European Spine Journal
https://doi.org/10.1007/s00586-019-06189-6

ORIGINAL ARTICLE

External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion

Ayesha Quddusi[1] · Hubert A. J. Eversdijk[2] · Anita M. Klukowska[2,3] · Marlies P. de Wispelaere[4] · Julius M. Kernbach[5] · Marc L. Schröder[2] · Victor E. Staartjes[2,6,7]

Received: 19 May 2019 / Revised: 16 September 2019 / Accepted: 13 October 2019

Abstract

Objective Patient-reported outcome measures following elective lumbar fusion surgery demonstrate major heterogeneity. Individualized prediction tools can provide valuable insights for shared decision-making. We externally validated the spine surgical care and outcomes assessment programme/comparative effectiveness translational network (SCOAP-CERTAIN) model for prediction of 12-month minimum clinically important difference in Oswestry Disability Index (ODI) and in numeric rating scales for back (NRS-BP) and leg pain (NRS-LP) after elective lumbar fusion.

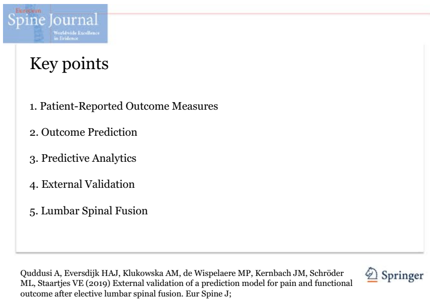
Methods Data from a prospective registry were obtained. We calculated the area under the curve (AUC), calibration slope and intercept, and Hosmer–Lemeshow values to estimate discrimination and calibration of the models.

Results We included 100 patients, with average age of 50.4 ± 11.4 years. For 12-month ODI, AUC was 0.71 while the calibration intercept and slope were 1.08 and 0.95, respectively. For NRS-BP, AUC was 0.72, with a calibration intercept of 1.02, and slope of 0.74. For NRS-LP, AUC was 0.83, with a calibration intercept of 1.08, and slope of 0.95. Sensitivity ranged from 0.64 to 1.00, while specificity ranged from 0.38 to 0.65. A lack of fit was found for all three models based on Hosmer–Lemeshow testing.

Conclusions The SCOAP-CERTAIN tool can accurately predict which patients will achieve favourable outcomes. However, the predicted probabilities—which are the most valuable in clinical practice—reported by the tool do not correspond well to the true probability of a favourable outcome. We suggest that any prediction tool should first be externally validated before it is applied in routine clinical practice.

Graphic abstract

These slides can be retrieved under Electronic Supplementary Material.



Key points

1. Patient-Reported Outcome Measures
2. Outcome Prediction
3. Predictive Analytics
4. External Validation
5. Lumbar Spinal Fusion

Quddusi A, Eversdijk HAJ, Klukowska AM, de Wispelaere MP, Kernbach JM, Schröder ML, Staartjes VE (2019) External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. Eur Spine J;

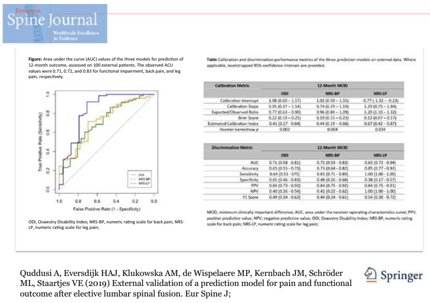


Figure: Area under the curve (AUC) values of the three models for prediction of 12-month outcome, assessed on 100 external patients. The observed ACU values were 0.71, 0.72, and 0.83 for functional impairment, back pain, and leg pain, respectively.

ODI, Oswestry Disability Index; NRS-BP, numeric rating scale for back pain; NRS-LP, numeric rating scale for leg pain;

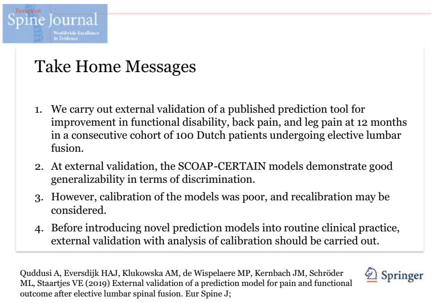
Table Calibration and discrimination performance metrics of the three prediction models on external data. Where applicable, bootstrapped 95% confidence intervals are provided.

Calibration Metric	12-Month MCID		
	ODI	NRS-BP	NRS-LP
Calibration Intercept	1.08 (0.60 – 1.57)	1.02 (0.50 – 1.55)	-0.77 (-1.32 – -0.23)
Calibration Slope	0.95 (0.37 – 1.54)	0.74 (0.29 – 1.19)	1.29 (0.75 – 1.84)
Expected/Observed Ratio	0.77 (0.63 – 0.90)	0.96 (0.84 – 1.09)	1.20 (1.10 – 1.32)
Brier Score	0.22 (0.19 – 0.25)	0.19 (0.15 – 0.23)	0.12 (0.07 – 0.17)
Estimated Calibration Index	0.41 (0.17 - 0.64)	0.44 (0.19 – 0.66)	0.67 (0.42 – 0.87)
Hosmer-Lemeshow p	0.002	0.004	0.034

Discrimination Metric	12-Month MCID		
	ODI	NRS-BP	NRS-LP
AUC	0.71 (0.58 - 0.81)	0.72 (0.59 - 0.83)	0.83 (0.72 - 0.94)
Accuracy	0.65 (0.55 - 0.73)	0.73 (0.64 - 0.82)	0.85 (0.77 - 0.92)
Sensitivity	0.64 (0.53 - 075)	0.81 (0.71 - 0.89)	1.00 (1.00 - 1.00)
Specificity	0.65 (0.46 - 0.83)	0.48 (0.26 - 0.68)	0.38 (0.17 - 0.57)
PPV	0.84 (0.73 - 0.93)	0.84 (0.75 - 0.92)	0.84 (0.75 - 0.91)
NPV	0.40 (0.26 - 0.54)	0.42 (0.22 - 0.62)	1.00 (1.00 - 1.00)
F1 Score	0.49 (0.34 - 0.62)	0.44 (0.24 - 0.61)	0.54 (0.30 - 0.72)

