Machine learning–augmented objective functional testing in the degenerative spine: quantifying impairment using patient-specific five-repetition sit-to-stand assessment

Victor E. Staartjes, BMed,^{1–3} Anita M. Klukowska, BSc, BMBS,^{2–4} Moira Vieli, BMed,¹ Christiaan H. B. van Niftrik, MD, PhD,¹ Martin N. Stienen, MD,⁵ Carlo Serra, MD,¹ Luca Regli, MD,¹ W. Peter Vandertop, MD, PhD,² and Marc L. Schröder, MD, PhD³

¹Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland; ²Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Neurosurgery, Amsterdam Movement Sciences, Amsterdam; ³Department of Neurosurgery, Bergman Clinics, Amsterdam, The Netherlands; ⁴Department of Surgery, Royal Derby Hospital, Derby, United Kingdom; and ⁵Department of Neurosurgery, Cantonal Hospital St. Gallen, St. Gallen, Switzerland

OBJECTIVE What is considered “abnormal” in clinical testing is typically defined by simple thresholds derived from normative data. For instance, when testing using the five-repetition sit-to-stand (5R-STs) test, the upper limit of normal (ULN) from a population of spine-healthy volunteers (10.5 seconds) is used to identify objective functional impairment (OFI), but this fails to consider different properties of individuals (e.g., taller and shorter, older and younger). Therefore, the authors developed a personalized testing strategy to quantify patient-specific OFI using machine learning.

METHODS Patients with disc herniation, spinal stenosis, spondylolisthesis, or discogenic chronic low-back pain and a population of spine-healthy volunteers, from two prospective studies, were included. A machine learning model was trained on normative data to predict personalized “expected” test times and their confidence intervals and ULNs (99th percentiles) based on simple demographics. OFI was defined as a test time greater than the personalized ULN. OFI was categorized into types 1 to 3 based on a clustering algorithm. A web app was developed to deploy the model clinically.

RESULTS Overall, 288 patients and 129 spine-healthy individuals were included. The model predicted “expected” test times with a mean absolute error of 1.18 (95% CI 1.13–1.21) seconds and R^2 of 0.37 (95% CI 0.34–0.41). Based on the implemented personalized testing strategy, 191 patients (66.3%) exhibited OFI. Type 1, 2, and 3 impairments were seen in 64 (33.5%), 91 (47.6%), and 36 (18.8%) patients, respectively. Increasing detected levels of OFI were associated with statistically significant increases in subjective functional impairment, extreme anxiety and depression symptoms, being bedridden, extreme pain or discomfort, inability to carry out activities of daily living, and a limited ability to work.

CONCLUSIONS In the era of “precision medicine,” simple population-based thresholds may eventually not be adequate to monitor quality and safety in neurosurgery. Individualized assessment integrating machine learning techniques provides more detailed and objective clinical assessment. The personalized testing strategy demonstrated concurrent validity with quality-of-life measures, and the freely accessible web app (<https://neurosurgery.shinyapps.io/5RSTS/>) enabled clinical application.

<https://thejns.org/doi/abs/10.3171/2021.8.FOCUS21386>

KEYWORDS machine learning; precision medicine; personalized medicine; objective functional impairment; sit to stand; impairment; artificial intelligence

STANDARDIZED outcome assessment has evolved from radiological and physician-rated outcomes toward patient-reported outcome measures—not only in clinical practice, but importantly, also in quality and safety improvement programs and in scientific research.^{1–5} Ac-

curate capture of clinical outcomes is a necessary step toward monitoring trends in neurosurgical quality and safety improvement programs, including detection of trends or spikes in poor outcomes and infection or complication rates. In addition, standardized outcome measurement en-

ABBREVIATIONS ADL = activities of daily living; ODI = Oswestry Disability Index; OFI = objective functional impairment; MAE = mean absolute error; RMDQ = Roland-Morris Disability Questionnaire; RMSE = root-mean-square error; ULN = upper limit of normal; 5R-STs = five-repetition sit-to-stand.

SUBMITTED June 30, 2021. **ACCEPTED** August 25, 2021.

INCLUDE WHEN CITING DOI: 10.3171/2021.8.FOCUS21386.

ables the setting of benchmarks for surgical quality among individual centers and surgeons, assessment of the efficacy of new interventions, checklists, and protocols, and identification of systematic human errors.^{3,6}

Up to now, both patient-reported and objective outcome measures have relied on single, fixed thresholds derived from normative populations to distinguish between healthy and unhealthy individuals, or between a good and bad outcome. For example, in degenerative lumbar spine disease, the presence of objective functional impairment (OFI) is normally determined by comparing the five-repetition sit-to-stand (5R-STs) test time of a particular patient with the upper limit of normal (ULN) of test times in a spine-healthy population (10.5 seconds).^{7–12} If the patient takes longer than these 10.5 seconds to complete the 5R-STs, OFI can be diagnosed and further classified based on fixed thresholds.^{8,12} The advantages of such thresholds are their simplicity, generalizability, ease of derivation and validation, and simple anchoring to a representative normative population. However, there are inherent disadvantages. Differences in test properties among individuals become obvious when considering the example of body height, which is one of the most powerful determinants of 5R-STs performance, as tall patients need to cover a longer distance standing up and sitting down from a chair that has a standardized height.^{8,13,14}

