Predicting patient-reported outcomes following lumbar spine surgery: development and external validation of multivariable prediction models

M. Halicka[a], M. Wilby[b], R. Duarte[c], C. Brown[a]

[a]Department of Psychological Sciences, University of Liverpool, Liverpool, UK
[b]Department of Neurosurgery, The Walton Centre NHS Foundation Trust, Liverpool, UK
[c]Liverpool Reviews & Implementation Group (LRiG), University of Liverpool, Liverpool, UK

Corresponding author: Monika Halicka, Department of Psychological Sciences, University of Liverpool, Eleanor Rathbone Building, Bedford Street South, Liverpool, L69 7ZA, UK, +44(0)1517941398, mon.halicka@gmail.com

This research was funded via the Translational Research Access Programme (TRAP), Faculty of Health & Life Sciences, University of Liverpool, UK.

The authors have no conflicts of interest to declare.

Abstract

Background: This study aimed to develop and externally validate prediction models of spinal surgery outcomes based on a retrospective review of a prospective clinical database, uniquely comparing multivariate regression and machine learning approaches, and identifying the most important predictors.

Methods: Outcomes were change in back and leg pain intensity and Core Outcome Measures Index (COMI) from baseline to the last available postoperative follow-up (3-24 months), defined as minimal clinically important change (MCID) and continuous change score. Eligible patients underwent lumbar spine surgery for degenerative pathology between 2011 and 2021. Data were split by surgery date into development (N=2691) and validation (N=1616) sets. Multivariate logistic and linear regression, and random forest classification and regression models, were fit to the development data and validated on the external data.

Results: All models demonstrated good calibration in the validation data. Discrimination ability (area under the curve) for MCID ranged from 0.63 (COMI) to 0.72 (back pain) in regression, and from 0.62 (COMI) to 0.68 (back pain) in random forests. The explained variation in continuous change scores spanned 16%-28% in linear, and 15%-25% in random forests regression. The most important predictors included age, baseline scores on the respective outcome measures, type of degenerative pathology, previous spinal surgeries, smoking status, morbidity, and duration of hospital stay.

Conclusions: The developed models appear robust and generalisable across different outcomes and modelling approaches but produced only borderline acceptable discrimination ability, suggesting the need to assess further prognostic factors. External validation showed no advantage of the machine learning approach.

1. Introduction

Chronic back pain is the single greatest cause of years lived with disability worldwide (Rice et al., 2016) with annual direct healthcare costs in the UK of £1,632 million (Maniadakis & Gray, 2000). Spinal surgery is the largest component of healthcare expenditure for managing low back pain (costing £5000-£10,000 per patient) and its rate has doubled over 15-years from 2.5 to 4.9 per 10,000 adults (Weir et al., 2017). However, success rates of lumbar spine surgery are highly variable - only about 60% of patients achieve reductions in pain of at least a minimal clinically important difference (MCID), and 1/5 experience persistent long-term pain after surgery (Hegarty & Shorten, 2012; Weir et al., 2017; Zweig et al., 2017).

Reliable predictive factors could maximise patient benefit and cost-effectiveness of surgery. Although several systematic reviews concluded that medical, sociodemographic, and psychological factors are linked to improvement in pain and disability following lumbar spine surgery (Dorow et al., 2017; Halicka et al., *under review*; Rushton et al., 2018; Wilhelm et al., 2017; Wilson et al., 2016), we lack clear clinical guidelines on the reliable predictors (de Campos, 2017) and predictive factors are rarely formally documented.

Statistical prediction models enable probabilistic estimation of treatment outcome given a set of preoperative patient data. Previously developed regression-based models demonstrated good ability to discriminate between patients who did and did not achieve MCID in pain or disability after lumbar spine surgery (Abbott et al., 2011; Cobo Soriano et al., 2010; Khor et al., 2018; McGirt et al., 2015, 2017). However, performance of the existing prediction models has been rarely quantified in patient data other than that used to develop them. In the few available examples, external validation revealed poorer discrimination ability in the new data (Janssen et al., 2021; Staub et al., 2020), or poor calibration leading to under- and overestimation of outcome probabilities (Quddusi et al., 2020).

