North American Spine Society Journal (NASSJ) 13 (2023) 100188

Contents lists available at ScienceDirect

North American Spine Society Journal (NASSJ)

journal homepage: www.elsevier.com/locate/xnsj

Clinical Studies

Validating the predictive precision of the dialogue support tool on Danish patient cohorts

Casper Friis Pedersen[a,*], Mikkel Østerheden Andersen[a], Leah Yacat Carreon[a], Søren Eiskjær[b]

[a] University of Southern Denmark, Center for Spine Surgery and Research, Spinecenter of Southern Denmark, Lillebaelt Hospital, Oestre Hougvej 55, DK-5500 Middelfart, Denmark
[b] Aalborg University, Dept. of Orthopedic Surgery, Hobrovej 18-22, DK-9000 Aalborg, Denmark

ARTICLE INFO

Keywords:
Spine surgery
Outcome
Prediction
Dialogue support
Shared decision-making
Validation

ABSTRACT

Background: Despite advances in surgical techniques and diagnostics, some patients remain unsatisfied with the result following spine surgery. One way to improve patient satisfaction may be found in better alignment of expectations. Prognostic tools might prove useful in strengthening surgeon-patient communication prior to surgery. The purpose of this study is to assess the predictive capabilities of the Swedish based Dialogue Support (DS) tool for spine surgery on a Danish population.

Methods: The study included the diagnoses lumbar disc herniation, lumbar spinal stenosis, and lumbar degenerative disc disease. A total of 5.954 patients were retrieved from the Danish national spine registry (DaneSpine). For each group, 200 random cases with complete preoperative and 1 year follow-up data were selected. Two outcome measures were used: Global assessment of pain (GA pain) and satisfaction with outcome. Predictions were produced by manual entry in the DS application. Goodness of fit tests were used to compare the predicted distribution of proportions with successful outcomes (GA pain) to the actual distribution in the three samples. Binomial tests were performed to evaluate the predicted proportion of satisfied patients. Furthermore, ROC-curves, calibration plots, and metrics were calculated to assess the predictive performance.

Results: ROC curves showed comparable AUC values with the values reported by the developing authors of the DS from 0.62 to 0.73 (GA pain) and 0.64 to 0.70 (satisfaction with outcome). The calibration plots, however, revealed a low degree of concordance. For GA pain sensitivity varied from 92.4% to 99.3%, and specificity from 1.5% to 13.4%. For satisfaction, sensitivity varied from 97.1% to 99.2% and specificity from 0.0% to 2.9%.

Conclusions: The predictive capabilities of the DS tool could not be generalized to the Danish sample cohorts. Further research on larger samples, provided full access to the underlying algorithms can be obtained, could produce a different result.

Background

Internationally, there are several large registries that collect patient-reported data to monitor patient outcomes after surgery for lumbar degenerative diseases (e.g. DaneSpine [1], Swespine [2], NORspine [3], Spine Tango [4], QOD [5]). Despite all the best efforts, there is still a considerable proportion of patients who are dissatisfied with their surgical outcome [6–10].

Even a technically successful operation based on correct indications is not a guarantee for a satisfactory result measured by patient reported outcomes measures (PROMs). The proportion of dissatisfied patients can probably be reduced by strengthening preoperative communication

FDA device/drug status: Not applicable.

Author disclosures: ***CFP:*** Nothing to disclose. ***MØA:*** Nothing to disclose. ***LYC:*** Consulting: National Spine Health Foundation (B); Orthopaedic Research Foundation (B). Scientific Advisory Board/Other Office: University of Louisville Institutional Review Board (Nonfinancial); The Spine Journal (Nonfinancial); Spine (Nonfinancial): Spine Deformity (Nonfinancial); American Spine Registry (Nonfinancial). Research Support (Investigator Salary, Staff/Materials: Pfizer (D); TSRH (B); Alan L. & Jacqueline B. Stuart Spine Research (C); Cerapedics (D); Scoliosis Research Society (E); Medtronic (D); SDU Faculty Scholarship (E); Johnson & Johnson (E); Cerapedics (F); IRSs Kursus og rejsepulje (B); TrygFonden (F); Region Syddanmark PhD puljen (E); SLB Forskningsrad (E); Sygeforsikring Donation (F); Sundhedsstyrelsen (F); SLB Forskningsrad Projektstotte (C). ***SE:*** Nothing to disclose.

* Corresponding author at: Center for Spine Surgery and Research, Spinecenter of Southern Denmark, Lillebaelt Hospital, Oestre Hougvej 55, DK-5500 Middelfart, Denmark.

E-mail address: casper.friis.pedersen@rsyd.dk (C.F. Pedersen).

https://doi.org/10.1016/j.xnsj.2022.100188
Received 23 August 2022; Received in revised form 21 November 2022; Accepted 28 November 2022
Available online 2 December 2022

between patient and surgeon. To ensure that, prior to surgery, patients have realistic expectations regarding the outcome of the surgical procedure, as well as insight into the surgical risks. To support and strengthen this communication, it would be beneficial to have support tools.

The Dialogue Support (DS) is an online application predicting outcome 1 year after surgery for spinal disorders. It has been publicly available since October 2020 courtesy of the EUROSPINE steering board and the Swespine society of spinal surgeons. The application was developed to support shared decision-making with the patient when discussing surgery for different spinal disorders. The underlying prediction models were based on pre- and postoperative data on 77.743 patients enrolled in the Swedish national spine surgery registry from 2007 to 2019 (Swespine). Included diagnosis groups were lumbar disc herniation (LDH), lumbar spinal stenosis (LSS), lumbar degenerative disc disease (DDD) and cervical radiculopathy (CR) [11]. Swespine collects patient self-reported data on demographics, comorbidity and PROMs by questionnaires. Diagnosis, type of surgery and complications are recorded by surgeons. Follow-up questionnaires are distributed and collected at 1, 2, 5 and 10 years postoperatively [12].

