European Spine Journal (2020) 29:1702–1708
https://doi.org/10.1007/s00586-020-06343-5

ORIGINAL ARTICLE

Initial classification of low back and leg pain based on objective functional testing: a pilot study of machine learning applied to diagnostics

Victor E. Staartjes[1,2,3,6] · Ayesha Quddusi[4] · Anita M. Klukowska[2,5] · Marc L. Schröder[2]

Received: 7 August 2019 / Revised: 8 October 2019 / Accepted: 12 February 2020 / Published online: 18 February 2020

Abstract

Objective The five-repetition sit-to-stand (5R-STS) test was designed to capture objective functional impairment and thus provided an adjunctive dimension in patient assessment. The clinical interpretability and confounders of the 5R-STS remain poorly understood. In clinical use, it became apparent that 5R-STS performance may differ between patients with lumbar disk herniation (LDH), lumbar spinal stenosis (LSS) with or without low-grade spondylolisthesis, and chronic low back pain (CLBP). We seek to evaluate the extent of diagnostic information contained within 5R-STS testing.

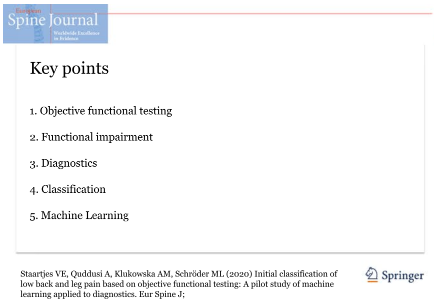
Methods Patients were classified into gold standard diagnostic categories based on history, physical examination, and imaging. Crude and adjusted comparisons of 5R-STS performance were carried out among the three diagnostic categories. Subsequently, a machine learning algorithm was trained to classify patients into the three categories using only 5R-STS test time and patient age, gender, height, and weight.

Results From two prospective studies, 262 patients were included. Significant differences in crude and adjusted test times were observed among the three diagnostic categories. At internal validation, classification accuracy was 96.2% (95% CI 87.099.5%). Classification sensitivity was 95.7%, 100%, and 100% for LDH, LSS, and CLBP, respectively. Similarly, classification specificity was 100%, 95.7%, and 100% for the three diagnostic categories.

Conclusion 5R-STS performance differs according to the etiology of back and leg pain, even after adjustment for demographic covariates. In combination with machine learning algorithms, OFI can be used to infer the etiology of spinal back and leg pain with accuracy comparable to other diagnostic tests used in clinical examination.

Graphic abstract

These slides can be retrieved under Electronic Supplementary Material.



Key points

1. Objective functional testing
2. Functional impairment
3. Diagnostics
4. Classification
5. Machine Learning

Staartjes VE, Quddusi A, Klukowska AM, Schröder ML (2020) Initial classification of low back and leg pain based on objective functional testing: A pilot study of machine learning applied to diagnostics. Eur Spine J;

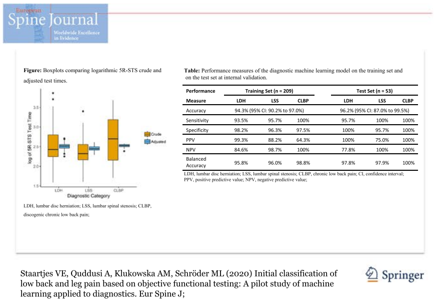


Figure: Boxplots comparing logarithmic 5R-STS crude and adjusted test times.

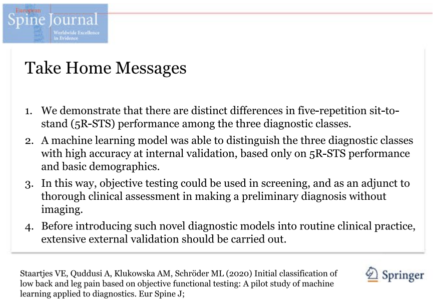
LDH, lumbar disc herniation; LSS, lumbar spinal stenosis; CLBP, discogenic chronic low back pain;

Table: Performance measures of the diagnostic machine learning model on the training set and on the test set at internal validation.