Machine learning has been found to improve predictive performance relative to regression (Couronné et al., 2018), including for spinal surgery outcomes (Karhade et al., 2021; Siccoli et al., 2019; Staartjes et al., 2019). This may stem from the ability to capture nonlinear or interactive effects, often characteristic of clinical data. Conversely, machine learning algorithms are prone to overfitting smaller datasets and comparisons to standard regression have lacked external validation; thus, such predictions may not extrapolate well to new data. External validation is rare but necessary before any prediction models can be implemented in clinical practice (Moons et al., 2015; Siontis et al., 2015).

Our primary objective was to develop and externally validate prediction models of patient-reported spinal surgery outcomes, for the first time comparing the performance of multivariate regression and machine learning approaches. Secondarily, we identify the most relevant predictors from routinely collected medical and patient-reported information, since existing models (particularly those from the US; Karhade et al., 2021; Khor et al., 2018; McGirt et al., 2015, 2017) may not translate well to UK cohorts due to different healthcare systems, type of data recorded, and cultural and demographic differences.

2. Methods

2.1. Source of data

This study was based on retrospective review of a prospective clinical database from a single Neurosurgery Department at the Walton Centre NHS Foundation Trust (United Kingdom). The

Walton Centre contributes data to the Eurospine's international Spine Tango registry (Röder et al., 2002), which governs standardised data collection protocols. Specifically, patients complete a self-assessment form (Core Outcome Measures Index, COMI) (Deyo et al., 1998; Ferrer et al., 2006) at the preoperative consultation and postoperative follow-ups at 3, 12, and 24 months. The surgeon completes the surgery form before discharge, detailing the patient's history, type of pathology and surgery, and hospital stay (Röder et al., 2005). The Walton Centre's database was reviewed on 26/04/2021 to extract lumbar spine surgery cases with degenerative disease as the main pathology. This data included patients operated between 4/04/2011 and 30/03/2021, with the last follow-up dated 21/04/2021. Patients partaking in the Spine Tango registry consented for their pseudonymised data being used for research purposes, however, due to confidentiality, the data are not publicly available. The study received approval from an institutional ethics committee (ref. 8224).

2.2. Participants

Eligible patients had lumbar disc herniation and/or stenosis and underwent elective spinal decompression surgery with or without fusion. These most common degenerative pathologies and surgical measures were selected to obtain a representative sample and minimise clinical heterogeneity. Eligible patients completed preoperative and at least one postoperative self-assessment form. If there were multiple surgery cases per patient, only chronologically first eligible surgery was included.

2.3. Outcomes

The spinal surgery outcomes were defined as reduction in back and leg pain intensity and reduction in COMI from baseline to the last available follow-up. Due to the self-reported nature of outcomes, their assessment was not blinded.

Back and leg average pain intensity in the past week was measured on 0 (*no pain*) to 10 (*worst pain I can imagine*) numerical rating scales embedded within COMI. This scale is a recommended outcome measure of pain in studies evaluating effectiveness of treatments for chronic pain (Dworkin et al., 2005). COMI (Deyo et al., 1998) is a multidimensional instrument consisting of measures of pain, function, symptom-specific well-being, quality of life, social disability, and work disability. Average COMI score can range from 0 to 10, where higher scores indicate worse level of functioning. This instrument is an official outcome measure of Eurospine's international spine registry and has been extensively validated in patients with back pain (Ferrer et al., 2006; Mannion et al., 2005).

Both continuous and dichotomous outcomes were considered for their precision and clinical utility, respectively. Continuous outcomes represent numerical differences between the baseline and follow-up back pain, leg pain, and COMI. Positive scores correspond to improvement (i.e., reduction in pain and functional impairment). Dichotomous outcomes represent achievement of MCID between baseline and follow-up back pain and leg pain (reduction of ≥2 points; Farrar et al., 2001), and COMI (reduction of ≥2.2 points; Mannion et al., 2009).

2.4. Predictors

As candidate predictors we assessed preoperative factors, and controlled for potential intra- and postoperative confounders (see **Table 1** for a full list). Assessment of all predictors was blinded to

outcomes assessed ≥3 months later, but not blinded to other predictors, either self-reported by the patients before surgery, or recorded by the surgeon before discharge.