Sweden and Denmark are highly comparable societies in terms of culture, language, social security systems, public health care, and health insurance systems [13,14]. The Danish national spine registry, DaneSpine was acquired by the Danish Spine Society from the Swedish Society of Spinal Surgeons in 2009 [15]. DaneSpine shares the same timing of follow-up, structure regarding questionnaires and measures with Swespine. It is almost an exact copy of the latter. It is therefore feasible to investigate if the DS predictions can be generalized to a Danish population.

The object of this study was to assess the predictive capabilities of the DS at 1 year follow-up for spine surgery in terms of calibration and discrimination when applied to sample data of Danish patients from the DaneSpine registry. The DS is based on logistic regression models. To make any reasonable comparisons it was hypothesized that the population characteristics applied by the predictive algorithms would at least be approximately reflected, on average, in the Danish samples.

Methods

Patient sample

The study included the following diagnosis groups: lumbar disc herniation (LDH), lumbar spinal stenosis (LSS) and lumbar degenerative disc disease (DDD). 2.845 (LDH), 2.531 (LSS) and 578 (DDD) patients operated between 2010 and 2020 at Spine Center of Southern Denmark, Middelfart with complete preoperative and follow-up data were identified in DaneSpine [16]. For each of the three diagnosis groups the first 200 patients in a computer-generated random list (reordering) were selected from the data. The decision to use a smaller sample size of 600 patients in total was a necessity for practical reasons. Access to the underlying predictive algorithms of the DS could not be obtained from the authors. As a result, the only option left was to obtain the predictions for the Danish cohorts by manual entry in the DS tool and logging the results for each case. A time-consuming task that took approximately 4 minutes per case or 40 hours in total. This seriously limited the sample size we were able to produce.

Outcome measures

The DS predicts two outcome measures, global assessment (GA pain) [17] and satisfaction with outcome. GA pain is a Likert scale with six ordinal categories: “How is your back/leg pain today as compared to before the surgery?” where *no back/leg pain* before the surgery=0, *completely pain free*=1, *much better*=3, *unchanged*=4, *worse*=5. Patients answering *no back/leg pain* before the surgery were excluded from the predictive modeling. For both LDH and LSS the DS predicts the outcome of GA leg pain. For the DDD group, GA back pain is predicted. Satisfaction with outcome is an ordinal Likert scale with three categories (*satisfied, hesitant,* and *dissatisfied*). Satisfaction with outcome is predicted for all diagnosis groups. The DS presents the main results as *Proportion satisfied patients* and *Proportion with successful outcome*. Proportion satisfied patients is defined as satisfaction dichotomized as success (*satisfied*) and failure (*hesitant* or *dissatisfied*). Proportion with successful outcome is defined as GA pain dichotomized into success (*completely pain-free, much better*) and failure (*somewhat better, unchanged, worse*). GA pain is also presented in the DS as a pie chart with corresponding probabilities for the six ordinal categories. However, the confidence intervals for the probabilities are not presented.

Retrieval of predictions

Direct access to the underlying predictive algorithms of the DS could not be obtained. As a result, predictions for the Danish cohorts were collected by manual entry in the web-based instrument available at the following link: https://app.molnify.com/app/7wqw6owgrznr76bkaqc6l4bs7q

The variables that must be entered in the DS for predicting Proportion with successful outcome/GA pain and Proportion satisfied patients are operated levels (LSS, DDD), age, gender, employment status, disability status, retirement status, smoking status, previous spine surgery, quality of life (EQ-5D) [18], comorbidity, walking distance, duration of leg pain, duration of back pain, preoperative leg and back pain scores [19] and functional impairment ODI [20]. For each of the 600 cases, we recorded the dichotomized results of Proportion satisfied patients and Proportion with successful outcome calculated and presented by the DS. The individual probabilities presented for each of the six GA pain categories were also recorded. To compare the predicted Proportion satisfied patients and Proportion with successful outcome with the actual results, we dichotomized the follow-up data for our cohorts following the definition of outcome in the DS as previously described. In this external assessment, we have followed the TRIPOD recommendations [21] (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) as far as possible given the available information provided by Fritzell et al. The development authors of the DS were not part of this paper.

Data handling, analysis and statistics

All data handling and analysis were done in RStudio 2021.09.1 Build 372 (R Base version 4.1.2) [22,23]. The following packages were used: ggplot2, pROC [24,25]. On a group level, Chi-square goodness of fit tests were used to compare the predicted distribution of *Proportion with successful outcome* to the actual distribution among the cohorts. Binomial tests were performed to evaluate the prediction of *Proportion of satisfied patients*. On the individual level, calibration plots, ROC-curves and various performance metrics including Matthews Correlation Coefficient were produced to assess the predictive capabilities of both outcome measures [26–28]. Calibration plots illustrates the agreement between predicted probabilities and observed events. For binary outcomes the predicted probabilities can be divided into strata of equally sized groups (bins). For each bin the average predicted probability is calculated and presented on the x-axis. The corresponding estimated observed probability of the outcome is calculated from the binary encoding (0,1) in each bin and presented as the ratio of positives on the y-axis. The plotted values can be compared with a diagonal line representing perfect agreement [29]. Each datapoint of the presented calibration plots in this study represents 5% of cases on average (n = 10), out of 3 × 200 observations. Class discrimination is evaluated by ROC-curves and performance metrics. Discrimination refers to the ability of the predictive models to separate outcomes with a positive class (success) from outcomes with a negative class (failure) [30]. Success was defined with a threshold of a predictive probability of P>0.50.

Table 1a
Characteristics of the study population given as Mean (SD) or Proportions (GA pain).