Performance	Training Set (n = 209)			Test Set (n = 53)		
Measure	LDH	LSS	CLBP	LDH	LSS	CLBP
Accuracy	94.3% (95% CI: 90.2% to 97.0%)			96.2% (95% CI: 87.0% to 99.5%)		
Sensitivity	93.5%	95.7%	100%	95.7%	100%	100%
Specificity	98.2%	96.3%	97.5%	100%	95.7%	100%
PPV	99.3%	88.2%	64.3%	100%	75.0%	100%
NPV	84.6%	98.7%	100%	77.8%	100%	100%
Balanced Accuracy	95.8%	96.0%	98.8%	97.8%	97.9%	100%

LDH, lumbar disc herniation; LSS, lumbar spinal stenosis; CLBP, chronic low back pain; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value;

Staartjes VE, Quddusi A, Klukowska AM, Schröder ML (2020) Initial classification of low back and leg pain based on objective functional testing: A pilot study of machine learning applied to diagnostics. Eur Spine J;



Take Home Messages

1. We demonstrate that there are distinct differences in five-repetition sit-to-stand (5R-STS) performance among the three diagnostic classes.
2. A machine learning model was able to distinguish the three diagnostic classes with high accuracy at internal validation, based only on 5R-STS performance and basic demographics.
3. In this way, objective testing could be used in screening, and as an adjunct to thorough clinical assessment in making a preliminary diagnosis without imaging.
4. Before introducing such novel diagnostic models into routine clinical practice, extensive external validation should be carried out.

Staartjes VE, Quddusi A, Klukowska AM, Schröder ML (2020) Initial classification of low back and leg pain based on objective functional testing: A pilot study of machine learning applied to diagnostics. Eur Spine J;

Keywords Objective functional testing · Functional impairment · Diagnostics · Classification · Machine learning

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00586-020-06343-5) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

Introduction

Objective functional testing has recently received more attention in the clinical assessment of patients suffering from back and leg pain [1–3]. Tests like the timed-up-and-go (TUG) and 6-minute-walking (6MWT) tests have already become standards in both spinal clinical practice and research [4, 5]. These tests correlate well with patient-reported measures of pain and subjective functional impairment and are robust to mental status as a confounder. In addition, these tests are able to capture deficits and complications, such as foot drop, tingling, or limping, which are not always picked up by questionnaires [2]. Tests for OFI are also more popular with patients compared to a battery of questionnaires [6]. In combination with well-validated questionnaires for pain severity, subjective functional impairment, and health-related quality of life, objective functional testing provides a holistic description of a patients' health state.

Recently, the five-repetition sit-to-stand test (5R-STS), which has already seen broad use for many other diseases such as Parkinson's disease or chronic obstructive pulmonary disease, has been validated for use in patients with back and leg pain [1]. The 5R-STS provides a simple and quick assessment of OFI feasible during busy clinical practice, is easy to administer, and has shown excellent test–retest reliability [1, 7–9].

It has been observed that the degree of OFI differs somewhat between the various causes of back and leg pain [1, 4]. While there is no clear distinction, it appears that clusters of patients exist, which differ in diagnosis, age, gender, and body metrics [1, 7]. It is currently unknown if these clusters have any prognostic clinical impact. However, it is conceivable that the degree of OFI could provide hints for the initial suspected cause of back and leg pain, without the need for imaging. Our hypothesis is that it is possible to accurately classify patients who present with back and leg pain into a suspected diagnosis. Quick obtainment of a suspected diagnosis could potentially guide further assessment and treatment, save costs, or even enable more accurate diagnostics in regions with limited resources, where imaging may not always be available [10].

Machine learning is continually gaining importance in medical predictive analytics. Machine learning algorithms can uncover highly complex interactions among variables that allow for accurate prognosis, analysis of medical images, and other novel applications [11, 12]. Consequently, the rationale of this study was to evaluate the degree to which a machine learning algorithm can correctly classify patients suffering from back and leg pain into an initial diagnostic category based on objective functional testing.