2.5. Sample size

To estimate the minimum required sample size and minimise the risk of overfitting, at least 20 participants per predictor should be included for continuous outcomes (Harrell, 2015), and 10 or more events (i.e., achieving MCID) per predictor for dichotomous outcomes (Moons et al., 2014). We considered 19 predictors, however, since several categorical factors had more than two levels, each additional factor level was counted separately, resulting in 34 predictors in total. Therefore, a minimum of 680 participants would be required for continuous outcomes, and 340 participants who achieve MCID for dichotomous outcomes.

2.6. Data processing and statistical analysis

Data was processed and analysed using R software, version 4.1.1 (R Core Team, 2021). Details of data processing and handling of predictors are described in **Methods S1.** Except for categorising portion of the BMI data to match the variable types across two versions of the surgery form, other continuous factors were treated as continuous in the analysis. Due to severely skewed distribution, hospital duration was log-transformed (base 10). Analysis scripts are available at https://osf.io/x3rwg/?view_only=958cac2415e54281980c83ff7f4d786c.

2.6.1. Handling missing data

As per our eligibility criteria, only patients with complete outcome data were included in the analysis. We report attrition and summary statistics on the portion of excluded data, and between-group comparisons with the final included sample.

Missing predictor data was assumed to be missing at random and addressed through Multivariate Imputation by Chained Equations (MICE) with 40 iterations using *mice* R package, (Buuren & Groothuis-Oudshoorn, 2011). The MICE procedure is explained in **Methods S2**. Values were missing in 11 out of 19 predictors, with partial missing rates from <1% to 36% (see Results for details), and a total missing rate of 3.75% across all predictors. Multiple imputation has been found to remain unbiased up to 50% missing rates (Haji-Maghsoudi et al., 2013; Marshall et al., 2010). A series of diagnostic checks detailed in **Methods S2** demonstrated good convergence and no apparent biases of the MICE algorithm, with overlapping range and distributions of the imputed and observed data (**Figs. S1-S4**), and the missingness of each variable was associated with other factors in the imputation dataset (**Table S1**), therefore supporting the missing at random assumption. We also conducted a sensitivity complete case analysis excluding cases with missing data on any of the predictors.

2.6.2. Model development and validation data

The dataset was split into model development and validation samples based on the date of surgery (2011-2017 and 2018-2021, respectively). For large datasets, a non-random split has been recommended, e.g., by time (Moons et al., 2014). The chosen time-split coincided with the introduction of a new version of the surgery form. The setting, eligibility criteria, predictors, and outcomes used were consistent across the development and validation data, except for some differences between the surgery form versions (see **Methods S1**).

2.6.3. Multivariate regression analysis

As our primary analysis, we developed multivariate logistic regression models for each dichotomous outcome, and multivariate linear regression models for each continuous outcome. Models were

fitted using R functions *glm* (with binomial family and logit link function) and *lm* (for linear regression) within the *stats* package (R Core Team, 2021). Full model approach (i.e., including all candidate predictors) was used to estimate prediction accuracy based on the available set of routinely collected data in combination and to assess relative contribution of specific preoperative factors while controlling for potential confounders. Each model based on the development sample was then fitted to the predictor values in the validation data to predict the outcomes of interest in new, untrained data. The predicted outcomes were compared with the actual (observed) outcome values.

Model performance on the development and validation data was expressed by calibration and discrimination measures with 95% confidence intervals (CI) where applicable. For calibration plots, we divided each dataset into 10 deciles according to the predicted probability of MCID (or predicted change score for linear regression), and plotted the mean predicted versus mean observed probabilities (or change scores) for each decile. Perfect fit would be reflected by all points being aligned on a line with intercept 0 and slope 1. We tested whether the 95% CIs of the intercept and slope of the calibration line of best fit include these values. Calibration of the logistic regression models was additionally tested with Hosmer and Lemeshow χ^2 tests to assess the differences between the observed and predicted numbers of MCID and no-MCID ($p > 0.05$ would indicate adequate fit). For linear regression models, correlation between ungrouped observed and predicted outcomes was expressed as Pearson's *r*. Discrimination of the logistic regression models was assessed via Receiver-Operating Characteristic curve (ROC) plots with estimated Area Under the Curve (AUC; c-index), and classification accuracy expressed as sensitivity and specificity estimates at the optimal ROC cut-off (probability threshold maximising both indices). AUC can range from 0 to 1 and a value of 0.5 corresponds to chance discrimination, while 0.7-0.8 is considered acceptable, and >0.8 excellent discrimination (Hosmer et al., 2013). Nagelkerke pseudo-R^2 and deviance were also reported as overall performance measures. For linear regression, discrimination was quantified by R^2, *F*-test, and root mean square error (RMSE) for each model. To express adjusted contribution of each predictor to the outcome of interest, we presented log odds and odds ratios (for dichotomous outcomes) and unstandardised and standardised regression coefficients (for continuous outcomes) with 95% CIs.

