptive use and sit-to-stand-to-sit after lumbar microdiscectomy: the effect of surgical approach and early physiotherapy. Clin Biomech. 2016;32:40–8.
17. Silva KN, Imoto AM, Almeida GJ, Atallah AN, Peccin MS, Trevisani VFM. Balance training (proprioceptive training) for patients with rheumatoid arthritis. CDSR. 2010;5:1–10.
18. Maldaner N, Sosnova M, Zeitlberger AM, Ziga M, Gautschi OP, Regli L, et al. Responsiveness of the self-measured 6-minute walking test and the timed up and go test in patients with degenerative lumbar disorders. J Neurosurg. 2021;1:1–8.
19. Gautschi OP, Joswig H, Corniola MV, Smoll NR, Schaller K, Hildebrandt G, et al. Pre-and postoperative correlation of patient-reported outcome measures with standardized timed up and go (TUG) test results in lumbar degenerative disc disease. Acta Neurochir. 2016;158:1875–81.
20. Jakobsson M, Brisby H, Gutke A, Lundberg M, Smeets R. One-minute stair climbing, 50-foot walk, and timed up-and-go were responsive measures for patients with chronic low back pain undergoing lumbar fusion surgery. BMC Musculoskelet Disord. 2020;20(1):1–12.
21. Stienen MN, Maldaner N, Sosnova M, Zeitlberger AM, Ziga M, Weyerbrock A, et al. External validation of the timed up and go test as measure of objective functional impairment in patients with lumbar degenerative disc disease. Neurosurg. 2021;88(2):142–9.
22. Gautschi OP, Corniola MV, Joswig H, Smoll NR, Chau I, Jucker D, et al. The timed up and go test for lumbar degenerative disc disease. J Clin Neurosci. 2015;22(12):1943–8.
23. Stienen MN, Maldaner N, Joswig H, Corniola MV, Bellut D, Prömmel P, et al. Objective functional assessment using the "timed up and go" test in patients with lumbar spinal stenosis. Neurosurg Focus. 2019;46(5):E4.
24. Maldaner N, Sosnova M, Ziga M, Zeitlberger AM, Bozinov O, Gautschi OP, et al. External validation of the minimum clinically important difference in the timed-up-and-go test after surgery for lumbar degenerative disc disease. Spine. 2021;47(4):337–42.
25. Gautschi OP, Stienen MN, Corniola MV, Joswig H, Schaller K, Hildebrandt G, et al. Assessment of the minimum clinically important difference in the timed up and go test after surgery for lumbar degenerative disc disease. Neurosurgery. 2017;80(3):380–5.
26. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012;21:651–7.
27. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4(1):1–9.
28. Clarke M, Clarke TT, Clarke L. Cochrane systematic reviews as a source of information for practice and trials. Trials. 2011;12:49.
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5:1–10.
30. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42.
31. Goldsmith MR, Bankhead CR, Austoker J. Synthesising quantitative and qualitative research in evidence-based patient information. J Epidemiol Community Health. 2007;61(3):262.
32. Beheshti A, Chavanon M-L, Christiansen H. Emotion dysregulation in adults with attention deficit hyperactivity disorder: a meta-analysis. BMC Psychiatry. 2020;20(1):1–11.
33. Cohen J. A power primer. Psychol Bull. 1992;112(1):155–9.
34. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420.
35. Pathak A, Wilson R, Sharma S, Pryymachenko Y, Ribeiro DC, Chua J, et al. Measurement properties of the patient-specific functional scale and its current uses: an updated systematic review of 57 studies using COSMIN guidelines. JOSPT. 2022;52(5):262–75.
36. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166.e7-166.e16.
37. Mannion AF, Porchet F, Kleinstück F, Lattig F, Jeszenszky D, Bartanusz V, et al. The quality of spine surgery from the patient's perspective. Part 1: the core outcome measures index in clinical practice. Eur Spine J. 2009;18:367–73.
38. Perez-Cruet MJ, Hussain NS, White GZ, Begun EM, Collins RA, Fahim DK, et al. Quality-of-life outcomes with minimally invasive transforaminal lumbar interbody fusion based on long-term analysis of 304 consecutive patients. Spine. 2014;39(3):191–8.
39. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53(5):459–68.

Publisher's Note