Instead of fixed thresholds, dynamic thresholds that respect a patient's demographics could allow for a more accurate grading of OFI. Some developments in this direction have been made, such as the introduction of tables reporting fixed grading thresholds distinguished by male and female, or younger and older than 65 years.^{15,16} However, memorizing a range of fixed thresholds makes clinical application cumbersome. A still more-detailed and more-personalized testing strategy could improve upon fixed thresholds by enabling the grading of disease tailored to a particular patient, instead of groups or subgroups of patients. The future of medicine is moving toward more personalized healthcare analytics in the era of personalized or precision medicine.¹⁷ We aimed to implement this rationale by developing a machine learning–based personalized testing strategy to quantify impairment using patient-specific 5R-STs assessment.

Methods

Study Design

To train and validate the patient-specific objective functional testing model, data from two prospective studies including both patients with spinal disease and spine-healthy volunteers were pooled.^{8,9} Between October 2017 and June 2018, all participants were seen at a specialized outpatient spine surgery clinic.

We trained a machine learning model to predict a personalized “expected” or “normal” test time from basic demographic data, including age, height, weight, BMI, sex, and smoking status. This individually predicted 5R-STs test time can be used as a benchmark of the performance that a patient would be expected to achieve without the presence of disease, or in case of full recovery, afterward (e.g., surgery for lumbar disc herniation).⁸

Subsequently, individualized thresholds such as the personalized ULN can be calculated, representing the 99th percentile of the 5R-STs test time that would be expected among individuals with the same demographics in the normative population. If patients can perform the 5R-STs within their personalized ULN, the presence of OFI can be ruled out. Instead, if patients perform more slowly than this personalized ULN, the presence of OFI can be diagnosed, and the type of OFI can then be assessed using a clustering method (V. E. Staartjes et al., unpublished data). This method applies unsupervised clustering using a k-means matching algorithm and classifies patients with OFI into 3 clinically distinct OFI types. Types 1 and 2 represent relatively mild to moderate impairment, with type 2 additionally representing a higher likelihood of extreme anxiety and depression symptoms, being bedridden, and an inability to work. A type 3 OFI corresponds to severe impairment that is associated with an even higher magnitude of the aforementioned symptoms.

This report was compiled according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.¹⁸

Ethics Approval

The two prospective studies (ClinicalTrials.gov identifiers NCT03303300 and NCT03321357) were approved by the local IRB (Medical Research Ethics Committees United). Informed consent was obtained from all participants.

Study Population

All included patients were scheduled for surgery and were assessed during outpatient consultations. Inclusion criteria were the presence of lumbar disc herniation, lumbar spinal stenosis, spondylolisthesis, or discogenic chronic low-back pain. Patients with synovial facet cysts causing radiculopathy were not included. Patients with hip or knee prosthetics, and those requiring walking aides were excluded to eliminate these confounders. Individuals with missing 5R-STs data were excluded. A normative reference population of spine-healthy individuals was also included, most of whom were partners of the patients with similar demographics, employees of the department, or other volunteers.

Measurements and Data Collection

The 5R-STs was performed according to a previously published testing protocol.^{8,9,19} Most importantly, an armless, hard-seated chair of standard height (48 cm) was firmly placed against a wall, stable shoes were worn, and patients were instructed and motivated to perform the test “as fast as possible.” The 5 repetitions were timed from the “go” command to the completed fifth stand (5R-STs test time). If the patient was unable to perform the test in 30 seconds, or not at all, this was noted and the test score was recorded as 30 seconds.⁸ Some patients and volunteers performed the test twice, in which case the mean test time was used.

A range of questionnaires was additionally used. All participants provided information on baseline sociodemographic data, as well as numeric rating scale scores

for back and leg pain severity, and they completed Dutch versions of the Oswestry Disability Index (ODI), Roland-Morris Disability Questionnaire (RMDQ), and Euro-QOL-5D-3L to capture subjective functional impairment as well as health-related quality of life.

Statistical Analysis

Analyses were carried out using R version 4.0.5 (The R Foundation).²⁰ A $p \leq 0.05$ on two-tailed tests was considered statistically significant. Data are reported as mean $\pm$ standard deviation for continuous and numbers (percentages) for categorical data. Variables or patients with missing data in more than 25% of the fields were excluded from the analysis. Missing data that were assumed to be missing (completely) at random were imputed using the k-nearest neighbor imputation, with $k = 5$.²¹ Baseline characteristics of the patient and control cohorts were compared using Pearson's chi-square tests or Welch's two-sample t-test. Patients without OFI and those with the 3 types of OFI were compared using Pearson's chi-square test or one-way ANOVA.