MCID, minimum clinically important difference; AUC, area under the receiver operating characteristics curve; PPV, positive predictive value; NPV; negative predictive value; ODI, Oswestry Disability Index; NRS-BP, numeric rating scale for back pain; NRS-LP, numeric rating scale for leg pain;

Quddusi A, Eversdijk HAJ, Klukowska AM, de Wispelaere MP, Kernbach JM, Schröder ML, Staartjes VE (2019) External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. Eur Spine J;



Take Home Messages

1. We carry out external validation of a published prediction tool for improvement in functional disability, back pain, and leg pain at 12 months in a consecutive cohort of 100 Dutch patients undergoing elective lumbar fusion.
2. At external validation, the SCOAP-CERTAIN models demonstrate good generalizability in terms of discrimination.
3. However, calibration of the models was poor, and recalibration may be considered.
4. Before introducing novel prediction models into routine clinical practice, external validation with analysis of calibration should be carried out.

Quddusi A, Eversdijk HAJ, Klukowska AM, de Wispelaere MP, Kernbach JM, Schröder ML, Staartjes VE (2019) External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. Eur Spine J;

Keywords Outcome prediction · Lumbar fusion · Patient-reported outcome · External validation · Predictive analytics

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00586-019-06189-6) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

Published online: 22 October 2019

Introduction

Prediction models, when externally validated, can be used in clinical practice for calculating individualized prognosis and enabling personalized risk–benefit estimation, instead of having to rely on generalized values reported in the literature [1–6]. Recently, there has been great interest in using machine learning (ML), as well as more conventional statistical modelling methods to devise models for predicting prognosis in various surgical procedures, including length of stay, rate of readmission, surgical site of infection, and post-operative surgical complications [7–9]. However, these models are usually only internally validated. Without external validation, applying prediction models in clinical practice in other cohorts than the derivation cohort can be pernicious [6, 10–12].

Khor et al. [2] recently proposed a predictive model for preoperatively estimating improvement in patient-reported outcome measures (PROMs) at 12 months after elective lumbar fusion for degenerative conditions. They used a statewide multicentre cohort to identify several factors including age, sex, race, insurance status, smoking status, among several others, that had an association with PROMs. Their fully developed model has been made freely available through a patient-facing web app [2]. The models have not been evaluated on external data and therefore ought to be used with caution on new patients from external centres [13, 14].

The aim of our study was to perform an external validation of the prediction tool developed by Khor et al. on a consecutive cohort of Dutch patients undergoing elective lumbar fusion [2].

Methods

Overview

The prediction model recently published by Khor et al. [2] has been developed on subsets of 1965 adult candidates for lumbar surgery prospectively collected at the Spine Surgical Care and Outcomes Assessment Program (SCOAP) and the survey centre at the Comparative Effectiveness Translational Network (CERTAIN), with patients originating from fifteen Washington state hospitals. The SCOAP-CERTAIN model has been incorporated into a user-friendly web app, available at https://becertain.shinyapps.io/lumbar_fusion_calculator, and the model has not been externally or prospectively internally validated as of yet. We compared the predicted probabilities generated by the SCOAP-CERTAIN model to the true 12-month pain and functional outcomes observed in our series to assess the model's external validity [13, 14].

Patient population

From a prospective registry, we identified a consecutive series of 100 patients [15] who had undergone elective, posterior lumbar spinal fusion for degenerative disease between 2014 and 2018. All patients were operated in a Dutch specialist short-stay spine centre under application of an Enhanced Recovery After Surgery (ERAS) protocol [16], and underwent robot-guided, minimally invasive (MI) transforaminal lumbar interbody fusion (MI-TLIF) or posterior lumbar interbody fusion (MI-PLIF) by a single senior neurosurgeon (M.L.S.) as previously described [17]. Adult patients with complete data were considered for inclusion. Primary indications for surgery included chronic low back pain (CLBP) caused by degenerative disc disease (DDD) as well as spondylolisthesis with or without concomitant central stenosis. Secondary diagnoses included coexistent disc herniation, radiculopathy, and failed back surgery syndrome (FBSS). In patients with additional low-grade spondylolisthesis, the decision of whether to add a fusion procedure to the decompression alone was based upon a validated decision-making protocol [18]. The study has been constructed according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement [19]. All patients included in the registry provided written informed consent. The prospective registry was authorized by the local institutional review board (Medical Research Ethics Committees United, Registration Number W16.065), and this study was carried out in accordance with the 2013 Declaration of Helsinki.

Data collection

Data collection was performed according to the specifications set by Khor et al. [2] Clinical and radiological baseline data were obtained at the first outpatient visit by the treating surgeon. Patients underwent magnetic resonance imaging (MRI) and a full clinical workup. The collected variables consisted of baseline patient-reported outcome measures (PROMs), as well as gender, age, smoking status (active/previous/never), insurance, ethnicity, American Society of Anesthesiologists (ASA) grade, opioid consumption, presence of asthma, and prior spine surgery. The "Medicaid" value of the variable "insurance" in the SCOAP-CERTAIN model most closely represents the Dutch health insurance system and thus was chosen for all patients. At 12 months post-operatively, PROMs were collected again in a paper-based fashion at a clinical follow-up visit by the treating surgeon [20].

Outcome measures

Patient-reported outcome

For PROM measurement, patients completed a standardized questionnaire including numeric rating scales (NRS) for back pain (NRS-BP) and leg pain (NRS-LP) severity, ranging from 0 to 10, and a validated Dutch version of the Oswestry Disability Index (ODI) to capture functional disability, ranging from 0 to 100, with higher values representing increasing severity [21]. According to Khor et al., we defined clinical success as achievement of the minimum clinically important difference (MCID) threshold of a ≥ 15-point reduction for ODI, and a ≥ 2-point reduction for NRS back and leg pain severity [2]. In cases where baseline ODI was < 15 or NRS was < 2 at baseline already (minimal pain/disability), and MCID would thus be impossible, no prediction was made in concordance to the output of the SCOAP-CERTAIN online calculator ("Cannot compute your chance of improvement—You are already at minimal disability").