Materials and methods

Design

Pooled data from two prospective studies formed the basis of this study [1, 7]. Between October 2017 and June 2018, patients were seen at a specialized short-stay spine clinic. They completed a variety of questionnaires, as well as a test for OFI (5R-STS). As a "gold standard," the primary cause of back and leg pain was determined by a combination of clinical assessment, history, and diagnostic magnetic resonance imaging (MRI, Magnetom Essenza, Siemens, 1.5 T) and classified as either lumbar disk herniation (LDH), lumbar spinal stenosis (LSS) with or without low-grade spondylolisthesis, or discogenic chronic low back pain (CLBP). The study was compiled according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement [13].

Ethical approval

The two prospective studies (ClinicalTrials.gov Identifiers: NCT03303300 and NCT03321357*)* were approved by the local institutional review board (Medical Research Ethics Committees United, Registration Numbers: W17.107 and W17.134) and were conducted according to the Declaration of Helsinki. Informed consent was obtained from all participants.

Study population

All enrolled patients were scheduled for surgery and were assessed during outpatient consultations. Inclusion criteria were the presence of LDH, LSS, or discogenic CLBP. 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. We also excluded all healthy volunteers—recruited in the control group—from this analysis.

Primary outcome measures

The primary endpoint of this study was the accuracy of the machine learning model in classifying the diagnostic categories on the test set (internal validation). Internal or external validation—the final testing of a model on new data—is vital to evaluate the out-of-sample error of a given model. If the out-of-sample error is not higher compared to the model's error on the training data, overfitting can be ruled out, and the model is said to generalize well to new data.

Overfitting occurs when the model is too closely fitted to the training data and then consequently demonstrates high or near-perfect performance on the training data [14]. However, because it is "overfitted" to the training data, it will perform poorly on new, unseen testing data. The proper use of internal or external validation enables the diagnosis of overfitting.

Secondary outcome measures

The 5R-STS was performed according to the protocol described by Jones et al. [1, 8]. Measurements were obtained during the initial clinical visit, under instruction from a licensed physiotherapist. The participants were asked to sit down on an armless chair of standard height (48 cm) and with a hard seat, firmly placed against a wall. The participants were instructed to fold their arms across their chest and to keep their feet flat on the ground. Participants were required to wear stable shoes for the test. To become familiarized with the movement, the participants were asked to stand up fully and sit back down again once without using their upper limbs. If assistance was required, or if the maneuver could not be completed, the test was abandoned. Otherwise, the patients were asked to, starting on the command "go," stand up fully and sit down again, landing on the seat firmly, five times as fast as possible. Using a stopwatch, we timed the five repetitions from the initial command to the completed fifth stand. This time was recorded as the participant's score. If the patient was unable to perform the test in 30 s, or not at all, this was noted and the test score was recorded as 30 s [1].

A range of PROMs were additionally used. Patients were asked to complete questionnaires containing baseline sociodemographic data, as well as numeric rating scales (NRS) for back and leg pain severity, and validated Dutch versions of the Oswestry Disability Index (ODI), Roland-Morris Disability Questionnaire (RMDQ), and EuroQOL-5D-3L (EQ-5D) to capture subjective functional impairment as well as HRQOL. Participants filled out the questionnaires right after initially performing the test during the clinical visit.

Analytical methods

Data were reported as mean ± standard deviation for continuous and numbers (percentages) for categorical data. Analyses were carried out using R version 3.5.2 (The R Foundation for Statistical Computing, Vienna, Austria) [15]. Kruskal–Wallis H or Chi-square tests with Yates' correction for continuity were performed to test for differences in 5R-STS performance and basic demographic parameters among the diagnostic categories. Adjusted 5R-STS test times were calculated using a linear regression model adjusted for gender, age, height, and weight [1, 16, 17]. A $p \leq 0.05$ on two-sided tests was considered significant.