2.6.4. Random forests

As our secondary analysis, we used random forests (RF), which demonstrated superior predictive performance over other machine learning algorithms (Breiman, 2001; Christodoulou et al., 2019; van der Ploeg et al., 2016). RF is a non-linear classification and regression algorithm based on an ensemble of deep decision trees. Multiple decision trees are trained on different randomly selected bootstrap samples of the same training dataset. Each time a decision tree split is performed, the best split variable (Breiman, 2001) is chosen from a random subset of the original predictor set. Each tree gives an outcome prediction on the leftover data which was not used during training (out-of-bag, OOB). The predictions are then aggregated (bagging) by assigning class labels (MCID vs. no-MCID) by majority vote, or by averaging the continuous dependent variable (change score) across all the trees. Prediction errors are calculated as OOB error (misclassification rate) for classification or OOB mean square error for regression. We used Breiman's implementation of RF in *randomForest* R package (Breiman, 2001; Wright et al., 2021). The RF algorithm hyperparameters were tuned to maximise the predictive accuracy as described in **Methods S3** and **Table S2**. Using the final hyperparameters and 500 trees for each model, we applied RF classification to predict MCID outcomes, and RF regression to predict continuous outcomes in the development data. Then we used each developed model to predict the same outcomes in the validation data. Although the leg pain outcome was characterised by a moderate class imbalance, with MCID rates of 67-68%, no

adjustments such as downsampling non-events were made, as these could distort the true outcome rates, lead to inadequate clinical predictions, and increase the risk of overfitting.

RF discrimination in the development and validation data was quantified by OOB errors and AUC with sensitivity and specificity at the optimal ROC cut-off for classification models; and by RMSE, pseudo-R^2, and Pearson's *r* for observed against predicted outcomes for regression models. Calibration plots were prepared following the procedures described for the logistic and linear regression models. We also presented relative variable importance based on mean decrease in accuracy (loss in prediction performance) when a particular variable is omitted from the training data for each development model.

3. Results

3.1. Participants

Fig. S5 illustrates the flow of participants through the eligibility screening process. Out of 6810 screened surgery cases, 4307 unique patients were included in the analysis. Descriptive characteristics, missingness rates, and statistical comparisons between included and excluded participants due to missing baseline and/or follow-up assessment are reported in **Results S1** and **Table S3**. Although there were statistically significant differences on 10 predictors, as expected in large datasets even when effect sizes of these differences are small, the included data were representative: the included data covered the full range of possible predictor values in the excluded data, and all levels of categorical factors present in the excluded data were well-represented in the included data.

Development dataset included 2691 and validation dataset 1616 patients, thus each was more than sufficient to fit regression models with 19 specified predictors. Patient characteristics are presented in **Table S4** and any group differences are described in **Results S1**. On average, patients in the development compared to the validation sample achieved less reduction in back pain (mean [SD]: 2.15 [3.30] vs. 2.51 [3.50]) and leg pain (3.68 [3.69] vs. 3.94 [3.75]), but did not differ in COMI change (3.22 [3.07] vs. 3.34 [3.18]) or MCID rates in COMI (57% vs. 58%), back pain (53% vs. 57%), or leg pain (67% vs. 68%). There were slight statistically significant differences on most of the candidate predictors; yet, importantly, the validation data was within the range of the development data for all variables.

3.2. Development and validation of regression models

Regression diagnostic checks are detailed in **Results S2** and **Figs. S7-11**. In summary, there was no multicollinearity among the predictors, no severe deviations from linearity between the continuous predictors and logit of MCID or continuous outcomes, residual variance was homogenous across different levels of categorical predictors, there were no highly influential values, models had good fit across the range of observations, and standardised residuals in linear models showed acceptable homoscedasticity and normal distribution, with slight deviations on the tails.