	Lumbar herniated disc			Lumbar spinal stenosis			Degenerative disc disease		
	Danish cohort	**Dialogue Support cohort**	**Dif.**	**Danish cohort**	**Dialogue Support cohort**	**Dif.**	**Danish cohort**	**Dialogue Support cohort**	**Dif.**
Number of patients, (n)	200	9.571		200	21.687		200	3.367	
Age, years, mean (SD)	47.3 (11.0)	44.6 (13.0)	2.0	62.9 (11.0)	65.2 (10.5)	2.3	49.5 (10.2)	44.7 (10.3)	4.8
Gender, females, %	46.5	45.0	1.5	44.0	52.0	7.0	49.0	54.0	5.0
Unemployed, %	5.0	9.0	4.0	4.5	10.0	5.5	19.0	10.0	9.0
Disability pension, %	5.0	23.0	18.0	5.5	15.0	9.5	7.5	28.0	20.5
Smoker, %	28.0	12.0	16.0	23.5	8.0	15.5	20.0	7.0	13.0
Previous spine surgery, %	16.0	11.0	5.0	24.0	20.0	4.0	48.0	30.0	18.0
Comorbidity, %	8.5	11.0	2.5	14.5	25.0	10.5	12.0	12.0	0.0
Operated levels = 1, %				63.0	55.0	8.0	68.5	56.0	12.5
Operated levels = 2, %				28.5	31.0	2.5	20.5	40.0	19.5
Operated levels = 3, %				7.5	11.0	3.5	9.5	4.0	5.5
Operated levels = 4, %				0.5	2.0	1.5	1.0	0.0	1.0
Operated levels = 5, %				0.5	0.0	0.5	0.5	0.0	0.5
Quality of life (EQ-5D-3L), mean (SD)	0.377 (0.323)	0.270 (0.340)	0.107	0.466 (0.305)	0.370 (0.320)	0.096	0.347 (0.313)	0.350 (0.320)	0.030
Functional impairment (ODI), mean (SD)	44.5 (17.7)	47.5 (18.2)	3.0	37.6 (14.3)	41.9 (15.5)	4.3	42.9 (15.2)	42.4 (13.2)	0.5
Walking distance, 0 - 100 m, %	31.0	30.0	1.0	22.0	34.0	12.0	25.0	10.0	15.0
Walking distance, 100 - 500 m, %	26.0	20.0	6.0	37.0	30.0	7.0	15.5	17.0	1.5
Walking distance, 0.5 - 1 km, %	17.0	15.0	2.0	17.0	16.0	1.0	29.0	21.0	8.0
Walking distance, > 1 km, %	26.0	35.0	9.0	24.0	19.0	5.0	30.5	52.0	21.5
Duration of pain in legs, No pain, %	1.5	0.0	1.5	1.0	2.0	1.0	10.5	23.0	12.5
Duration of pain in legs, < 3 months, %	30.5	16.0	14.5	6.5	2.0	4.5	9.5	2.0	7.5
Duration of pain in legs, 3 - 12 months, %	52.0	58.0	6.0	38.0	28.0	10.0	27.5	14.0	13.5
Duration of pain in legs, 1 - 2 years, %	9.5	14.0	4.5	27.5	28.0	0.5	24.5	18.0	6.5
Duration of pain in legs, > 2 years, %	6.5	12.0	5.5	27.0	40.0	13.0	28.0	44.0	16.0
Duration of pain in back, No pain, %	6.5	6.0	0.5	7.0	5.0	2.0	6.0	0.0	6.0
Duration of pain in back, < 3 months, %	17	11.0	6.0	4.5	2.0	2.5	3.0	0.0	3.0
Duration of pain in back, 3 - 12 months, %	48.0	51.0	3.0	25.0	20.0	5.0	20.5	9.0	11.5
Duration of pain in back, 1 - 2 years, %	11.5	14.0	2.5	21.5	22.0	0.5	18.5	17.0	1.5
Duration of pain in back, > 2 years, %	17.0	18.0	1.0	42.0	52.0	10.0	52.0	74.0	22.0
Preoperative VAS pain (legs), mean (SD)	64.1 (24.1)	69.0 (22.6)	4.9	65.8 (21.3)	66.2 (22.7)	0.4	54.9 (28.9)	38.4 (28.6)	16.5
Preoperative VAS pain (back), mean (SD)	43.4 (28.1)	47.4 (28,6)	4.0	53.7 (26.2)	57.3 (26.0)	3.6	58.9 (25.0)	64.5 (19.6)	5.6
GA pain, completely disappeared	31.0	37.0	6.0	27.5	28.0	0.5	19.2	0.0	19.2
GA pain, much improved	36.0	40.0	4.0	32.0	32.0	0.0	30.3	0.0	30.3
GA pain, somewhat improved	22.0	13.0	9.0	19.0	18.0	1.0	23.2	0.0	23.2
GA pain, unchanged	6.0	5.0	1.0	11.0	13.0	2.0	15.7	0.0	15.7
GA pain, worse	5.0	4.0	1.0	10.5	10.0	0.5	11.6	0.0	11.6

Abbreviations: SD, standard deviation; ODI, Oswestry Disability Index; VAS, visual analogue pain scale.

Results

Study population characteristics

In the following, the complete eligible study population of the DS model development data across diagnosis is compared to the Danish validation cohorts for GA pain (Table 1a) and proportion satisfied patients (Table 1b).

Apart from VAS leg pain in the DDD group, both Quality of life (EQ-5D-3L), ODI and VAS baselines were quite similar in the Danish sample and the DS modelling data. There were however far more smokers and previously operated in the Danish sample across groups, but fewer disability pensioners. The distribution of GA pain results at 1 year follow-up were quite similar with the exception being the DDD group where Fritzell et al. reports the distribution as 0% in all five categories. In the DDD group fewer patients in the Danish cohort were satisfied at 1 year follow-up (Tables 1a & 1b).