Model Development

To predict personalized expected 5R-STs test times along with their 95% CIs and ULN (99th percentile), a quantile regression model with a least absolute shrinkage and selection operator (LASSO) penalty was trained for the 2.5th, 50th, 97.5th, and 99th quantiles (tau).^{22,23} This machine learning algorithm was trained on data from a representative cohort of spine-healthy volunteers of all ages. The model was internally validated using repeated fivefold cross-validation with 10 repeats to assess out-of-sample performance. Resampled root-mean-square error (RMSE), mean absolute error (MAE), and R^2 , along with their 95% CIs were obtained using 1000 repetitions of a bootstrap with replacement. Agreement of predicted and actual test times in the normative population was further evaluated using Bland-Altman analysis.²⁴

If patients were able to perform the 5R-STs within their personalized ULN (actual 5R-STs test time $\leq$ personalized ULN), OFI was ruled out. Whenever OFI was diagnosed (actual 5R-STs test time $>$ personalized ULN), we applied a clustering algorithm (V. E. Staartjes et al., unpublished data) to identify OFI types 1 to 3.

A web app allowing for measurement of the 5R-STs and automatizing the prediction and clustering process was constructed. The calculations run server-side, and the web app can easily be used on mobile devices.

Results

Cohorts

Detailed characteristics of the volunteer and patient cohorts are provided in Table 1. Among the 129 spine-healthy volunteers, 167 of 3354 (5.0%) data fields were missing. Similarly, among the 288 patients with spinal disease, 215 of 7488 (2.9%) data fields were missing. The mean age for the volunteer cohort was 40 ± 19 years, and for the patient cohort it was 47 ± 13 years ($p < 0.001$). Sixty volunteers (47%) and 141 patients (49%) with spinal disease were male ($p = 0.722$). The mean 5R-STs test time recorded in the

TABLE 1. Baseline characteristics of the spine-healthy volunteers and patients with lumbar degenerative disease, pooled from two prospective studies

Parameter	Volunteer Cohort (n = 129)	Patient Cohort (n = 288)	p Value
Mean 5R-STs test time, sec	6.27 (1.84)	13.50 (6.44)	<0.001*
Mean age, yrs	40.48 (18.80)	47.12 (13.38)	<0.001*
Male sex	60 (46.5)	141 (49.0)	0.722
Mean height, cm	171.90 (9.99)	175.83 (10.14)	<0.001*
Mean weight, kg	71.06 (13.93)	78.40 (13.67)	<0.001*
Mean BMI, kg/m ²	24.01 (4.04)	25.27 (3.34)	0.001*
Smoking status			<0.001*
Active	19 (14.7)	81 (28.1)	
Ceased	27 (20.9)	88 (30.6)	
Never	83 (64.3)	119 (41.3)	
Prior spine op	7 (5.4)	55 (19.1)	0.001*
Indication for op			0.001*
Disc herniation		201 (69.8)	
Spinal stenosis		57 (19.8)	
Spondylolisthesis		15 (5.2)	
Discogenic chronic low-back pain		15 (5.2)	
Highest level of education			0.225
Elementary school	4 (3.1)	4 (1.4)	
High school	44 (34.1)	122 (42.4)	
Higher education	77 (59.7)	149 (51.7)	
Post-doctoral	4 (3.1)	13 (4.5)	
Analgesic drug use			<0.001*
Not regularly	108 (83.7)	50 (17.4)	
Weekly	9 (7.0)	26 (9.0)	
Daily	12 (9.3)	212 (73.6)	
Ability to work			<0.001*
Full	122 (94.6)	76 (26.4)	
Limited	5 (3.9)	64 (22.2)	
Unable	2 (1.6)	148 (51.4)	
Mean EQ-5D-3L index	0.95 (0.14)	0.38 (0.30)	<0.001*
Mean EQ-5D-3L thermometer	84.78 (12.37)	49.46 (17.81)	<0.001*
Mean NRS back pain severity	0.96 (1.82)	5.95 (2.64)	<0.001*
Mean NRS leg pain severity	0.52 (1.36)	7.47 (1.88)	<0.001*
Mean ODI score	2.53 (6.72)	45.12 (17.02)	<0.001*
Mean RMDQ score	0.64 (1.86)	12.06 (5.35)	<0.001*

NRS = Numeric Rating Scale.

Values represent the number of patients (%) or mean (SD) unless indicated otherwise. Data are presented after imputation for missing data.

* $p \leq 0.05$.

volunteer cohort was 6.3 ± 1.8 seconds, while the mean test time was 13.5 ± 6.4 seconds among patients ($p < 0.001$).

Personalized Test Time Quantiles

Expected Test Times

To assess model fit at internal validation, we compared

actual test times and predicted ($\tau = 0.50$, 50th percentile) test times during cross-validation (Table 2). In terms of classic performance measures, RMSE was 1.48 (95% CI 1.43–1.53), MAE was 1.18 (95% CI 1.13–1.21), and R^2 was 0.37 (95% CI 0.34–0.41). Correspondingly, the correlation R of actual and predicted test times was 0.61 (95% CI 0.58–0.64). Bland-Altman analysis (Fig. 1) revealed a mean bias of -0.02 seconds, with a 95% limit of agreement of -2.77 to 2.74 seconds.

Personalized ULN

The mean personalized ULN, derived through prediction of the 99th percentile of the expected test time, for the entire patient cohort was 10.0 ± 1.3 seconds and ranged from 7.2 to 13.1 seconds (Fig. 2).