3.2.1. Binary outcomes (MCID)

Table 1 presents log-odds with 95% CIs for each candidate predictor in the development models for achievement of MCID in COMI, back pain, and leg pain intensity, while **Fig. S12** illustrates the corresponding ORs. Independent of other factors included in the models, older age was associated with higher odds of achieving MCID after surgery across all outcomes. Decompression with fusion surgery was related to higher odds of MCID in COMI and leg pain, whereas blood loss of 100-500 ml – higher odds of MCID in back pain. Additionally, higher baseline COMI, back pain, and leg pain

predicted better odds of improvement in their corresponding outcomes. In contrast, patients with spinal stenosis, history of previous surgeries, currently smoking, and with higher morbidity class had lower odds of achieving MCID after surgery across all outcomes. Disc herniation with stenosis, higher baseline COMI, and presence of any complications also predicted lower odds of MCID in back pain. Finally, patients with higher baseline back pain and longer hospital stay had lower odds of MCID in COMI and leg pain. Adjusted effect sizes of these predictors were small (ORs >0.4 and <2.5).

Table 1. Results of the multivariate logistic regression models on MCID outcomes and multivariate linear regression models on continuous outcomes in the development data and models' performance in the validation data.

		COMI MCID	***Back pain MCID***	***Leg pain MCID***	***COMI change***	***Back pain change***	***Leg pain change***
Predictors		*Log-Odds (95% CI)*	*Log-Odds (95% CI)*	*Log-Odds (95% CI)*	*Beta (95% CI)*	*Beta (95% CI)*	*Beta (95% CI)*
(Intercept)		-0.68 (-1.57, 0.20)	**-1.29 (-2.20, -0.38)**	-0.72 (-1.67, 0.23)	-0.56 (-1.77, 0.65)	-0.93 (-2.17, 0.30)	-0.59 (-1.97, 0.79)
Age (years)		**0.01 (0.00, 0.02)**	**0.01 (0.01, 0.02)**	**0.01 (0.00, 0.02)**	**0.01 (0.00, 0.02)**	**0.02 (0.01, 0.03)**	**0.02 (0.00, 0.03)**
Gender (ref. female)	male	0.13 (-0.03, 0.29)	0.16 (-0.00, 0.33)	0.17 (-0.01, 0.34)	**0.23 (0.01, 0.46)**	**0.28 (0.05, 0.51)**	**0.36 (0.11, 0.61)**
Degen. disease (ref. disc herniation)	stenosis	**-0.5 (-0.75, -0.26)**	**-0.46 (-0.71, -0.21)**	**-0.27 (-0.53, -0.01)**	**-0.64 (-0.96, -0.31)**	**-0.51 (-0.84, -0.18)**	**-0.67 (-1.04, -0.29)**
	disc herniation & stenosis	-0.25 (-0.51, 0.02)	**-0.3 (-0.57, -0.03)**	-0.12 (-0.41, 0.18)	**-0.37 (-0.73, -0.01)**	**-0.49 (-0.86, -0.13)**	**-0.45 (-0.86, -0.04)**
Previous surgeries (ref. no)	yes	**-0.4 (-0.60, -0.20)**	**-0.25 (-0.46, -0.04)**	**-0.41 (-0.62, -0.19)**	**-0.74 (-1.02, -0.46)**	**-0.43 (-0.72, -0.15)**	**-0.89 (-1.21, -0.57)**
Treatment duration (ref. <3 months)	3-12 months	-0.05 (-0.23, 0.14)	-0.17 (-0.35, 0.02)	-0.06 (-0.26, 0.14)	-0.2 (-0.45, 0.05)	-0.17 (-0.43, 0.08)	-0.16 (-0.44, 0.13)
	>12 months	0 (-0.24, 0.23)	-0.12 (-0.35, 0.12)	-0.01 (-0.26, 0.24)	-0.1 (-0.41, 0.22)	-0.19 (-0.51, 0.13)	-0.17 (-0.53, 0.19)