A machine learning model was trained to classify patients into one of the three above-mentioned diagnostic categories. Only the 5R-STS test time, patient age, gender, height, and weight were provided as inputs to the model. Data were randomly split into a training and a test set, in a 80/20 ratio. Random upsampling was applied to the training set to reduce the bias introduced by class imbalance [18, 19]. Class imbalance is present when the classes of the endpoint are not equally distributed, which can lead to artificially high performance in terms of overall accuracy, with poor sensitivity, specificity, or balanced accuracy. Class imbalance and its deleterious effects are often not properly prevented or diagnosed [18, 19].

Bootstrap resampling was applied using 25 repetitions with replacement. Resampling is vital to better estimate out-of-sample error during training, which often prevents or reduces overfitting. Bootstrap resampling works by repeatedly and randomly drawing samples with replacement from the training set, and continually evaluating the error of the model on these drawn samples.

A range of different models were tested, including neural networks, extreme gradient boosting, naïve Bayes classifiers, k-nearest neighbors, and fuzzy rule-based systems [20]. Hyperparameters were tuned until a final, best model based on logarithmic loss was selected. As opposed to the parameters of a model, which are optimized during training, the hyperparameters are those aspects of a given algorithm that need to be set before training. In neural networks, for example, the number of hidden layers and the number of neurons per layer need to be specified beforehand and are thus considered hyperparameters. This final model was then evaluated on the test set (internal validation) for accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and balanced accuracy. The statistical code (Supplementary Content 1) is provided.

Results

Cohort

A total of 262 patients were included. Detailed patient characteristics are provided in Table 1. There was no missing data. The mean age was 47.9 ± 13.4, with 134 (51.1%) male patients. In terms of diagnostic categories, 188 patients (71.8%) presented with LDH, 59 (22.5%) with LSS with or without spondylolisthesis, and 15 (5.7%) with CLBP. Age ($p < 0.001$) and height ($p = 0.029$) differed among the diagnostic categories (Table 2).

Table 1 Baseline patient characteristics. Continuous variables are presented as mean ± SD and categorical variables as frequency (percentage)

Parameter	Value ($n = 262$)
Age (years), mean ± SD	47.9 ± 13.4
Male gender, *n* (%)	134 (51.1)
Height (cm), mean ± SD	176.2 ± 10.1
Weight (kg), mean ± SD	78.9 ± 13.1
BMI (kg/m^2), mean ± SD	25.4 ± 3.2
Active smoker, *n* (%)	78 (29.8)
Diagnostic category, *n* (%)	
Lumbar disk herniation	188 (71.8)
Lumbar spinal stenosis	59 (22.5)
Chronic low back pain	15 (5.7)
Index level, *n* (%)	
L2-L3	8 (3.1)
L3-L4	29 (11.1)
L4-L5	100 (38.2)
L5-S1	125 (47.7)
VAS back pain, mean ± SD	5.9 ± 2.7
VAS leg pain, mean ± SD	7.4 ± 1.9
Oswestry Disability Index, mean ± SD	45.6 ± 17.3
Roland–Morris Disability Index, mean ± SD	12.2 ± 5.4
EQ-5D index, mean ± SD	0.37 ± 0.30
EQ-VAS, mean ± SD	50.1 ± 18.0

BMI body mass index, *VAS* visual analog scale

5R-STS performance among diagnostic categories

Crude 5R-STS test times (Table 2) differed significantly among the three diagnostic categories (Fig. 1), with LDH patients taking a mean of 13.9 ± 7.08 s, while patients with LSS and CLBP took 11.1 ± 5.18 s and 17.2 ± 10.51, respectively ($p = 0.041$). When calculating test times corrected for gender, age, height, and weight, the difference remained statistically significant, with smaller confidence intervals ($p < 0.001$).