In Silico Application of Personalized Testing Strategy

Test Performance

All 288 patients were run through the web app to evaluate the results of the personalized testing strategy. The mean 5R-STST test time was 13.5 ± 6.4 seconds, ranging from 4.9 to 30.0 seconds. The mean deviation of actual test time from a particular patient's personalized ULN (Fig. 2) was 3.5 ± 6.7 seconds (range -7.2 to 21.6 seconds), leading to a diagnosis of OFI in 191 patients (66.3%).

Cluster Assignment

Among the 191 patients with OFI, 64 patients (34%) had type 1 impairment, and 91 (48%) and 36 (19%) had type 2 and 3 impairments, respectively.

TABLE 2. Performance measures of the quantile regression model during repeated fivefold cross validation

Performance Measure	Cross-Validation Performance (95% CI)
RMSE	1.48 (1.43–1.53)
MAE	1.18 (1.13–1.21)
R^2	0.37 (0.34–0.41)

The actual 5R-STST performance of the volunteer control cohort ($n = 129$) is compared with the corresponding predictions of the expected median test time ($\tau = 0.50$, 50th percentile). Bland-Altman analysis revealed a mean bias of -0.02 seconds, with a 95% limit of agreement of -2.77 to 2.74 seconds.

Test Interpretation

Table 3 demonstrates the final classification of all 288 patients using the machine learning-augmented testing strategy. Subjective functional impairment scores (ODI and RMDQ) increased with severity of OFI, as did rates of extreme anxiety and depression symptoms, being bed-ridden, extreme pain or discomfort, and inability to carry out activities of daily living (ADL) (all $p \leq 0.003$). Limited ability or inability to work also increased steadily with OFI severity ($p = 0.012$). Analgesic drug use was similar among all classifications ($p = 0.499$).

Back pain severity correlated with severity of OFI ($p < 0.001$) while leg pain did not ($p = 0.173$). Chronic low-back pain was by far the most common among patients with a type 3 OFI, while patients without OFI had a significantly higher rate of lumbar spinal stenosis ($p < 0.001$).

The mean age was significantly higher among patients

FIG. 1. Performance of the quantile regression model. **Left:** The actual 5R-STST performance of the volunteer cohort ($n = 129$) is compared with the corresponding predictions ($\tau = 0.50$, 50th percentile). Correlation was 0.61 (95% CI 0.58–0.64). **Right:** Bland-Altman analysis revealed a mean bias of -0.02 seconds, with a 95% limit of agreement of -2.77 to 2.74 seconds.

FIG. 2. Histograms of the personalized ULNs generated for the entire patient cohort (**left**; $n = 288$) as well as the personalized performance of the patient cohort, expressed as the deviation of the actual test time from each patient's personalized ULN (**right**). The **thick black line** indicates the median.

without OFI, while there were no significant differences in age among the 3 types of OFI ($p < 0.001$). A mean BMI of around 25 kg/m² was observed in patients without OFI and those with type 3 OFI, while those with type 1 and type 2 OFI are clearly demographically distinguished by normal-weight and overweight patients, respectively (Fig. 3A). The rate of active smokers increased steadily with severity of OFI ($p = 0.028$).

Deployment

A web app containing detailed testing instructions and providing capabilities for testing (either measuring the 5R-STSTest time using an integrated stopwatch or entering a previously measured test time), automated generation of personalized “expected” test time as well as personalized ULN, and automated interpretation (presence and type of OFI) was constructed. Details of 5 example patients from our cohort are presented in Table 4. The web app is freely available online (<https://neurosurgery.shinyapps.io/5RSTST/>).

Discussion

Using data from two prospective cohort studies, we have developed and internally validated a personalized testing strategy based on machine learning. Based on age, sex, height, weight, BMI, and smoking status, precise predictions of personalized “expected” test times and their ULNs can be generated for each patient. Patients requiring longer to complete the 5R-STSTest than their personalized ULN are deemed to be objectively functionally impaired. The extent of OFI can then be further classified using a clustering process. All steps of the testing process have been implemented in a freely accessible web app.

What is considered “abnormal” in clinical testing is usually defined by simple thresholds derived from normative data.²⁵ For instance, when using the 5R-STSTest, the ULN from a population of spine-healthy volunteers (10.5 seconds) is used to identify OFI.⁸ This approach is simple and effective, yet it fails to consider the radically different 5R-STSTest testing properties of different individuals. For instance, height is known to influence 5R-STSTest performance significantly.^{8,13,14} Since chairs of standardized height are used, the distance that needs to be covered with each sit-to-stand action is proportional to body height. Thus, a tall individual with the same health status as a comparable shorter individual will usually still require significantly longer to complete the 5R-STSTest. Apart from such obvious differences in testing properties, what is considered normal should optimally be based on a normative population that is as similar to the test subject as possible. One would expect a completely healthy 21-year-old rugby player to perform the 5R-STSTest more quickly than an otherwise healthy 78-year-old obese retiree, although both performances could be seen as normal for their specific situations. For this reason, ULNs should be derived from many individuals without functional impairment of different age ranges, nutritional status, et cetera. Of course, one could simply calculate multiple ULNs for younger and older, normal-weight and obese, male and female, or tall and short individuals. This would require generating an exponentially growing number of different thresholds for each subset, eventually also running into sample size limitations. Memorization and clinical application would also be increasingly cumbersome. A more elegant and detailed way of arriving at a personalized threshold for each patient is to model the effects of the most important demographic parameters for different quantiles of the norma-