Outcome at 1 year

In Table 2, outcome at 1-year follow-up for the Danish validation cohorts is presented.

The Danish cohorts achieved mean improvements across groups at 1 year follow-up on EQ-5D-3L, walking distance, VAS pain (leg/back) and ODI.

Model performance

The distribution of proportion with successful outcome/GA pain among LDH patients differs from the predicted distribution: $\chi^2 = 16.59$, df = 4, p=0.002. The distribution of proportion with successful outcome/GA pain among LSS patients was statistically significantly different from the predicted distribution: $\chi^2 = 16.592$, df = 4, p=0.002. The distribution of proportion with successful outcome/GA pain among DDD patients was statistically significantly different from the predicted distribution: $\chi^2 = 48.291$, df = 4, p<0.000.

The proportion of satisfied LDH patients was 73.5% (CI: 66.8;79.5%). This was statistically different (p=0.034) from the predicted target of 79.8%. The proportion of satisfied LSS patients was 70.0% (CI: 63.1;76.3%). This was not statistically different (p=0.583) from the predicted target of 71.7%. The proportion of satisfied DDD patients was 59.5% (CI: 52.3;66.4%). This was statistically different (p<0.000) from the predicted target of 77.3%. Detailed proportions can be found in the appendix (Table 5 & 6).

Quality of class probabilities

The calibration plots compare how well the predicted probabilities of the models fits the relative frequencies of the observed outcomes when ranked and grouped in bins. The diagonal lines represent perfect fit. Points below the diagonal can be interpreted as predicted probabili-

Table 1b
Characteristics of the study population given as Mean (SD) or Proportions (Satisfaction).

	Lumbar herniated disc			Lumbar spinal stenosis			Degenerative disc disease		
	Danish cohort	**Dialogue Support cohort**	**Dif.**	**Danish cohort**	**Dialogue Support cohort**	**Dif.**	**Danish cohort**	**Dialogue Support cohort**	**Dif.**
Number of patients, (n)	200	9.721		200	22.522		200	3.443	
Age, years, mean (SD)	47.3 (11.0)	44.7 (13.1)	2.6	62.9 (11.0)	65.5 (10.5)	2.6	49.5 (10.2)	44.8 (10.4)	4.7
Gender, females, %	46.5	45.0	1.5	44.0	52.0	8.0	49.0	55.0	6.0
Unemployed, %	5.0	9.0	4.0	4.5	10.0	5.5	19.0	10.0	9.0
Disability pension, %	5.0	23.0	18.0	5.5	15.0	9.5	7.5	28.0	20.5
Smoker, %	28.0	12.0	16.0	23.5	8.0	15.5	20.0	7.0	13.0
Previous spine surgery, %	16.0	11.0	5.0	24.0	19.0	5.0	48.0	29.0	19.0
Comorbidity, %	8.5	11.0	2.5	14.5	25.0	10.5	12.0	12.0	0.0
Operated levels = 1, %				63.0	55.0	8.0	68.5	56.0	12.5
Operated levels = 2, %				28.5	31.0	2.5	20.5	39.0	18.5
Operated levels = 3, %				7.5	11.0	3.5	9.5	4.0	5.5
Operated levels = 4, %				0.5	2.0	1.5	1.0	0.0	1.0
Operated levels = 5, %				0.5	0.0	0.5	0.5	0.0	0.5
Quality of life (EQ-5D-3L), mean (SD)	0.377 (0.323)	0.270 (0.340)	0.107	0.466 (0.305)	0.380 (0.320)	0.086	0.347 (0.313)	0.350 (0.320)	0.003
Functional impairment (ODI), mean (SD)	44.5 (17.7)	47.5 (18.3)	3.0	37.6 (14.3)	41.5 (15.6)	3.9	42.9 (15.2)	42.3 (13.7)	0.6
Walking distance, 0 - 100 m, %	31.0	30.0	1.0	22.0	34.0	12.0	25.0	10.0	15.0
Walking distance, 100 - 500 m, %	26.0	20.0	6.0	37.0	30.0	7.0	15.5	17.0	1.5
Walking distance, 0.5 - 1 km, %	17.0	15.0	2.0	17.0	16.0	1.0	29.0	21.0	8.0
Walking distance, > 1 km, %	26.0	35.0	9.0	24.0	20.0	4.0	30.5	53.0	22.5
Duration of pain in legs, No pain, %	1.5	1.0	0.5	1.0	3.0	2.0	10.5	23.0	12.5
Duration of pain in legs, < 3 months, %	30.5	16.0	14.5	6.5	2.0	4.5	9.5	2.0	7.5
Duration of pain in legs, 3 - 12 months, %	52.0	58.0	6.0	38.0	28.0	10.0	27.5	14.0	13.5
Duration of pain in legs, 1 - 2 years, %	9.5	13.0	3.5	27.5	28.0	0.5	24.5	18.0	6.5
Duration of pain in legs, > 2 years, %	6.5	12.0	5.5	27.0	39.0	12.0	28.0	43.0	15.0
Duration of pain in back, No pain, %	6.5	6.0	0.5	7.0	5.0	2.0	6.0	0.0	6.0
Duration of pain in back, < 3 months, %	17	11.0	6.0	4.5	2.0	2.5	3.0	0.0	3.0
Duration of pain in back, 3 - 12 months, %	48.0	50.0	2.0	25.0	20.0	5.0	20.5	9.0	11.5
Duration of pain in back, 1 - 2 years, %	11.5	14.0	2.5	21.5	22.0	0.5	18.5	17.0	1.5
Duration of pain in back, > 2 years, %	17.0	18.0	1.0	42.0	51.0	9.0	52.0	74.0	22.0
Preoperative VAS pain (legs), mean (SD)	64.1 (24.1)	69.0 (23.1)	4.9	65.8 (21.3)	64.6 (24.0)	1.2	54.9 (28.9)	38.4 (28.7)	16.5
Preoperative VAS pain (back), mean (SD)	43.4 (28.1)	47.4 (28.6)	4.0	53.7 (26.2)	57.0 (26.2)	3.3	58.9 (25.0)	64.5 (19.7)	5.6
Satisfied	73.5	79.0	5.5	70.0	66.0	4.0	59.5	76.0	16.5

Abbreviations: SD, standard deviation; ODI, Oswestry Disability Index; VAS, visual analogue pain scale.