Machine learning-based diagnostic classification

The final model was a fuzzy rule-based classification system based on Chi's method [20]. Fuzzy rule-based classifiers are any models that apply fuzzy logic—as opposed to traditional, so-called "crisp" logic—to arrive at their classification [20, 21]. In crisp logic, e.g., a binary outcome such as occurrence of a complication is either true or not (crisp label). In fuzzy logic, varying degrees of the outcome are possible, such as "very likely", "not likely", or "very unlikely". Thus, the degree of membership to a class is provided (soft label) [21].

At internal validation (test set), the diagnostic classification accuracy was 96.2% (95% CI 87.0–99.5%), indicating excellent discrimination (Table 3). Classification sensitivity was 95.7%, 100%, and 100% for LDH, LSS, and CLBP, respectively. Similarly, classification specificity was 100%, 95.7%, and 100% for the three diagnostic categories. Balanced accuracy—calculated as the average of the proportions of correctly classified patients of each diagnostic category individually—was 97.8%, 97.9%, and 100%. The difference in performance among the training and test set was minimal, indicating that overfitting was negligible. The confusion matrices are provided in Supplementary Content 2.

Table 2 Comparison of 5R-STS performance and basic demographic characteristics among patients with lumbar disk herniation, lumbar spinal stenosis with or without spondylolisthesis, and chronic low back pain

Parameter	Diagnostic category			
	LDH ($n = 188$)	LSS ($n = 59$)	CLBP ($n = 15$)	*p*
5R-STS performance (s)				
Crude test time, mean ± SD	13.9 ± 7.08	11.1 ± 5.18	17.2 ± 10.51	0.041*
Log crude test time, mean ± SD	2.53 ± 0.44	2.37 ± 0.39	2.68 ± 0.61	
Adjusted test time, mean ± SD	13.8 ± 1.05	12.8 ± 1.02	13.9 ± 0.82	< 0.001*
Log adjusted test time, mean ± SD	2.62 ± 0.08	2.55 ± 0.08	2.63 ± 0.06	
Basic demographic parameters				
Male gender, *n* (%)	98 (52.1)	29 (49.2)	7 (46.7)	0.866
Age (years), mean ± SD	44.6 ± 12.1	59.4 ± 12.1	43.9 ± 10.1	< 0.001*
Height (cm), mean ± SD	177.2 ± 9.8	172.7 ± 10.7	177.5 ± 8.1	0.029*
Weight (kg), mean ± SD	79.0 ± 13.1	79.4 ± 13.7	76.1 ± 10.6	0.679

Continuous variables are presented as mean ± SD and categorical variables as frequency (percentage). Adjusted 5R-STS test times were corrected for gender, age, height, and weight using a linear regression model

OFI objective functional impairment, *5R-STS* five-repetition sit-to-stand test

*$p \leq 0.05$

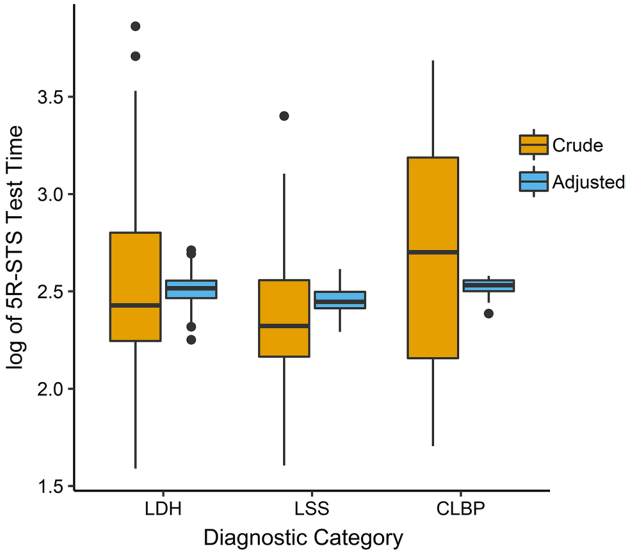


Fig. 1 Boxplots comparing logarithmic 5R-STS crude and adjusted test times. LDH, lumbar disk herniation; LSS, lumbar spinal stenosis; CLBP, discogenic chronic low back pain