TABLE 3. Classification of patients according to personalized ULN and cluster assignment

Parameter	No OFI (n = 97)	Type 1 OFI (n = 64)	Type 2 OFI (n = 91)	Type 3 OFI (n = 36)	p Value
Mean ULN	10.66 (1.37)	9.50 (1.15)	9.78 (1.19)	9.74 (1.03)	<0.001*
Mean 5R-STs test time, sec	8.35 (1.78)	13.15 (3.19)	13.86 (3.48)	27.07 (4.32)	<0.001*
Mean age, yrs	53.58 (14.14)	42.85 (12.36)	44.51 (11.62)	43.88 (10.95)	<0.001*
Male sex	56 (57.7)	13 (20.3)	47 (51.6)	25 (69.4)	<0.001*
Mean height, cm	175.71 (10.74)	169.86 (7.83)	178.92 (8.14)	181.92 (8.66)	<0.001*
Mean weight, kg	79.81 (12.90)	65.74 (6.08)	87.48 (8.74)	80.97 (12.38)	<0.001*
Mean BMI, kg/m ²	25.83 (3.16)	22.86 (2.41)	27.36 (2.35)	24.47 (3.38)	<0.001*
Smoking status					0.028*
Active	16 (16.5)	16 (25.0)	34 (37.4)	15 (41.7)	
Ceased	36 (37.1)	20 (31.2)	23 (25.3)	9 (25.0)	
Never	45 (46.4)	28 (43.8)	34 (37.4)	12 (33.3)	
Prior spine op	12 (12.4)	11 (17.2)	22 (24.2)	10 (27.8)	0.099
Indication for op					<0.001*
Disc herniation	52 (53.6)	49 (76.6)	72 (79.1)	28 (77.8)	
Spinal stenosis	33 (34.0)	11 (17.2)	11 (12.1)	2 (5.6)	
Spondylolisthesis	7 (7.2)	2 (3.1)	5 (5.5)	1 (2.8)	
Discogenic chronic low-back pain	5 (5.2)	2 (3.1)	3 (3.3)	5 (13.9)	
History of symptoms					0.749
≤6 wks	2 (2.1)	2 (3.1)	5 (5.5)	1 (2.8)	
6 wks–6 mos	14 (14.4)	9 (14.1)	14 (15.4)	7 (19.4)	
6 mos–1 yr	21 (21.6)	21 (32.8)	20 (22.0)	10 (27.8)	
>1 yr	60 (61.9)	32 (50.0)	52 (57.1)	18 (50.0)	
Analgesic drug use					0.499
Not regularly	14 (14.4)	11 (17.2)	14 (15.4)	3 (8.3)	
Weekly	7 (7.2)	3 (4.7)	12 (13.2)	4 (11.1)	
Daily	76 (78.4)	50 (78.1)	65 (71.4)	29 (80.6)	
Ability to work					0.012*
Full	35 (36.1)	18 (28.1)	17 (18.7)	6 (16.7)	
Limited	27 (27.8)	11 (17.2)	18 (19.8)	8 (22.2)	
Unable	35 (36.1)	35 (54.7)	56 (61.5)	22 (61.1)	
Mean NRS back pain severity	5.04 (2.72)	6.09 (2.74)	6.26 (2.31)	7.22 (2.27)	<0.001*
Mean NRS leg pain severity	7.19 (1.88)	7.80 (1.32)	7.44 (1.98)	7.75 (2.35)	0.173
Mean ODI score	38.43 (16.31)	46.41 (16.00)	46.35 (15.59)	59.00 (14.60)	<0.001*
Mean RMDQ score	9.48 (4.97)	12.66 (5.22)	12.78 (4.93)	16.50 (3.65)	<0.001*
Health-related quality of life					
Extreme anxiety & depression symptoms	1 (1.0)	3 (4.7)	7 (7.7)	5 (13.9)	0.003*
Bedridden	4 (4.1)	3 (4.7)	6 (6.6)	12 (33.3)	<0.001*
Extreme pain or discomfort	43 (44.3)	38 (59.4)	56 (61.5)	34 (94.4)	<0.001*
Unable to carry out ADL	17 (17.5)	16 (25.0)	24 (26.4)	21 (58.3)	<0.001*
Unable to care for oneself	1 (1.0)	0 (0.0)	0 (0.0)	2 (5.6)	0.002*
Mean EQ-5D-3L index	0.49 (0.28)	0.35 (0.28)	0.37 (0.30)	0.13 (0.23)	<0.001*
Mean EQ-5D-3L thermometer	54.04 (16.47)	45.27 (16.64)	50.29 (18.52)	42.08 (18.41)	0.001*

NRS = Numeric Rating Scale.

Values represent the number of patients (%) or mean (SD) unless indicated otherwise. Data are presented after imputation for missing data.