Table 2
Follow-up (1-year) Characteristics as Mean (SD) or Proportions of the Danish validation cohorts.

Characteristic	Disc herniation	Spinal stenosis	Degenerative disc disease
Number of patients, (n)	200	200	200
Δ Quality of life (EQ-5D-3L), mean (SD)	0.356 (0.396)	0.271 (0.357)	0.447 (0.598)
Walking distance improvement, n (%)	131 (65.5)	114 (57.0)	99 (49.5)
Δ VAS pain (legs), mean (SD)	-41.1 (34.3)	-32.7 (32.7)	-22.4 (35.6)
Δ VAS pain (back), mean (SD)	-17.3 (28.9)	-24.9 (30.8)	-20.0 (32.1)
Δ Functional impairment (ODI), mean (SD)	-26.7 (21.6)	-18.5 (17.5)	-15.0 (19.0)
Proportion with successful outcome (GA pain), n (%)	132 (66.0)	115 (57.5)	80 (40.0)
Proportion satisfied patients, n (%)	147 (73.5)	140 (70.0)	119 (59.5)

Abbreviations: SD, standard deviation; ODI, Oswestry Disability Index; VAS, visual analogue pain scale. Conventions: Δ-values are reported as the difference between postoperative outcome and preoperative outcome (Follow-up - Baseline).

ties being too large and points above as probabilities being too small, (Fig. 1).

On average the DS predicted higher probabilities of success than was observed in the Danish cohorts indicated by the majority of points below the diagonal reference line. Performance metrics (Table 3a and 3b) supports the findings illustrated by the calibration plots. The DS was poor at detecting true negatives (failures) in the Danish cohorts across diagnosis with rates ranging from 0 – 13.4%.

Table 3a
Predictive performance – Danish cohorts (GA pain).

Metrics	Disc herniation	Spinal stenosis	Degenerative disc disease
True positive rate (Sensitivity)	99.3%	92.4%	97.0%
True negative rate (Specificity)	1.5%	13.4%	12.7%
Precision	67.5%	63.2%	52.5%
Accuracy	67.5%	63.5%	55.0%
Balanced accuracy	50.4%	52.9%	54.8%
Positive likelihood ratio	1.024	1.17	1.126
Negative likelihood ratio	0,504	0.565	0.238
MCC	0.088	0.196	0.194

Abbreviations: MCC, Matthews correlation coefficient.

Table 3b
Predictive performance – Danish cohorts (Satisfaction).

Metrics	Disc herniation	Spinal stenosis	Degenerative disc disease
True positive rate (Sensitivity)	98.6%	97.1%	99.2%
True negative rate (Specificity)	2.0%	2.9%	0.0%
Precision	74.4%	70.8%	59.3%
Accuracy	74.0%	70.0%	59.0%
Balanced accuracy	50.3%	50.0%	49.6%
Positive likelihood ratio	1.046	1.041	0.992
Negative likelihood ratio	0.671	1.000	n/a
MCC	0.122	0.089	-0.058

Abbreviations: MCC, Matthews correlation coefficient.