Discussion

Among 262 patients from two prospective studies on the 5R-STS, there were significant differences in OFI between patients presenting with LDH, LSS with or without low-grade spondylolisthesis, and discogenic CLBP. These differences were more pronounced after correction of 5R-STS test times for demographic data. Subsequently, a machine learning model was trained on 209 patients to classify the patients into these three diagnostic categories based only on 5R-STS test time and patient age, gender, height, and weight. When predicting diagnostic categories on 53 new patients, the machine learning model exhibited the same performance, classifying diagnostic category with excellent discrimination. These findings suggest that patients show different levels of OFI depending on their specific pathology and that objective functional testing based on the 5R-STS not only gives an objective measure of the functional impairment but also contains at least some information on the possible diagnosis.

The mean age of patients with LSS is significantly higher than in the LDH and CLBP group, whereas this group has the shortest 5R-STS performance. This is remarkable since mobility in older patients is expected to be lower than in the younger population, with an expected higher test time. Apparently, patients with neurogenic claudication, as the prevailing symptom in LSS, do better during sit-to-stand movements, in contrast to the TUG test, in which patients need to stand up and walk [2]. This could be due to the stenosis being partially relieved on repeatedly standing and sitting. Patients in the discogenic CLBP group had the longest test times. This can be explained by the prevailing high intensity of the back pain in this patient population. This relationship is supported by prior findings, where it has been proven that the 5R-STS test is more related to the intensity of back pain than to leg pain [1, 7, 9]. In order to use objective tests in the future, it is necessary to stratify the test results for different potential confounders such as age and BMI, in order to more accurately predict OFI among different spinal pathologies [2].

Grading of OFI based on the 5R-STS has up to now been achieved through a fixed cutoff for the presence of OFI of 10.4 s [1]. However, not all patients can be expected to—without presence of specific OFI—perform an objective functional test in the same timespan. A healthy obese 75-year-old and an 22-year-old athlete should ideally not be judged to have OFI or not by the same static cutoff. Gautschi et al. [2] have suggested multiple cutoffs for patients over and under 65 years and for male and female patients to tackle this problem. Potentially, ML algorithms can suggest personalized "expected" cutoffs for each individual patient based on demographics [1]. In any case, our results show that not only age and gender, but also diagnostic category has to be taken into account when interpreting 5R-STS results.

The ML algorithm that was trained and internally validated in this study demonstrated good sensitivity and specificity in categorizing patients into three diagnostic categories

Table 3 Performance measures of the diagnostic machine learning model on the training set and on the test set at internal validation

Performance measure	Training set ($n=209$)			Test set ($n=53$)		
	LDH	LSS	CLBP	LDH	LSS	CLBP
Accuracy	94.3% (95% CI 90.2–97.0%)			96.2% (95% CI 87.0–99.5%)		
Sensitivity	93.5%	95.7%	100%	95.7%	100%	100%
Specificity	98.2%	96.3%	97.5%	100%	95.7%	100%
PPV	99.3%	88.2%	64.3%	100%	75.0%	100%
NPV	84.6%	98.7%	100%	77.8%	100%	100%
Balanced Accuracy	95.8%	96.0%	98.8%	97.8%	97.9%	100%

LDH lumbar disk herniation, *LSS* lumbar spinal stenosis, *CLBP* chronic low back pain, *CI* confidence interval, *PPV* positive predictive value, *NPV* negative predictive value based on only 5R-STS performance and basic demographic data. Although this was certainly not the primary focus of our study, such an algorithm could be used by clinicians to rule in or rule out suspected diagnoses with a certain degree of confidence and without the need for extensive imaging.