Statistical analysis

Continuous data are presented as mean ± standard deviation, and categorical data as numbers and percentages. There was no missing data. The biostatistician was not blinded in terms of outcomes or predictor variables. Area under the receiver operating characteristics curve (AUC) was obtained by comparing the predicted probabilities with the true MCID outcome at 12 months. Similarly, calibration was assessed visually through inspection of calibration curves, and quantitatively through calibration intercept ("calibration-in-the-large") and slope [10, 11, 22]. A perfectly calibrated model has an intercept of 0.0, with a slope of 1.0. A Hosmer–Lemeshow test for goodness-of-fit was carried out, with a $p > 0.2$ indicating no lack of fit [5, 23]. In terms of calibration, we also assessed expected/observed event ratios (*E*/*O*-ratios) [13], as well as the Brier Score [24] and the Estimated Calibration Index [25].

Khor et al. [2] provide no threshold for binary classification. Accordingly, the threshold for binary classification was set at 0.5. This threshold is the most commonly observed threshold in logistic regression models when no specific threshold is specified. In addition, the threshold of 0.5 appeared to correspond closely to the post hoc identified optimal AUC-anchored thresholds ("closest to (0,1) criterion"). Subsequently, the binary classifications were compared to the true observed MCID outcome, and accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), as well as F1 Score were calculated.

Whenever applicable, bootstrapped 95% confidence intervals (CIs) based on 1000 resamples with replacement are provided. All analyses were carried out in *R* version 3.5.4 (The R Foundation for Statistical Computing, Vienna, Austria) [26]. The complete statistical code is provided in Supplementary Content 1.

Results

Among the 100 included patients, who had complete data, the mean age was 50.4 ± 11.4 years, and 51 patients (51%) were male. Minimal disability in terms of ODI was observed in only one case (1%) at baseline. No patients had minimum NRS-BP or NRS-LP at baseline. Detailed baseline characteristics of the development cohort reported by Khor et al. [2] and of the current external validation cohort are shown in Table 1. Notably, the patients in our cohort were around 10 years younger on average (61.3 ± 12.5 vs. 50.4 ± 11.4 years), and far fewer of our patients had higher ASA Scores (32% vs. 2%). In addition, radiculopathy was present more often in the development cohort (92% vs. 5%), as was stenosis (77% vs. 46%).

Patient-reported outcome

At 12 months post-operatively, ODI scores had improved a mean of − 29.3 ± 20.7 from baseline, with NRS-BP and NRS-LP improving − 3.7 ± 3.1 and − 4.4 ± 3.2, respectively. Achievement of the MCID was seen in 73 patients (73%) for ODI. In addition, 77 patients (77%) achieved MCID for NRS-BP, while 76 patients (76%) achieved MCID for NRS-LP. Table 2 summarizes outcome measures in the development cohort [2] and the external validation cohort.

Calibration

A detailed overview of calibration measures is provided in Table 3. A calibration intercept of 1.08 (95% CI 0.60–1.57) and slope of 0.95 (95% CI 0.37–1.54) were observed for prediction of MCID in ODI at 12 months, with a Hosmer–Lemeshow $p = 0.002$ (Fig. 1). The low *E*/*O*-ratio of 0.77 (95% CI 0.63–0.90) indicates a model that underestimated the probability of a favourable outcome.

Similarly, for NRS-BP, we observed a calibration intercept of 1.02 (95% CI 0.50–1.55), slope of 0.74 (95% CI 0.29–1.19), and *E*/*O*-ratio of 0.96 (95% CI 0.84–1.09), with a corresponding Hosmer–Lemeshow $p = 0.004$ (Fig. 2).

For prediction of MCID in NRS-LP in our cohort, we found a calibration intercept of − 0.77 (95% CI − 1.32 to − 0.23) and slope of 1.29 (95% CI 0.75–1.84). The Hosmer–Lemeshow *p* value was 0.034. Overall, the model appears to overestimate the probability of a favourable

Table 1 Baseline patient characteristics of the development and external validation cohorts

Parameter	Development cohort[a]	External validation cohort
Age, mean ± SD	61.3 ± 12.5	50.4 ± 11.4
Male gender, *n* (%)	639 (40)	51 (51)
ASA score ≥ 3, *n* (%)	510 (32)	2 (2)
Smoking status, *n* (%)		
Current smoker	205 (13)	30 (30)
Previous	607 (38)	18 (18)
Never	721 (46)	52 (52)
Unknown	50 (3)	0 (0)
Medicaid, *n* (%)	131 (8)	100 (100)
Caucasian ethnicity, *n* (%)	1422 (90)	94 (94)
Opioid consumption, *n* (%)	889 (56)	25 (25)
Asthma, *n* (%)	219 (14)	1 (1)
Prior spine surgery, *n* (%)	395 (25)	22 (22)
Diagnosis, *n* (%)		
Spondylolisthesis	1033 (65)	79 (79)
Disc herniation	220 (14)	8 (8)
FBSS	238 (15)	14 (14)
Stenosis	1223 (77)	46 (46)
Pseudarthrosis	75 (5)	0 (0)
Radiculopathy	1461 (92)	5 (5)
DDD	473 (30)	35 (35)
Surgical approach, *n* (%)		
MI-TLIF	N.R.	62 (62)
MI-PLIF	N.R.	38 (38)

SD standard deviation, *ASA* American Society of Anesthesiologists, *FBSS* failed back surgery syndrome, *DDD* degenerative disc disease, *MI-TLIF* minimally invasive transforaminal lumbar interbody fusion, *MI-PLIF* minimally invasive posterior lumbar interbody fusion, *N.R.*, not reported

[a] The patient characteristics of the development cohort are provided for comparison, and are taken from the original report of the SCOAP-CERTAIN model (Khor et al. [2])

Table 2 Tabulation of outcome measures in the development and external validation cohorts