* p ≤ 0.05.

tive population. Some machine learning methods such as quantile regression enable this approach and can generate precise ULNs for each individual.^{22,23}

Our model demonstrated its capacity to predict per-

sonalized expected test times (50th percentile) with an accuracy of within 1.2 seconds of the actual test time, as well as predicting individualized ULNs (99th percentile).⁸ When defining the presence of OFI as an actual 5R-STs

FIG. 3. Scatterplots demonstrating clusters of functional impairment among the patient cohort ($n = 288$) in terms of selected continuous variables. **A:** BMI. **B:** Personalized ULN. **C:** EQ-5D-3L index. **D:** Actual test time.

performance that is slower than the personalized ULN, we observed that a slightly higher percentage of around two-thirds of the patient cohort was deemed to be impaired. This compares with approximately 50% to 60% of the spinal patient population that was deemed to be objectively functionally impaired using the standard ULN of 10.5 seconds.^{8,12} Those patients that were additionally classified as having OFI by our personalized testing strategy—and not by the usual fixed 10.5-second cutoff—were mostly younger and shorter patients who would indeed realistically be expected to complete the test in 7 or 8 seconds. Conversely, some very tall patients who normally would initially have been classified as impaired were now deemed not to

have OFI, because a test time of 13 seconds, for example, is still considered normal, given their height. Hence, one can argue that using personalized cut offs for objective tests of function seems to increase the diagnostic yield of these tests, which is of obvious value for both clinical care and research.

Whenever OFI was diagnosed, it was classified as types 1, 2, or 3 using a clustering algorithm. As discussed previously, the 3 types roughly correspond to different levels of impairment, with type 3 indicating severe OFI. Patients diagnosed with types 1 and 2 OFI often show similar levels of impairment, especially when considering 5R-STs test times only, but those with a type 2 OFI diagnosis have

TABLE 4. 5R-STs web app information, including demographics, clinical characteristics, test performance, and health-related quality of life, for 5 example patients

Parameter	Pt 1	Pt 2	Pt 3	Pt 4	Pt 5
Principal complaint	Neurogenic claudication	Neurogenic claudication	Radiating leg pain	Radiating leg pain	Chronic low-back pain
Age, yrs	48	69	56	35	32
Sex	M	F	M	F	M
Height, cm	185	168	180	185	188
Weight, kg	78	68	91	115	88
BMI, kg/m ²	22.8	24.1	28.1	33.6	24.9
Smoking status	Never	Ceased	Ceased	Active	Ceased
Actual 5R-STs test time, sec	4.98	12.6	16.30	17.08	30 (unable to complete test)
Predicted test time, sec (95% CI)	6.99 (4.21–10.09)	7.85 (5.01–12.01)	7.30 (4.87–10.61)	7.13 (6.80–7.98)	6.37 (4.18–8.32)
Personalized ULN, sec	10.24	12.06	10.86	8.33	8.53
OFI	No	Yes	Yes	Yes	Yes
Unsupervised cluster assignment	No impairment	Type 1	Type 2	Type 2	Type 3
Extreme anxiety & depression	No	No	No	No	Yes
Bedridden	No	No	No	No	Yes
Unable to care for oneself	No	No	No	No	No
Unable to carry out ADL	No	No	No	Yes	Yes
Unable to work	No	No	No	Yes	Yes

Pt = patient.

a slightly higher likelihood of extreme anxiety and depression symptoms, being bedridden, and an inability to work. In addition, patients with type 2 OFI were virtually all overweight (BMI ≥ 25 kg/m²) and were on average taller and more likely to be male and actively smoking than those with type 1 OFI. These differences may underline the practical applicability of this grading versus just looking at the 5R-STs test time alone; patients with type 1 and 2 OFI had the same test times and reported virtually the same level of symptoms, yet those with type 1 OFI appeared to be slightly less troubled by their symptoms than patients with type 2 OFI.

Concurrent validity of an outcome measurement or classification is assessed by comparing a certain measurement of interest with other relevant parameters that one would expect to differ between the levels of that measurement.²⁶ Our personalized testing strategy demonstrated that multiple relevant anchors of health-related quality of life changed steadily from no OFI to OFI type 3, indicating concurrent validity. For instance, increasing levels of OFI were associated with increases in subjective functional impairment, extreme anxiety and depression symptoms, being bedridden, extreme pain or discomfort, inability to carry out ADL, and a limited ability to work. Differences were particularly pronounced between patients classified as being without impairment versus those with types 1 and 2 OFI, and between patients with type 1 and 2 OFI versus those with type 3 OFI. It is also known that low-back pain can lead to relatively more impairment in ADL than

can radiculopathy, particularly when performing the 5R-STs.^{27–30} Correspondingly, back pain severity increased with each level of OFI, while leg pain severity was not affected.¹²