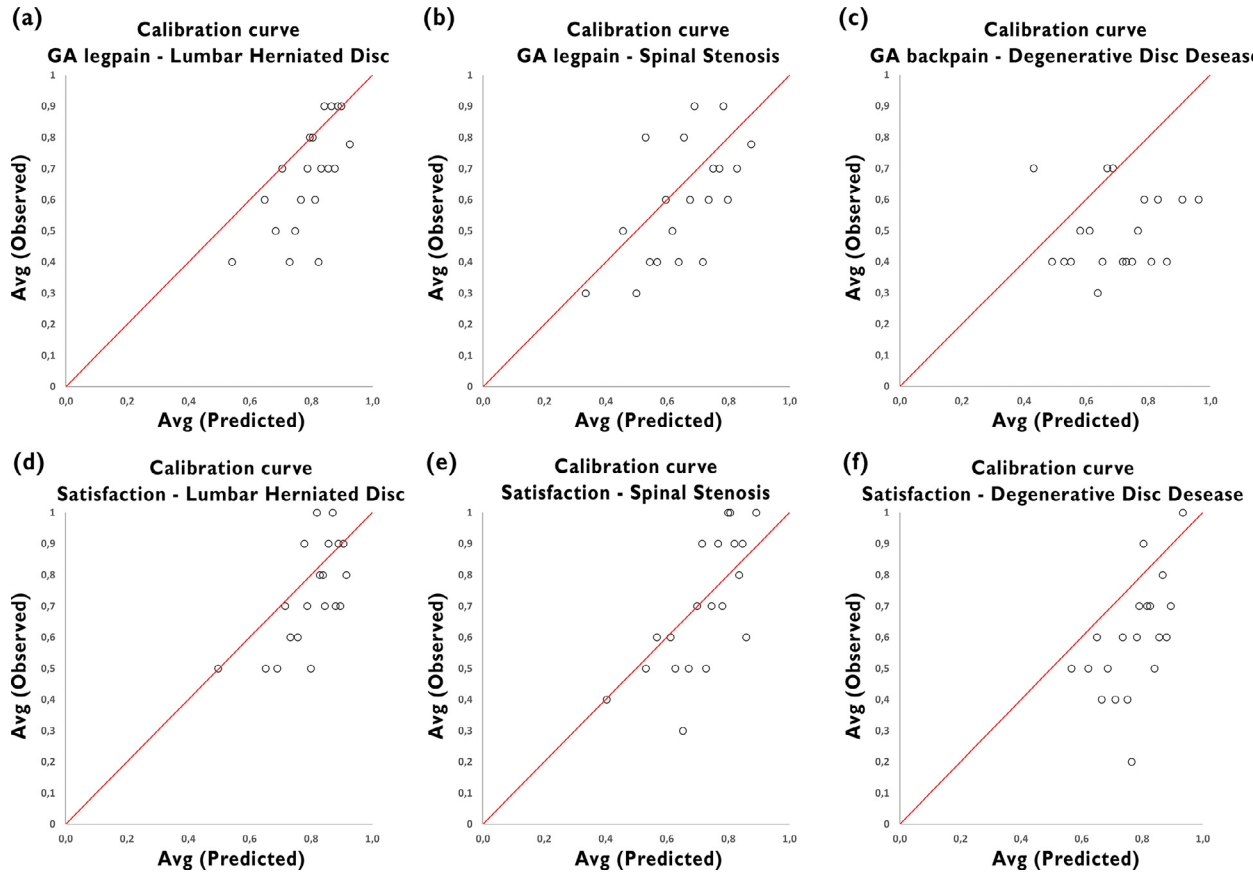

Fig. 1. Calibration plots of successful outcome/GA pain in the LDH (a), LSS (b) and DDD (c) group; of satisfaction in the LDH (d), LSS (e) and DDD (f) group. Danish validation cohorts.

Table 4
AUC values – Danish cohorts compared to the DS study.

	Lumbar herniated disc			Lumbar spinal stenosis			Degenerative disc disease		
	Danish cohort	**Dialogue Support cohort**	**Dif.**	**Danish cohort**	**Dialogue Support cohort**	**Dif.**	**Danish cohort**	**Dialogue Support cohort**	**Dif.**
GA pain	0.644	0.680	0.04	0.617	0.666	0.05	0.730	0.675	0.06
Satisfaction	0.654	0.663	0.01	0.700	0.652	0.05	0.633	0.598	0.04

Class discrimination

The ROC curves illustrate the ability of the predictive models to distinguish between true positives and false positives at different classification thresholds. The diagonal lines represent models with no discrimination ability. As seen in Fig. 2, AUC values varied across diagnosis groups and outcome from 0.62 to 0.73.

There were very small differences in the AUC values reported by Fritzell et al. and the Danish cohorts on both GA pain and satisfaction. Unfortunately, confidence intervals are not reported in the DS study.

Discussion

This is, to our knowledge, the first attempt to validate the DS on non-Swedish patients. External validation in clinical predictive modeling is paramount in assessing the stability and performance of novel algorithms [31]. This is what we aim to do – are the DS predictive models of use in a cohort of Danish patients? Fritzell et al. have devoted a substantial amount of time and effort to develop this tool and have made the web calculator available on the internet.

Our results show overall comparable ROC curves and AUC values with Fritzell et al. for both GA pain and satisfaction [11]. Fritzell et al. do not provide any performance metrics besides AUC values without confidence intervals (CI). In our samples lower bounds were not convincing, in some cases approaching random chance.

The calibration plots revealed a high degree of class imbalance favoring the positive class with the majority of points below the diagonal. The performance metrics on class discrimination confirm this. For both GA pain and satisfaction, specificity was very low. A reminder that AUC and accuracy may not always portray actual performance [32]. The algorithms largely failed to detect true negatives on an individual level with minor differences between groups. This contrasts to the calibration plots of the combined training and test data by Fritzell et al. (Fig. 2) where predicted and actual class probabilities are markedly more evenly spread around the diagonal reference line.

When comparing the predicted with the actual average distributions on the entire samples they differed significantly, with the exception of proportions of satisfied LSS patients. The latter showed a high degree of concordance with a predicted target matching actual satisfied.

The results are overall somewhat surprising. As mentioned, Sweden and Denmark are usually considered highly comparable societies. The national registries Swespine and DaneSpine are almost identical, shares the same structure, variables, and encoding schemes. Furthermore, previous comparative studies have shown very similar outcomes with no