Parameter	Development cohort[a]		External validation cohort	
PROMs	Baseline	12 months	Baseline	12 months
ODI				
Number of pts.	783	545	100	100
Median (range)	46 (2–100)	24 (0–90)	47 (12–96)	12 (0–60)
0–20, *n* (%)	55 (7)	248 (46)	5 (5)	70 (70)
21–40, *n* (%)	266 (34)	157 (29)	30 (30)	20 (20)
41–60, *n* (%)	307 (39)	109 (20)	48 (48)	10 (10)
61–100, *n* (%)	155 (20)	31 (6)	17 (17)	0 (0)
MCID achieved, *n* (%)	–	306 (58)	–	73 (73)
NRS-BP				
Number of pts.	1466	933	100	100
Median (range)	6 (0–10)	3 (0–10)	7 (0–10)	2 (0–10)
0–2, *n* (%)	229 (16)	565 (61)	8 (8)	54 (54)
3–6, *n* (%)	516 (35)	251 (27)	27 (27)	30 (30)
7–10, *n* (%)	712 (49)	117 (13)	65 (65)	16 (16)
MCID achieved, *n* (%)	–	616 (69)	–	77 (77)
NRS-LP				
Number of pts.	726	508	100	100
Median (range)	6 (0–10)	1 (0–10)	7 (0–10)	1 (0–10)
0–2, *n* (%)	143 (20)	345 (68)	11 (11)	64 (64)
3–6, *n* (%)	254 (35)	104 (21)	25 (25)	31 (31)
7–10, *n* (%)	329 (45)	59 (12)	64 (64)	5 (5)
MCID achieved, *n* (%)	–	355 (77)	–	76 (76)

PROMs patient-reported outcome measures, *ODI* Oswestry disability index, *NRS-BP* numeric rating scale for back pain, *NRS-LP* numeric rating scale for leg pain, *MCID* minimum clinically important difference

[a] These data are provided for comparison, and are taken from the original report of the SCOAP-CERTAIN model (Khor et al. [2])

outcome when looking at the calibration plot (Fig. 3) and the high *E*/*O*-ratio of 1.20 (95% CI 1.10–1.32).

Discrimination

A detailed overview of discrimination measures is provided in Table 4. Figure 4 demonstrates AUC curves for the three models at external validation. For prediction of ODI, we observed an AUC of 0.71 (95% CI 0.58–0.81), sensitivity of 0.64 (95% CI 0.53–0.75), and specificity of 0.65 (95% CI 0.46–0.83).

Similarly, for prediction of NRS-BP, AUC values of 0.72 (95% CI 0.59–0.83), sensitivity of 0.81 (95% CI 0.71–0.89), and specificity of 0.48 (95% CI 0.26–0.68) were identified at external validation.

Finally, prediction of NRS-LP yielded an AUC of 0.83 (95% CI 0.72–0.94), sensitivity of 1.00 (95% CI 1.00–1.00), and specificity of 0.38 (95% CI 0.17–0.57).

Overall, these values correspond well to those observed in the derivation cohort [2]. This means that the SCOAP-CERTAIN tool generalizes well to new patient data even in other cohorts with differing demographics and indications, especially in terms of its ability to binarily predict which patient will achieve a favourable outcome.

Table 3 Calibration performance metrics of the three prediction models on external data

Calibration metric	12-month MCID		
	ODI	NRS-BP	NRS-LP
Calibration intercept	1.08 (0.60–1.57)	1.02 (0.50–1.55)	− 0.77 (− 1.32 to − 0.23)
Calibration slope	0.95 (0.37–1.54)	0.74 (0.29–1.19)	1.29 (0.75–1.84)
Expected/Observed ratio[a]	0.77 (0.63–0.90)	0.96 (0.84–1.09)	1.20 (1.10–1.32)
Brier score[b]	0.22 (0.19–0.25)	0.19 (0.15–0.23)	0.12 (0.07–0.17)
Estimated calibration index[c]	0.41 (0.17–0.64)	0.44 (0.19–0.66)	0.67 (0.42–0.87)
Hosmer–Lemeshow *p*	0.002	0.004	0.034

Where applicable, bootstrapped 95% confidence intervals are provided

MCID minimum clinically important difference, *ODI* Oswestry disability index, *NRS-BP* numeric rating scale for back pain, *NRS-LP* numeric rating scale for leg pain

[a]The expected/observed ratio, or *E*/*O*-ratio, describes the overall calibration of a prediction model, and is defined as the ratio of expected positive (predicted positive) cases and observed positive (true positive) cases. A value of 1 is optimal

[b]The Brier score measures overall calibration and is defined as the average squared difference between predicted probabilities and true outcomes. It takes on values between 0 and 1, with lower values indicating better calibration

[c]The Estimated Calibration Index (ECI) is a measure of overall calibration, and is defined as the average squared difference of the predicted probabilities with their grouped estimated observed probabilities. It can range between 0 and 100, with lower values representing better overall calibration

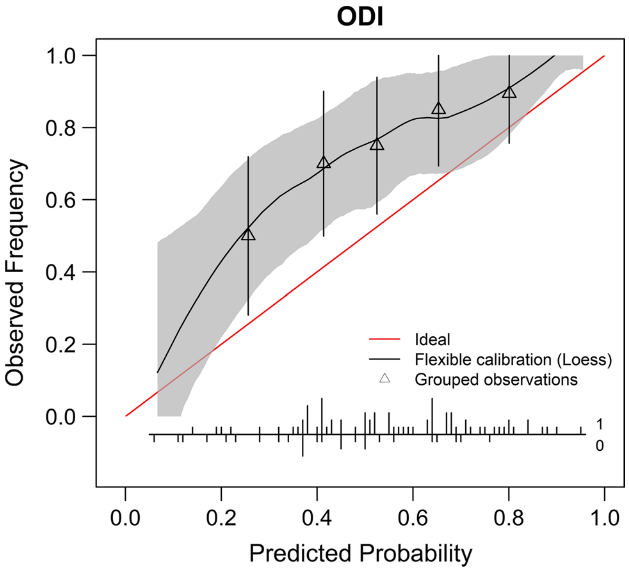


Fig. 1 Calibration plots for prediction of improvement in 12-month Oswestry Disability Index. Calibration intercept and slope were 1.08 and 0.95, respectively. *ODI* Oswestry disability index, *LOESS* locally estimated scatterplot smoothing