Machine learning methods are gaining popularity in all fields of medicine, especially their applications in natural language processing of electronic healthcare data, automated rating of performance, and interpretation of imaging. Our study shows that ML algorithms can combine data from objective clinical testing and patient history, reaching a precise diagnosis with a high level of accuracy, which creates precedence for the development of ML algorithms combing objective functional testing data with patient characteristics to reach a suspected diagnosis in other fields of medicine as well. Motion tracking-based 5R-STS assessment has been shown to be feasible [22]. With these advanced techniques, combined with the knowledge that classification of back and leg pain patients based on OFI is feasible, it is even conceivable that such initial diagnostic classifications may become more integrated in clinical practice. In the future, it may become possible to immediately suggest a suspected diagnosis with a measure of certainty based on the patients registered healthcare data and, e.g., on how the patient walks into the examination room and sits down or gets up from a chair. This can benefit patients and healthcare providers, especially in primary care or in remote areas where advanced imaging modalities may not be immediately available, as suggested by Munakomi [10].

The application of functional testing to diagnosis is certainly not ready for introduction into clinical practice as of yet, also because it has to be considered that imaging is and should in fact be performed in virtually all cases nowadays to rule out more delicate causes of back or leg pain and for surgical treatment, including grade II or higher spondylolisthesis. In addition, MRI or CT imaging nowadays is a prerequisite for surgery in virtually all cases. For example, the algorithm may misclassify patients as having LDH while they actually present with radiculopathy caused by higher-grade spondylolisthesis. At this point, it has to be stressed that it will forever remain hard for algorithms to outdo the basic tenets of taking a good history and examining patients thoroughly. Diagnoses made by ML algorithms in this context should never be seen as definite inferences, but rather as one additional, supportive “test” that may help guide decision-making, especially in rural areas where MRI may not be easily accessible.

In addition, and arguably more importantly, our study demonstrates that there are measurable differences in 5R-STS performance between patients with LDH, LSS, and discogenic CLBP. This finding has to be taken into account as potential confounders in future studies applying these functional tests as outcome measures.

Limitations

Our study, while based on two prospective studies, presents only single-center data. Although out-of-sample error was assessed in a held-out test set, external validation would be necessary before publishing the model or applying it in clinical practice elsewhere. Most likely, derivation of a multicenter model would increase generalizability, too. Moreover, inclusion of other covariates from the patient’s history as well as walking patterns could improve the algorithm’s robustness. We chose to only include age, gender, height, and weight alongside 5R-STS performance to assess if adjusted OFI is related to diagnosis in the purest of ways. We also had a smaller number of patients diagnosed with discogenic CLBP available, as compared to other diagnosis of LDH and LSS, making the evaluation for CLBP less secure. Furthermore, a larger training sample would be necessary to increase generalizability, and to assess calibration of the ML algorithm. Lastly, our model does not detect any “red flag” conditions such as cauda equina syndrome or spondylodiscitis.

Conclusions

In this study, we demonstrate that 5R-STS performance differs among patients with LDH, LSS, and discogenic CLBP, and that a simple test for objective functional impairment can help accurately classify patients presenting with back or leg pain into an initial diagnosis, when combined with a machine learning algorithm. In this way, objective testing could be used in screening, and as an adjunct to thorough clinical assessment in making a diagnosis without imaging. These findings may have implications in the initial diagnostic process, and may in the future be integrated with higher levels of automation.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Compliance with ethical standards

Conflict of interest The authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Staartjes VE, Schröder ML (2018) 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 29:380–387. https://doi.org/10.3171/2018.2.SPINE171416