As machine learning methods become more broadly adopted in many fields of medicine,^{17,31–33} it is feasible that clinical and scientific patient assessment—including laboratory studies, radiological studies, and physical examination—will move from simple fixed thresholds (e.g., a ULN for D-dimer of < 250 ng/mL²⁵) to personalized cutoffs based on comparable individuals from a normative population (e.g., with age-adjusted D-dimers).³⁴ We also expect that integration of other machine learning techniques will enable even more automated testing; the 5R-STs could be automatically rated using machine vision or accelerometers for motion tracking,³⁵ and demographic data about a particular patient could be pulled from electronic health records.³⁶ At an even higher level of abstraction, OFI could potentially be graded based on how patients walk into the examination room and sit down or get up from a chair. Nonetheless, the applications of personalized cutoffs and other extremely personalized measures in actual clinical practice and in quality and safety improvement, apart from their applications in research, are currently few and far between, and there is not yet enough evidence to support their adoption as standard of care. Even if clear prognostic subgroups can be defined and outcome measurements become more granular and specific, it does not necessarily follow that this would lead to any real-world benefit to patients.

Limitations

Our data originated from two prospective studies but were collected at a single Dutch center. Although we collected data from a normative population of all ages, the models developed on Dutch individuals may not necessarily generalize to other populations. However, the data that were used (demographics such as age, sex, and BMI, as well as 5R-STs testing) are not center-specific. Furthermore, the 5R-STs has demonstrably high interrater reliability.^{9,10} An external validation study would enable a definite statement on generalizability. Similarly, although out-of-sample error was properly assessed using cross validation in this study, a prospective validation study would provide further evidence on the out-of-sample performance (overfitting) of the quantile regression model. Patients with hip or knee prosthetics and those requiring walking aides were not included, and other comorbidities such as hip or knee arthritis and nonspinal neuropathies (e.g., diabetic polyneuropathy) were not systematically assessed. It is plausible that such comorbidities may skew 5R-STs performance toward higher test times. We could have included further input parameters into the quantile regression model to make its predictions even more accurate, but this would have come at the cost of ease-of-use. Perhaps even more importantly, the predictive value of OFI and its classification on outcomes after surgery must also be assessed. Lastly, we have not validated the personalized testing strategy in specific subgroups such as lumbar disc herniation or lumbar spinal stenosis, but it serves as a general model for frequent degenerative lumbar spine conditions. Both the prediction of personalized ULN and the clustering algorithm are independent of diagnosis or other clinical characteristics.

Conclusions

In the era of precision medicine, simple thresholds or even multiple thresholds for certain demographic subgroups, which may be hard to implement clinically, may eventually not be adequate to monitor quality and safety in neurosurgery. Individualized assessment integrating machine learning techniques provides more detailed and objective clinical assessment. We have developed and internally validated a method for generation of personalized reference ranges for the 5R-STs that allows for patient-specific quantification of impairment. If impairment is present, it can be further classified using a clustering algorithm. The personalized testing strategy demonstrated concurrent validity with quality-of-life measures. A freely accessible web app (<https://neurosurgery.shinyapps.io/5RSTS/>) enables clinical application of this personalized testing strategy.

Acknowledgments

We are grateful to all participating volunteers, and to Femke Beusekamp, BSc and Nathalie Schouman for study coordination and data collection. We also thank Marlies P. de Wispelaere, PDEng for her efforts in clinical informatics.

References

1. Falavigna A, Dozza DC, Teles AR, Wong CC, Barbagallo G, Brodke D, et al. Current status of worldwide use of Patient-Reported Outcome Measures (PROMs) in spine care. *World Neurosurg*. 2017;108:328-335.
2. Theodosopoulos PV, Ringer AJ, McPherson CM, Warnick RE, Kuntz C IV, Zuccarello M, Tew JM Jr. Measuring surgical outcomes in neurosurgery: implementation, analysis, and auditing a prospective series of more than 5000 procedures. *J Neurosurg*. 2012;117(5):947-954.
3. Theodosopoulos PV, Ringer AJ. Measuring outcomes for neurosurgical procedures. *Neurosurg Clin N Am*. 2015;26(2):P265-P269.
4. Fernández-Méndez R, Rastall RJ, Sage WA, Oberg I, Bullen G, Charge AL, et al. Quality improvement of neuro-oncology services: integrating the routine collection of patient-reported, health-related quality-of-life measures. *Neurooncol Pract*. 2019;6(3):226-236.
5. Asher AL, McCormick PC, Selden NR, Ghogawala Z, McGirt MJ. The National Neurosurgery Quality and Outcomes Database and NeuroPoint Alliance: rationale, development, and implementation. *Neurosurg Focus*. 2013;34(1):E2.
6. Rock AK, Opalak CF, Workman KG, Broadus WC. Safety outcomes following spine and cranial neurosurgery: evidence from the National Surgical Quality Improvement Program. *J Neurosurg Anesthesiol*. 2018;30(4):328-336.
7. Stienen MN, Ho AL, Staartjes VE, Maldaner N, Veeravagu A, Desai A, et al. Objective measures of functional impairment for degenerative diseases of the lumbar spine: a systematic review of the literature. *Spine J*. 2019;19(7):1276-1293.
8. Staartjes VE, Schröder ML. The five-repetition sit-to-stand test: evaluation of a simple and objective tool for the assessment of degenerative pathologies of the lumbar spine. *J Neurosurg Spine*. 2018;29(4):380-387.
9. Staartjes VE, Beusekamp F, Schröder ML. Can objective functional impairment in lumbar degenerative disease be reliably assessed at home using the five-repetition sit-to-stand test? A prospective study. *Eur Spine J*. 2019;28(4):665-673.
10. Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, Radwan H. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. *Spine (Phila Pa 1976)*. 1998;23(22):2412-2421.
11. Teixeira da Cunha-Filho I, Lima FC, Guimarães FR, Leite HR. Use of physical performance tests in a group of Brazilian Portuguese-speaking individuals with low back pain. *Physiother Theory Pract*. 2010;26(1):49-55.
12. Klukowska AM, Schröder ML, Stienen MN, Staartjes VE. Objective functional impairment in lumbar degenerative disease: concurrent validity of the baseline severity stratification for the five-repetition sit-to-stand test. *J Neurosurg Spine*. 2020;33(1):4-11.
13. Ng SSM, Cheung SY, Lai LSW, Liu ASL, Ieong SHI, Fong SSM. Association of seat height and arm position on the five times sit-to-stand test times of stroke survivors. *BioMed Res Int*. 2013;2013:642362.
14. Ng SSM, Cheung SY, Lai LSW, Liu ASL, Ieong SHI, Fong SSM. Five Times Sit-To-Stand test completion times among older women: influence of seat height and arm position. *J Rehabil Med*. 2015;47(3):262-266.
15. Stienen MN, Smoll NR, Joswig H, Corniola MV, Schaller K, Hildebrandt G, Gautschi OP. Validation of the baseline severity stratification of objective functional impairment in lumbar degenerative disc disease. *J Neurosurg Spine*. 2017;26(5):598-604.
16. 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):270-278.
17. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219.

18. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335(7624):806-808.
19. Jones SE, Kon SSC, Canavan JL, Patel MS, Clark AL, Nolan CM, et al. The five-repetition sit-to-stand test as a functional outcome measure in COPD. *Thorax*. 2013;68(11):1015-1020.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. Accessed September 9, 2021. <https://www.R-project.org/>
21. Kowarik A, Templ M. Imputation with the R package VIM. *J Stat Softw*. 2016;74:i07.
22. Koenker R, Chernozhukov V, He X, Peng L. *Handbook of Quantile Regression*. CRC Press; 2017.
23. Koenker R, Portnoy S, Ng PT, Melly B, Zeileis A, Grosjean P, et al. quantreg: Quantile regression. R-project.org. Accessed September 9, 2021. <https://CRAN.R-project.org/package=quantreg>
24. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307-310.
25. Pagana KD, Pagana TJ, Pagana TN. *Mosby's Diagnostic and Laboratory Test Reference*. Elsevier Health Sciences; 2018.
26. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol 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(7):737-745.
27. Staartjes VE, Klukowska AM, Schröder ML. Association of maximum back and leg pain severity with objective functional impairment as assessed by five-repetition sit-to-stand testing: analysis of two prospective studies. *Neurosurg Rev*. 2020;43(5):1331-1338.
28. Kothe R, Kohlmann T, Klink T, Rütger W, Klinger R. Impact of low back pain on functional limitations, depressed mood and quality of life in patients with rheumatoid arthritis. *Pain*. 2007;127(1-2):103-108.
29. Andersson GB. Epidemiological features of chronic low-back pain. *Lancet*. 1999;354(9178):581-585.
30. Leveille SG, Guralnik JM, Hochberg M, Hirsch R, Ferrucci L, Langlois J, et al. Low back pain and disability in older women: independent association with difficulty but not inability to perform daily activities. *J Gerontol A Biol Sci Med Sci*. 1999;54(10):M487-M493.
31. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920-1930.
32. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358.
33. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.
34. Righini M, Van Es J, Den Exter PL, Roy PM, Verschuren F, Ghuysen A, et al. Age-adjusted D-dimer cutoff levels to rule out pulmonary embolism: the ADJUST-PE study. *JAMA*. 2014;311(11):1117-1124.
35. Ejupi A, Brodie M, Gschwind YJ, Lord SR, Zagler WL, Delbaere K. Kinect-based five-times-sit-to-stand test for clinical and in-home assessment of fall risk in older people. *Gerontology*. 2015;62(1):118-124.
36. Staartjes VE, Stienen MN. Data mining in spine surgery: leveraging electronic health records for machine learning and clinical research. *Neurospine*. 2019;16(4):654-656.

Disclosures

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author Contributions

Conception and design: Staartjes, Schröder. Acquisition of data: Staartjes, Klukowska, Schröder. Analysis and interpretation of data: Staartjes, Schröder. Drafting the article: Staartjes. Critically revising the article: Klukowska, Vieli, van Niftrik, Stienen, Serra, Regli, Vandertop, Schröder. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Staartjes. Statistical analysis: Staartjes. Administrative/technical/material support: Staartjes, Serra, Regli, Schröder. Study supervision: Staartjes, Regli, Vandertop, Schröder.

Correspondence

Victor E. Staartjes: c/o Bergman Clinics, Naarden, The Netherlands. victor.staartjes@gmail.com.