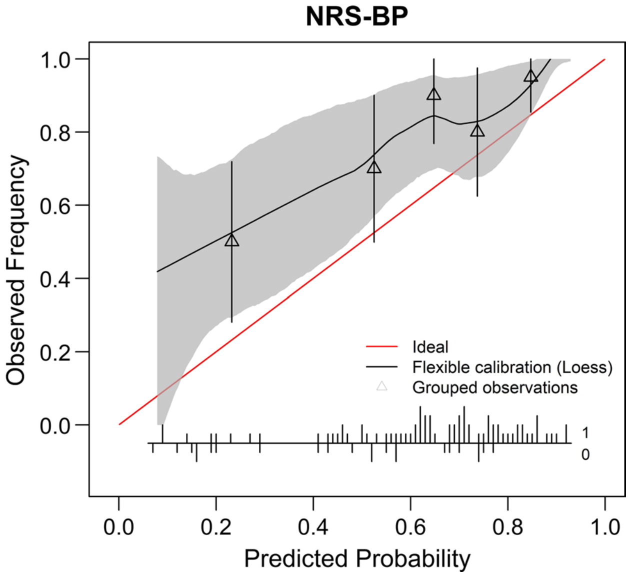


Fig. 2 Calibration plots for prediction of improvement in 12-month back pain. Calibration intercept and slope were 1.02 and 0.74, respectively. *NRS-BP* numeric rating scale for back pain, *LOESS* locally estimated scatterplot smoothing

Discussion

We carried out external validation of the SCOAP-CERTAIN models proposed by Khor et al. [2] on a cohort of 100 patients. We found good generalization of the models' discriminative ability, comparable to the values observed in the derivation cohort. This means that the tool is accurate in binarily predicting which patients will achieve a favourable clinical outcome, defined as the MCID. However, calibration and goodness-of-fit were poor for the three models for MCID in 12-month functional disability and back or leg pain. Thus, the tool was overall less accurate in generating predictions on how likely a favourable clinical outcome is.

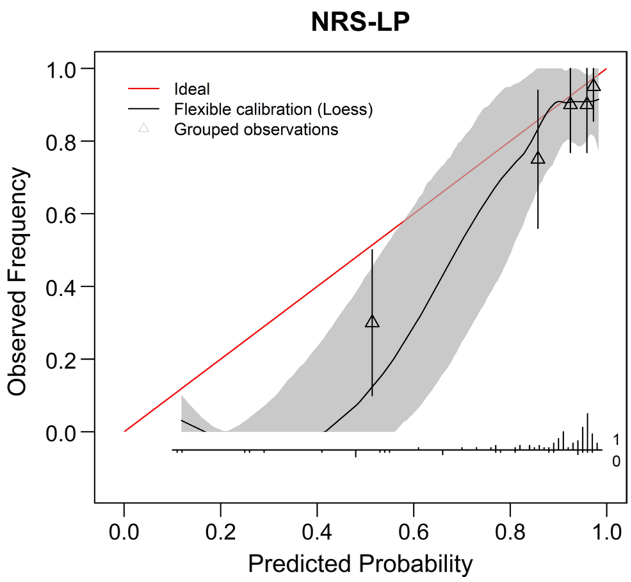


Fig. 3 Calibration plots for prediction of improvement in 12-month leg pain. Calibration intercept and slope were −0.77 and 1.29, respectively. *NRS-LP* numeric rating scale for leg pain, *LOESS* locally estimated scatterplot smoothing

In recent years, the medical profession has focused on delivery of much more individualized patient care, and PROMs are now viewed as an integral part of healthcare assessment, often replacing radiological outcome measures such as bony fusion [27, 28]. An algorithm that may potentially combine factors associated with PROMs after elective lumbar fusion, and predict MCID in these outcome measures, has the potential to greatly benefit patient care. In spine surgery, individualized prognosis based on prediction models has not yet seen routine clinical use, and there is a lack of properly validated models for this purpose [1–4]. A model that may allow the patient and the physician alike to estimate the chance of improvement after surgery would promote shared decision-making, allow for objective risk–benefit estimation on an individualized level, and potentially even account for risk factors for poor outcome preoperatively [1].

Surgical prediction tools may not only have the potential to help patients and physicians, but also the healthcare system by enhancing the cost-effectiveness of procedures. By estimating treatment effects before surgery, unnecessary and ineffective procedures can potentially be avoided, and patients can enjoy a more realistic and quantifiable prognostic assessment. In addition, adverse events can be anticipated, which can improve their management and, potentially, even allow their prevention in the first place [5, 29–32]. However, the outputs of prediction tools should never be considered absolute when reaching a decision. They should not trump a physician's clinical judgement, but rather be used only as an adjunct to the process of patient counselling and decision-making.

Many models based on ML or on statistical modelling techniques such as logistic regression are currently being published [7]. However, application of these models can only safely be considered after external validation in at least one centre outside of the development cohort. Most models remain only internally validated. While internal validation can provide some insights as to the generalizability of a model, all conclusions can only be made regarding the population of the derivation centre. It has been observed that even subtle differences in patient demographics or the incidence of the predicted outcome among cohorts can greatly bias predictions, or even render them useless [10, 22, 33, 34].