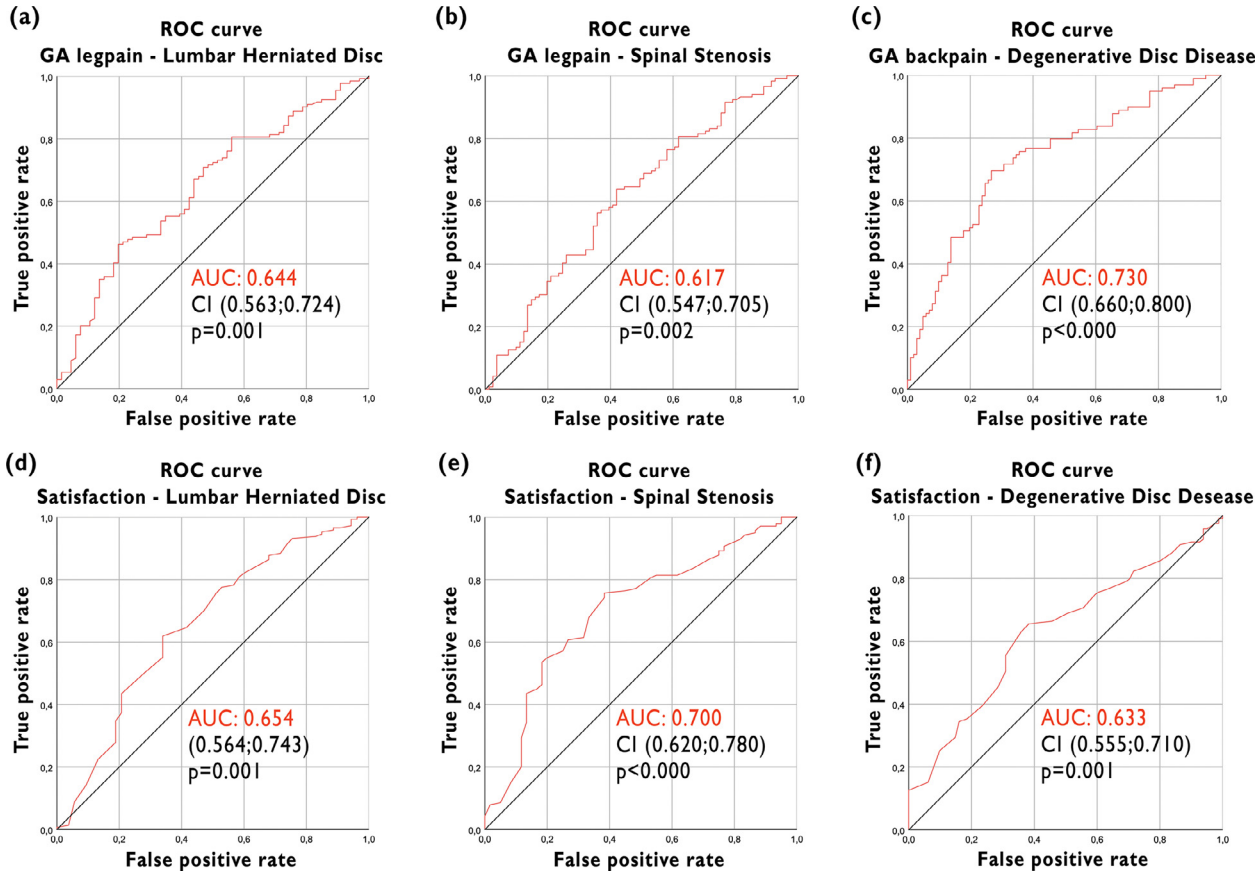

Fig. 2. ROC curves of successful outcome/GA pain in the LDH (a), LSS (b) and DDD (c) group; of satisfaction in the LDH (d), LSS (e) and DDD (f) group. Danish validation cohorts.

clinically relevant differences between the Nordic countries for both LDH, Stenosis and DDD [12,33,34].

One possible explanation could be differences in preoperative case-mix between our samples and the data utilized to model the DS algorithms. Comparing the descriptive statistics given by Fritzell et al. with our samples (Tables 1a and 1b) reveals only slight differences in the baseline measurements of EuroQol, ODI and VAS. None were clinically relevant with the exception of preoperative leg pain in DDD patients (Dialogue Support: 38.4; Study sample: 54.9). When comparing socio-demographic characteristics there were more smokers in our study samples across diagnosis groups (LHD: 28.0; Stenosis: 23.5; DDD: 20.0) compared to (LHD: 12.0; Stenosis: 8.0; DDD: 7.0). There were also more previously operated (LHD: 16.0%; Stenosis: 24.0%; DDD: 48.0%) compared to (LHD: 11.0%; Stenosis: 20.0%; DDD: 30.0%). There were however far fewer disability pensioners in our study samples (LHD: 5.0%; Stenosis: 5.5%; DDD: 7.5%) compared to (LHD: 23.0%; Stenosis: 15.0%; DDD: 28.0%). Whether these discrepancies in socio-demographic characteristics can explain the findings of this study remains unclear. Another possible explanation is the large disparity in sample size between the original Swedish development data and the current study. Statistically, it is very conceivable that random variation and selection bias may have influenced the outcomes in the present Danish samples. Complete access to the underlying algorithms of the DS is necessary to address this shortcoming.

Strength and limitations

This study is limited by relatively small sample sizes. Moreover, the samples were collected from a single Danish center. Thus, a possible selection bias cannot be ruled out. Ideally, the DS tool should be validated on larger, preferably national, Danish cohorts. This would for practical reasons require full access to the underlying predictive algorithms. Additionally, permissions would have to be obtained from all Danish spine centers due to the General Data Protection Regulation (GDPR) of the EU. Ideally, a minimum sample size for a reliable validation could be calculated by taking into account model specific properties and expected outcome proportions as suggested by Riley et al. [35], instead of relying on rule of thumb, e.g. of 100 - 200 events. The strength of this study is the high degree of comparability between the national spine registries Swespine and DaneSpine which allows direct application of Danish data on the DS without recoding or approximating input variables.

Conclusion

With the exception of proportion of average satisfied LSS patients, the predictive capabilities of the DS could not be generalized to sample data on Danish patient cohorts from the DaneSpine registry. Although AUC values were very similar to Fritzell et al., the detection rate of true negatives (failures) by the Dialogue Support tool was found to be inadequate. It remains to be determined if a true validation test on larger samples could yield a different result. For this reason, it is crucial that the underlying predictive algorithms of the DS are made available for other researchers to investigate its validity in populations outside Sweden.

Declarations of Competing Interests

One or more of the authors declare financial or professional relationships on ICMJE-NASSJ disclosure forms.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.xnsj.2022.100188.

References

[1] Andersen M, Nielsen M, Bech-Azeddine R, Helmig P ES. Spine Surgery in Demark. Yearly report (Rygkirurgi i Danmark. Årsrapport 2020). 2021.

[2] Fritzell P, Hagg O, Gerdhem P, Abbott A, Parai C TO, Stromqvist B, Mellgren L BC. Swespine 25 years. 2018 annual report follow up of spine surgery performed in Sweden in 2017. 2018.