BMI (ref. normal)	overweight	0.12 (-0.08, 0.32)	0.12 (-0.08, 0.32)	-0.05 (-0.26, 0.17)	0.11 (-0.16, 0.38)	0.03 (-0.24, 0.31)	-0.14 (-0.45, 0.17)
	obese	0 (-0.22, 0.21)	0.02 (-0.20, 0.24)	0.04 (-0.20, 0.28)	-0.19 (-0.48, 0.11)	-0.26 (-0.56, 0.05)	-0.04 (-0.38, 0.30)
Current smoker (ref. no)	yes	**-0.52 (-0.75, -0.28)**	**-0.41 (-0.65, -0.17)**	**-0.57 (-0.82, -0.33)**	**-0.73 (-1.05, -0.41)**	**-0.71 (-1.03, -0.38)**	**-1 (-1.36, -0.63)**
Morbidity (ref. ASA1)	ASA2	**-0.33 (-0.54, -0.13)**	**-0.32 (-0.53, -0.11)**	**-0.5 (-0.73, -0.27)**	**-0.44 (-0.71, -0.16)**	**-0.41 (-0.69, -0.13)**	**-0.57 (-0.89, -0.26)**
	ASA3/4	**-0.55 (-0.85, -0.26)**	**-0.4 (-0.71, -0.10)**	**-0.59 (-0.92, -0.27)**	**-0.94 (-1.35, -0.54)**	**-0.6 (-1.01, -0.19)**	**-0.91 (-1.37, -0.44)**
Baseline COMI		**0.22 (0.15, 0.29)**	**-0.1 (-0.17, -0.03)**	-0.05 (-0.13, 0.02)	**0.63 (0.54, 0.72)**	**-0.12 (-0.21, -0.02)**	-0.08 (-0.18, 0.02)
Baseline back pain		**-0.14 (-0.18, -0.11)**	**0.31 (0.27, 0.35)**	**-0.11 (-0.16, -0.07)**	**-0.24 (-0.28, -0.19)**	**0.62 (0.57, 0.67)**	**-0.21 (-0.27, -0.16)**
Baseline leg pain		0.02 (-0.02, 0.07)	-0.01 (-0.06, 0.04)	**0.37 (0.31, 0.42)**	0.04 (-0.03, 0.10)	-0.03 (-0.10, 0.03)	**0.83 (0.76, 0.91)**
Extent of surgery (ref. 1)	2	0.13 (-0.38, 0.62)	-0.01 (-0.52, 0.50)	0 (-0.55, 0.52)	0.28 (-0.41, 0.96)	0.18 (-0.51, 0.88)	0.22 (-0.56, 1.00)
	3	0.31 (-0.23, 0.85)	0.12 (-0.44, 0.67)	0.19 (-0.40, 0.76)	0.29 (-0.45, 1.02)	0.29 (-0.47, 1.04)	0.28 (-0.56, 1.12)
	≥4	0.01 (-0.70, 0.71)	-0.1 (-0.83, 0.63)	-0.29 (-1.04, 0.45)	-0.02 (-0.99, 0.95)	0.02 (-0.97, 1.01)	-0.11 (-1.22, 1.00)
Surgical measures (ref. decompression)	decompression & fusion	**0.68 (0.19, 1.19)**	0.37 (-0.12, 0.87)	**0.59 (0.06, 1.14)**	**1.14 (0.48, 1.79)**	**0.93 (0.27, 1.60)**	**1.22 (0.48, 1.97)**
Surgeon (ref. specialized spine)	board certified neuro	0.13 (-0.06, 0.33)	0.1 (-0.10, 0.29)	-0.06 (-0.27, 0.14)	0.1 (-0.16, 0.36)	0.07 (-0.20, 0.34)	0 (-0.30, 0.30)
	board certified ortho	0.11 (-0.70, 0.96)	0.09 (-0.76, 0.93)	0.46 (-0.47, 1.53)	-0.12 (-1.22, 0.97)	-0.02 (-1.14, 1.10)	0.6 (-0.65, 1.85)
	neuro in training	-0.05 (-0.28, 0.18)	-0.16 (-0.39, 0.07)	-0.14 (-0.39, 0.10)	-0.12 (-0.43, 0.19)	-0.04 (-0.36, 0.27)	-0.1 (-0.45, 0.25)
	ortho in training	-0.14 (-0.97, 0.69)	-0.64 (-1.52, 0.21)	-0.15 (-1.03, 0.77)	-0.42 (-1.57, 0.72)	-0.68 (-1.85, 0.48)	-0.02 (-1.33, 1.28)
	other	0.12 (-0.37, 0.63)	0.27 (-0.25, 0.79)	-0.14 (-0.66, 0.40)	0.03 (-0.66, 0.71)	0.21 (-0.49, 0.90)	0.08 (-0.70, 0.86)
Operation time (ref. <1)	1-2	-0.13 (-0.31, 0.05)	-0.05 (-0.24, 0.13)	-0.08 (-0.27, 0.12)	-0.19 (-0.43, 0.06)	-0.11 (-0.36, 0.14)	-0.19 (-0.47, 0.09)
	2-3	-0.09 (-0.48, 0.30)	-0.02 (-0.42, 0.38)	0.03 (-0.39, 0.46)	-0.15 (-0.68, 0.39)	-0.06 (-0.61, 0.48)	0.15 (-0.46, 0.76)
	>3	-0.6 (-1.26, 0.05)	-0.55 (-1.21, 0.11)	-0.29 (-0.98, 0.41)	**-1 (-1.88, -0.13)**	**-1.03 (-1.93, -0.14)**	-0.27 (-1.28, 0.73)