In addition, overfitting of the model to the development cohort can only reliably be detected after external validation, or at a minimum, prospective internal validation. Overfitting occurs when a model too closely approximates the development data—thus "learning by heart" the features

Table 4 Discrimination performance metrics of the three prediction models on external data

Discrimination metric	12-month MCID		
	ODI	NRS-BP	NRS-LP
AUC	0.71 (0.58–0.81)	0.72 (0.59–0.83)	0.83 (0.72–0.94)
Accuracy	0.65 (0.55–0.73)	0.73 (0.64–0.82)	0.85 (0.77–0.92)
Sensitivity	0.64 (0.53–075)	0.81 (0.71–0.89)	1.00 (1.00–1.00)
Specificity	0.65 (0.46–0.83)	0.48 (0.26–0.68)	0.38 (0.17–0.57)
PPV	0.84 (0.73–0.93)	0.84 (0.75–0.92)	0.84 (0.75–0.91)
NPV	0.40 (0.26–0.54)	0.42 (0.22–0.62)	1.00 (1.00–1.00)
F1 score[a]	0.49 (0.34–0.62)	0.44 (0.24–0.61)	0.54 (0.30–0.72)

Bootstrapped 95% confidence intervals are provided

MCID minimum clinically important difference, *AUC* area under the receiver operating characteristics curve, *PPV* positive predictive value, *NPV* negative predictive value, *ODI* Oswestry disability index, *NRS-BP* numeric rating scale for back pain, *NRS-LP* numeric rating scale for leg pain

[a] The F1 score is a composite metric, and is mathematically defined as the harmonic mean of PPV and sensitivity. Higher values represent better performance, with a maximum of 1

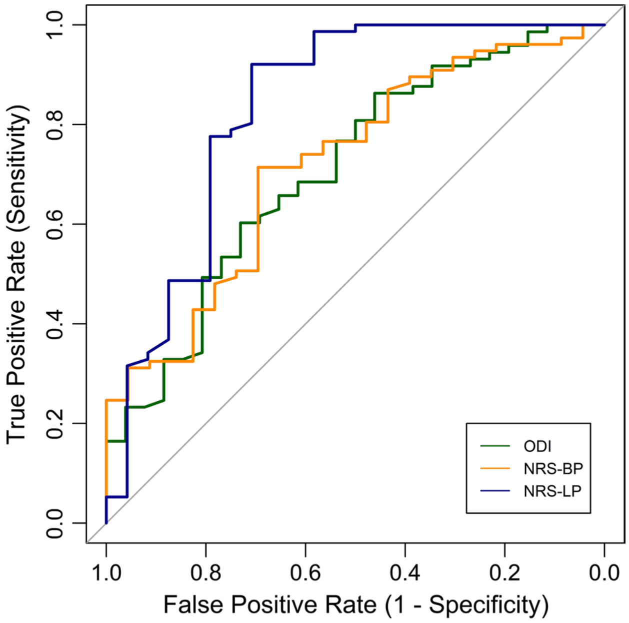


Fig. 4 Area under the curve (AUC) values of the three models for prediction of MCID in the 12-month outcome, assessed on 100 external patients. The observed AUC values were 0.71, 0.72, and 0.83 for achieving MCID in functional impairment, back pain, and leg pain, respectively. *ODI* Oswestry disability index, *NRS-BP* numeric rating scale for back pain, *NRS-LP* numeric rating scale for leg pain

of the patients in the development set [3]. If this occurs, the model will not generalize well to new patients, and any predictions made will be based solely on the memorized derivation set patients, instead of on generalizable, extracted features. The value of external validation lies within the fact that performance on new patients, unseen by the model, can be assessed, representing the situation of clinical application of a certain model.

Discrimination and calibration should be assessed. Discrimination denotes the ability of a model to accurately classify patients into those who experience MCID and those who do not. In contrast, the ability of a model to produce predicted probabilities that closely correlate to the true posteriors (observed frequency) is referred to as calibration.

Overall, the SCOAP-CERTAIN models displayed a good discrimination, with comparable values to those observed in the derivation cohort, with AUC values ranging from 0.66 to 0.79 [2]. However, calibration, as assessed by various metrics, was poor at external validation. For internal validation, Khor et al. report calibration intercepts of − 0.02 to 0.16, and slopes of 0.80–1.05 [2]. At external validation, we found that generally, calibration intercepts were within acceptable ranges, comparable to the internal validation cohort. The observed calibration slopes, however, demonstrated large heterogeneity, except for the model for MCID in the ODI. Testing for goodness-of-fit using the method described by Hosmer and Lemeshow corroborated the findings of generally poor calibration compared to the development cohort [23]. In addition, the calibration plots indicate shifts in overall calibration ("calibration-in-the-large"), which are corroborated by the calibration intercepts and *E*/*O*-ratios that were observed. This is especially true for the model predicting MCID in NRS-LP, which demonstrated to be overestimating the probability of a favourable outcome, predicting MCID far too often compared to what was truly observed. For clinical prediction models, calibration may play an arguably even more important role than discrimination alone, because clinicians and patients are usually not primarily interested in, e.g. a binary classification, but instead in the predicted probabilities of a certain endpoint [13, 35]. Therefore, poor calibration represents a major impediment to clinical and external applicability of prediction models. However, there are techniques that may help improve calibration. First, over- or underestimating models can be improved by simply adjusting their intercepts [34]. Second, whenever uniform deformations of the calibration curves are observed across all resamples during cross-validation or bootstrapping, rescaling of the predicted probabilities using Platt scaling or isotonic regression is possible [35].

When developing prediction models, taking into account class imbalance is crucial [36]. Class imbalance is present in binary classification tasks whenever one class (majority class) significantly outnumbers the other class (minority class). For example, when predicting a complication that occurs in only 10% of patients (minority class), even a zero-information model always voting for the majority class (no complication) will achieve an AUC of approximately 0.90, and accuracy of 90%, with high specificity but unemployable sensitivity [36]. The pernicious effects of class imbalance can be diagnosed by comparing sensitivity and specificity, or PPV and NPV. To force a model to actually extract generalizable features from imbalanced data, instead of simply always voting for the majority class, techniques such as random oversampling or synthetic minority oversampling (SMOTE) should be applied to prevent unbalanced models [36, 37].

Khor et al. did not report the sensitivity, specificity, PPV, or NPV for their model. In our cohort, we found that sensitivity for prediction of NRS-BP and NRS-LP was fair, while the specificity was poor. Both sensitivity and specificity for prediction of MCID in ODI were satisfactory. From our data, it can be concluded that the SCOAP-CERTAIN tool has high power to rule out MCID in NRS-BP and NRS-LP at 12 months but should not be used to rule in MCID.