[3] Solberg T OL. Yearly report from Norwegian Registry for Spine Surgery (NKR) 2017 (Årsrapport for 2017 med plan for forbedringstiltak). 2018.

[4] Aebi M, Grob D. SSE Spine Tango: a European spine registry promoted by the Spine Society of Europe (SSE). Eur Spine J 2004;13:661–2. doi:10.1007/S00586-004-0868-0.

[5] Asher AL, Speroff T, Dittus RS, Parker SL, Davies JM, Selden N, et al. The National Neurosurgery Quality and Outcomes Database (N2QOD): a collaborative North American outcomes registry to advance value-based spine care. Spine (Phila Pa 1976) 2014;39:S106–16. doi:10.1097/BRS.0000000000000579.

[6] Peul WC, Van Den Hout WB, Brand R, Thomeer RTWM, Koes BW. Prolonged conservative care versus early surgery in patients with sciatica caused by lumbar disc herniation: two year results of a randomised controlled trial. BMJ 2008;336:1355–8. doi:10.1136/BMJ.A143.

[7] Findlay GF, Hall BI, Musa BS, Oliveira MD, Fear SC. A 10-year follow-up of the outcome of lumbar microdiscectomy. Spine (Phila Pa 1976) 1998;23:1168–71. doi:10.1097/00007632-199805150-00019.

[8] Korres DS, Loupassis G, Stamos K. Results of lumbar discectomy: a study using 15 different evaluation methods. Eur Spine J 1992;1:20–4. doi:10.1007/BF00302137.

[9] Choi WS, Oh CH, Ji GY, Shin SC, Lee JB, Park DH, et al. Spinal canal morphology and clinical outcomes of microsurgical bilateral decompression via a unilateral approach for lumbar spinal canal stenosis. Eur Spine J 2014;23:991–8. doi:10.1007/S00586-013-3116-7.

[10] Strömqvist B, Fritzell P, Hägg O, Jönsson B, Sandén B. Swespine: the Swedish spine register : the 2012 report. Eur Spine J 2013;22:953–74. doi:10.1007/S00586-013-2758-9.

[11] Fritzell P, Mesterton J, Hagg O. Prediction of outcome after spinal surgery-using The Dialogue Support based on the Swedish national quality register. Eur Spine J 2021. doi:10.1007/S00586-021-07065-Y.

[12] Andersen MØ, Fritzell P, Eiskjaer SP, Lagerbäck T, Hägg O, Nordvall D, et al. Surgical treatment of degenerative disk disease in three scandinavian countries: an international register study based on three merged national spine registers. Glob Spine J 2019;9:850–8. doi:10.1177/2192568219838535.

[13] Health Statistics for the Nordic Countries 2017. 2017.

[14] Christiansen T, Lauridsen JT, Kifmann M, Lyttkens CH, Ólafsdóttir T VH. Public health, healthcare, health and inequality in health in the Nordic countries. 2018.

[15] [Implementation of the Danish national database Danespine for spinal surgery] - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/25346312/ (accessed January 3, 2022).

[16] DaneSpine. The Danish national database for spinal surgery. Data extraction, Oct. 2021. 2021.

[17] Parai C, Hägg O, Lind B, Brisby H. The value of patient global assessment in lumbar spine surgery: an evaluation based on more than 90,000 patients. Eur Spine J 2018;27:554–63. doi:10.1007/S00586-017-5331-0.

[18] EuroQolGroupEuroQol–a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.

[19] Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 1983;17:45–56. doi:10.1016/0304-3959(83)90126-4.

[20] Fairbank JCT, Pynsent PB. The Oswestry disability index. Spine (Phila Pa 1976) 2000;25:2940–53. doi:10.1097/00007632-200011150-00017.

[21] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73. doi:10.7326/M14-0698.

[22] RStudio TeamRStudio: integrated development for R. RStudio 2021:2021.

[23] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing 2021.

[24] Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag; 2016. p. 2016.

[25] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf 2011;12:1–8. doi:10.1186/1471-2105-12-77/TABLES/3.

[26] Tharwat A. Classification assessment methods. Appl Comput Informat 2018. doi:10.1016/J.ACI.2018.08.003.

[27] Kuhn M, Johnson K. Applied predictive modeling. 1st ed. Springer; 2013. doi:10.1007/978-1-4614-6849-3.

[28] Chicco D, Tötsch N, Jurman G. The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 2021;14:1–22. doi:10.1186/S13040-021-00244-Z/TABLES/5.

[29] Austin PC, Harrell FE, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat Med 2020;39:2714. doi:10.1002/SIM.8570.

[30] Steyerberg EW. Clinical prediction models. 2nd ed. Cham: Springer International Publishing; 2019. doi:10.1007/978-3-030-16399-0.

[31] Cabitza F, Campagner A, Soares F, García de Guadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed 2021;208. doi:10.1016/J.CMPB.2021.106288.

[32] Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 2008;17:145–51. doi:10.1111/j.1466-8238.2007.00358.x.

[33] Lønne G, Fritzell P, Hägg O, Nordvall D, Gerdhem P, Lagerbäck T, et al. Lumbar spinal stenosis: comparison of surgical practice variation and clinical outcome in three national spine registries. Spine J 2019;19:41–9. doi:10.1016/J.SPINEE.2018.05.028.

[34] Lagerbäck T, Fritzell P, Hägg O, Nordvall D, Lønne G, Solberg TK, et al. Effectiveness of surgery for sciatica with disc herniation is not substantially affected by differences in surgical incidences among three countries: results from the Danish, Swedish and Norwegian spine registries. Eur Spine J 2019;28:2562–71. doi:10.1007/S00586-018-5768-9.

[35] Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med 2021;40:4230–51. doi:10.1002/SIM.9025.