2. Gautschi OP, Smoll NR, Corniola MV et al (2016) Validity and reliability of a measurement of objective functional impairment in lumbar degenerative disc disease: the timed up and go (TUG) test. Neurosurgery 79:270–278. https://doi.org/10.1227/NEU.0000000000001195
3. Gautschi OP, Corniola MV, Schaller K et al (2014) The need for an objective outcome measurement in spine surgery—the timed-up-and-go test. Spine J 14:2521–2522. https://doi.org/10.1016/j.spinee.2014.05.004
4. Gautschi OP, Joswig H, Corniola MV et al (2016) 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 (Wien) 158:1875–1881. https://doi.org/10.1007/s00701-016-2899-9
5. Guyatt GH, Sullivan MJ, Thompson PJ et al (1985) The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. Can Med Assoc J 132:919–923
6. Joswig H, Stienen MN, Smoll NR et al (2017) Patients' preference of the timed up and go test or patient-reported outcome measures before and after surgery for lumbar degenerative disk disease. World Neurosurg 99:26–30. https://doi.org/10.1016/j.wneu.2016.11.039
7. Staartjes VE, Beusekamp F, Schröder ML (2019) 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 Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc. https://doi.org/10.1007/s00586-019-05897-3
8. Jones SE, Kon SSC, Canavan JL et al (2013) The five-repetition sit-to-stand test as a functional outcome measure in COPD. Thorax 68:1015–1020. https://doi.org/10.1136/thoraxjnl-2013-203576
9. Stienen MN, Ho AL, Staartjes VE et al (2019) Objective measures of functional impairment for degenerative diseases of the lumbar spine: a systematic review of the literature. Spine J. https://doi.org/10.1016/j.spinee.2019.02.014
10. Munakomi S (2019) Letter to the editor. reappraising role of clinical evaluations in degenerative lumbar spine pathologies. J Neurosurg Spine 30:860–861. https://doi.org/10.3171/2018.10.SPINE181282
11. Senders JT, Staples PC, Karhade AV et al (2018) Machine learning and neurosurgical outcome prediction: a systematic review. World Neurosurg 109:476–486.e1. https://doi.org/10.1016/j.wneu.2017.09.149
12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
13. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 350:g7594
14. Siccoli A, de Wispelaere MP, Schröder ML, Staartjes VE (2019) Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. Neurosurg Focus 46:E5. https://doi.org/10.3171/2019.2.FOCUS18723
15. Core Team R (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
16. Siccoli A, Staartjes VE, de Wispelaere MP, Schröder ML (2018) Gender differences in degenerative spine surgery: Do female patients really fare worse? Eur Spine J. https://doi.org/10.1007/s00586-018-5737-3
17. Siccoli A, Staartjes VE, de Wispelaere MP, Schröder ML (2018) Is elective degenerative lumbar spine surgery in older adults safe in a short-stay clinic? Data from an institutional registry. Eur Geriatr Med. https://doi.org/10.1007/s41999-018-0132-5
18. Staartjes VE, Schröder ML (2018) Letter to the Editor. Class imbalance in machine learning for neurosurgical outcome prediction: are our models valid? Neurosurg Spine. https://doi.org/10.3171/2018.5.SPINE18543
19. Batista GEAPA, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl 6:20–29. https://doi.org/10.1145/1007730.1007735
20. Riza LS, Bergmeir C, Herrera F, Benítez JM (2015) frbs: fuzzy rule-based systems for classification and regression in R. J Stat Softw. https://doi.org/10.18637/jss.v065.i06
21. Kuncheva L (2000) Fuzzy classifier design. Springer
22. Ejupi A, Brodie M, Gschwind YJ et al (2015) Kinect-based five-times-sit-to-stand test for clinical and in-home assessment of fall risk in older people. Gerontology 62:118–124. https://doi.org/10.1159/000381804

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Victor E. Staartjes[1,2,3,6] · **Ayesha Quddusi**[4] · **Anita M. Klukowska**[2,5] · **Marc L. Schröder**[2]

✉ Victor E. Staartjes
victor.staartjes@gmail.com

[1] Machine Intelligence in Clinical Neuroscience (MICN) Lab, Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland

[2] Department of Neurosurgery, Bergman Clinics, Amsterdam, The Netherlands

[3] Amsterdam UMC, Vrije Universiteit Amsterdam, Neurosurgery, Amsterdam Movement Sciences, Amsterdam, The Netherlands

[4] Center for Neuroscience, Queens University, Kingston, ON, Canada

[5] School of Medicine, University of Nottingham, Nottingham, UK

[6] Department of Neurosurgery, c/o Bergman Clinics, Naarden, Rijksweg 69, 1411 GE Naarden, The Netherlands