One likely reason for the differences in performance measures among the internal and external validation cohorts is the difference in endpoint incidence. The rates of MCID according to the definition by Khor et al. were higher in our cohort than in the development cohort, where MCID was achieved in 58.0–76.5% of patients [2]. It has been previously observed that differences in the incidence of the binary endpoint may distort calibration [33]. If necessary, models with large intercepts can be recalibrated using the techniques mentioned above [34, 35].

The reduced generalizability may also be explained by unclear or differing definitions for some of the input variables. While most variables can be uniformly defined, insurance status will have different definitions in different countries around the world. Since most countries, and even some provinces, have different insurance systems, it will be difficult for the user to decide what to choose as their insurance status when using this web-based tool. We chose the definition most closely resembling the input variable, applicable to the Dutch healthcare system. Khor et al. do not give detailed information on the invasiveness of their fusion procedures. In our cohort, most procedures were carried out in a minimally invasive fashion, which could potentially also explain some of the differences in performance between the cohorts. However, it has to be considered that, while there is some evidence that minimally invasive procedures reduce immediate post-operative pain and boost early recovery, there seems to be little to no effect on the long-term patient-reported outcome after lumbar fusion [38, 39].

In addition, our cohort included a large proportion of patients with CLBP due to DDD, as did the development cohort [2]. Still, the authors of the SCOAP-CERTAIN model excluded the presence of DDD as an input variable in the final model, and thus any even minor differences in the proportion of patients with DDD may bias predictions, as this factor is not being corrected for.

Limitations

Due to local insurance policy, patients aged over 80 years, with ASA classes over 3, and with a BMI over 33 are not allowed to undergo elective spine surgery in our short-stay setting [16]. For this reason, such patients were not available in our registry. This means that any findings as to the external validity of the model may not be extrapolated to these higher-risk patients. As expected, although our patients represent the exact patient population that the SCOAP-CERTAIN model has been developed for, the patient characteristics observed in our cohort differed in some cases from those observed in the derivation cohort. For example, in the current external validation cohort, the proportion of patients with radiculopathy was relevantly lower than in the derivation cohort. It is conceivable that at least part of the overestimation seen in the NRS-LP model could be explained by this difference in indications. In addition, patients in the external validation cohort were around a decade younger than in the original Khor et al. report [2]. However, as the deviations are within the expected variation of patient demographics, indications for surgery, and outcomes achieved seen among different surgical populations, our study represents a realistic use-case in which the SCOAP-CERTAIN model could be applied clinically. In addition, the original study did not report on how exactly PROMs were recorded, e.g. paper-based or web-based. The method of data collection could influence the observed outcomes [40]. Our study included a cohort of 100 consecutive patients taken from a prospective registry. While this sample size is usually sufficient for external validation of a prediction model, a larger sample size can often lead to more smooth calibration plots at graphical and statistical assessment [15, 23].

Conclusions

Using data from a prospective registry, we externally validated the SCOAP-CERTAIN prediction model. We conclude that the prediction tool generally had fair discrimination at external validation, with performance measures corresponding closely to those observed in the development cohort. However, calibration of the predicted probabilities was poor. As the predicted probabilities are arguably of greater interest to clinicians than binary classifications, and because the calibration of the prediction tool was poor, it may be premature to apply the SCOAP-CERTAIN prediction tool in clinical practice in its current form. We suggest that any prediction tool should first be externally validated before it is applied in routine clinical practice. We also suggest that future studies, whether carrying out external or internal validation, should ideally report sensitivity, specificity, PPV, and NPV of the assessed models.

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 they have no conflict of interest.

References

1. Steinmetz MP, Mroz T (2018) Value of adding predictive clinical decision tools to spine surgery. JAMA Surg. https://doi.org/10.1001/jamasurg.2018.0078
2. Khor S, Lavallee D, Cizik AM et al (2018) Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. JAMA Surg. https://doi.org/10.1001/jamasurg.2018.0072
3. 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

4. Staartjes VE, de Wispelaere MP, Vandertop WP, Schröder ML (2018) Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar diskectomy: feasibility of center-specific modeling. Spine J Off J North Am Spine Soc. https://doi.org/10.1016/j.spinee.2018.11.009
5. Janssen DMC, van Kuijk SMJ, d'Aumerie B, Willems P (2019) A prediction model of surgical site infection after instrumented thoracolumbar spine surgery in adults. Eur Spine J. https://doi.org/10.1007/s00586-018-05877-z
6. Janssen DMC, van Kuijk SMJ, d'Aumerie BB, Willems PC (2018) External validation of a prediction model for surgical site infection after thoracolumbar spine surgery in a Western European cohort. J Orthop Surg. https://doi.org/10.1186/s13018-018-0821-2
7. 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
8. Brusko GD, Kolcun JPG, Wang MY (2018) Machine-learning models: the future of predictive analytics in neurosurgery. Neurosurgery 83:E3–E4. https://doi.org/10.1093/neuros/nyy166
9. Cabitza F, Rasoini R, Gensini GF (2017) Unintended consequences of machine learning in medicine. JAMA 318:517–518. https://doi.org/10.1001/jama.2017.7797
10. Collins GS, Ogundimu EO, Le Manach Y (2015) Assessing calibration in an external validation study. Spine J 15:2446–2447. https://doi.org/10.1016/j.spinee.2015.06.043
11. Debray TPA, Vergouwe Y, Koffijberg H et al (2015) A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 68:279–289. https://doi.org/10.1016/j.jclinepi.2014.06.018
12. Tetreault LA, Côté P, Kopjar B et al (2015) A clinical prediction model to assess surgical outcome in patients with cervical spondylotic myelopathy: internal and external validations using the prospective multicenter AOSpine North American and international datasets of 743 patients. Spine J 15:388–397. https://doi.org/10.1016/j.spinee.2014.12.145
13. Riley RD, Ensor J, Snell KIE et al (2016) External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 353:i3140. https://doi.org/10.1136/bmj.i3140
14. Collins GS, de Groot JA, Dutton S et al (2014) External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 14:40. https://doi.org/10.1186/1471-2288-14-40
15. Collins GS, Ogundimu EO, Altman DG (2016) Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 35:214–226. https://doi.org/10.1002/sim.6787
16. Staartjes VE, de Wispelaere MP, Schröder ML (2019) Improving recovery after elective degenerative spine surgery: 5-year experience with an enhanced recovery after surgery (ERAS) protocol. Neurosurg Focus 46:E7. https://doi.org/10.3171/2019.1.FOCUS18646
17. Schröder ML, Staartjes VE (2017) Revisions for screw malposition and clinical outcomes after robot-guided lumbar fusion for spondylolisthesis. Neurosurg Focus 42:E12. https://doi.org/10.3171/2017.3.FOCUS16534
18. Staartjes VE, Schröder ML (2018) Effectiveness of a decision-making protocol for the surgical treatment of lumbar stenosis with grade 1 degenerative spondylolisthesis. World Neurosurg 110:e355–e361. https://doi.org/10.1016/j.wneu.2017.11.001
19. 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
20. Staartjes VE, Siccoli A, de Wispelaere MP, Schröder ML (2018) Patient-reported outcomes unbiased by length of follow-up after lumbar degenerative spine surgery: do we need 2 years of follow-up? Spine J Off J North Am Spine Soc. https://doi.org/10.1016/j.spinee.2018.10.004
21. Van Hooff ML, Spruit M, Fairbank JCT et al (2015) The Oswestry Disability Index (version 2.1a): validation of a Dutch language version. Spine 40:E83–E90. https://doi.org/10.1097/BRS.0000000000000683
22. Steyerberg EW, Vickers AJ, Cook NR et al (2010) Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiol Camb Mass 21:128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2
23. Hosmer DW, Lemeshow S, Sturdivant RX (2013) Assessing the fit of the model. In: Hosmer DW Jr, Lemeshow S, Sturdivant RX (eds) Applied logistic regression. Wiley, Hoboken, pp 153–225
24. Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78:1–3. https://doi.org/10.1175/1520-0493(1950)078%3c0001:VOFEIT%3e2.0.CO;2
25. Van Hoorde K, Van Huffel S, Timmerman D et al (2015) A spline-based tool to assess and visualize the calibration of multiclass risk predictions. J Biomed Inform 54:283–293. https://doi.org/10.1016/j.jbi.2014.12.016
26. Core Team R (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
27. Falavigna A, Dozza DC, Teles AR et al (2017) Current status of worldwide use of patient-reported outcome measures (PROMs) in spine care. World Neurosurg 108:328–335. https://doi.org/10.1016/j.wneu.2017.09.002
28. Glassman SD, Schwab F, Bridwell KH et al (2009) Do 1-year outcomes predict 2-year outcomes for adult deformity surgery? Spine J Off J North Am Spine Soc 9:317–322. https://doi.org/10.1016/j.spinee.2008.06.450
29. van Niftrik CHB, van der Wouden F, Staartjes VE et al (2019) Machine learning algorithm identifies patients at high risk for early complications after intracranial tumor surgery: registry-based cohort study. Neurosurgery. https://doi.org/10.1093/neuros/nyz145
30. Durand WM, DePasse JM, Daniels AH (2018) Predictive modeling for blood transfusion after adult spinal deformity surgery: a tree-based machine learning approach. Spine 43:1058. https://doi.org/10.1097/BRS.0000000000002515
31. Ehlers AP, Roy SB, Khor S et al (2017) Improved risk prediction following surgery using machine learning algorithms. eGEMs. https://doi.org/10.13063/2327-9214.1278
32. Kim JS, Merrill RK, Arvind V et al (2018) Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. Spine. https://doi.org/10.1097/BRS.0000000000002442
33. van Rein EAJ, van der Sluijs R, Voskens FJ et al (2019) Development and validation of a prediction model for prehospital triage of trauma patients. JAMA Surg. https://doi.org/10.1001/jamasurg.2018.4752
34. Janssen KJM, Moons KGM, Kalkman CJ et al (2008) Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 61:76–86. https://doi.org/10.1016/j.jclinepi.2007.04.018
35. Niculescu-Mizil A, Caruana R (2005) Predicting Good Probabilities with Supervised Learning. In: Proceedings of the 22nd international conference on machine learning. ACM, New York, pp 625–632
36. Staartjes VE, Schröder ML (2018) Letter to the editor. Class imbalance in machine learning for neurosurgical outcome

prediction: are our models valid? J Neurosurg Spine. https://doi.org/10.3171/2018.5.SPINE18543

37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
38. Goldstein CL, Phillips FM, Rampersaud YR (2016) Comparative effectiveness and economic evaluations of open versus minimally invasive posterior or transforaminal lumbar interbody fusion: a systematic review. Spine 41(Suppl 8):S74–S89. https://doi.org/10.1097/BRS.0000000000001462
39. Goldstein CL, Macwan K, Sundararajan K, Rampersaud YR (2016) Perioperative outcomes and adverse events of minimally invasive versus open posterior lumbar fusion: meta-analysis and systematic review. J Neurosurg Spine 24:416–427. https://doi.org/10.3171/2015.2.SPINE14973
40. Schröder ML, de Wispelaere MP, Staartjes VE (2018) Are patient-reported outcome measures biased by method of follow-up? Evaluating paper-based and digital follow-up after lumbar fusion surgery. Spine J Off J North Am Spine Soc 2:2. https://doi.org/10.1016/j.spinee.2018.05.002

Affiliations

Ayesha Quddusi[1] · Hubert A. J. Eversdijk[2] · Anita M. Klukowska[2,3] · Marlies P. de Wispelaere[4] · Julius M. Kernbach[5] · Marc L. Schröder[2] · Victor E. Staartjes[2,6,7]

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

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

[2] Department of Neurosurgery, Bergman Clinics, Naarden, Rijksweg 69, 1411 GE Naarden, Amsterdam, The Netherlands

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

[4] Department of Clinical Informatics, Bergman Clinics, Amsterdam, The Netherlands

[5] Department of Neurosurgery, Faculty of Medicine, RWTH Aachen University, Aachen, Germany

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

[7] Machine Intelligence in Clinical Neuroscience Lab, Department of Neurosurgery, University Hospital Zurich, Clinical Neuroscience Centre, University of Zurich, Zurich, Switzerland