Blood loss (ref. <100 ml)	100-500 ml	0.17 (-0.06, 0.41)	**0.25 (0.01, 0.49)**	0.2 (-0.05, 0.46)	0.27 (-0.05, 0.59)	0.25 (-0.07, 0.58)	0.18 (-0.19, 0.54)
	>500 ml	-0.76 (-1.72, 0.12)	-0.19 (-1.08, 0.68)	-0.43 (-1.31, 0.46)	-0.55 (-1.70, 0.59)	-0.61 (-1.78, 0.55)	-0.63 (-1.93, 0.68)
Complications (ref. no)	yes	-0.19 (-0.43, 0.05)	**-0.31 (-0.56, -0.06)**	-0.24 (-0.50, 0.01)	-0.15 (-0.48, 0.18)	-0.16 (-0.50, 0.18)	-0.29 (-0.66, 0.09)
(log) Hospital stay (days)		**-0.25 (-0.42, -0.08)**	-0.12 (-0.29, 0.05)	**-0.25 (-0.43, -0.08)**	**-0.41 (-0.64, -0.18)**	**-0.26 (-0.49, -0.02)**	**-0.33 (-0.59, -0.07)**
Follow-up (ref. 3 months)	12 months	-0.01 (-0.49, 0.47)	-0.07 (-0.56, 0.42)	0.02 (-0.50, 0.52)	-0.06 (-0.71, 0.59)	-0.25 (-0.91, 0.42)	0.03 (-0.72, 0.77)
	24 months	0.1 (-0.33, 0.52)	0.12 (-0.31, 0.55)	0.25 (-0.21, 0.69)	0.44 (-0.13, 1.01)	-0.03 (-0.61, 0.55)	0.39 (-0.26, 1.04)
Development performance (N = 2691)							
Nagelkerke R^2		0.10	0.17	0.17			
Null deviance (df = 2690) / Residual deviance (df = 2657)		3669 / 3456	3723 / 3359	3424 / 3077			
AUC (95% CI)		0.65 (0.63, 0.67)	0.69 (0.67, 0.71)	0.70 (0.68, 0.72)			
Sensitivity / Specificity (optimal cut-off)		0.66 / 0.56 (0.56)	0.75 / 0.51 (0.49)	0.72 / 0.56 (0.65)			
F-test					F(33,2657) = 13.35, p < 0.001	F(33,2657) = 24.11, p < 0.001	F(33,2657) = 23.62, p < 0.001
R^2 / adjusted R^2					0.142 / 0.132	0.230 / 0.221	0.227 / 0.217
RMSE					2.84	2.90	3.24
Calibration intercept (95% CI)		0.01 (-0.06, 0.09)	0.02 (-0.09, 0.12)	0.02 (-0.05, 0.08)	0.10 (-0.28, 0.48)	0.00 (-0.25, 0.26)	-0.03 (-0.48, 0.41)
Calibration slope (95% CI)		0.97 (0.85, 1.09)	0.97 (0.78, 1.16)	0.97 (0.88, 1.07)	0.97 (0.86, 1.08)	1.00 (0.90, 1.09)	1.01 (0.90, 1.12)
Hosmer-Lemeshow test		$\chi^2(8) = 5.29$, p = 0.727	$\chi^2(8) = 22.95$, p = 0.003	$\chi^2(8) = 5.50$, p = 0.703			
Pearson's r (observed vs. predicted)					0.38	0.48	0.48
Validation performance (N = 1616)							
AUC (95% CI)		0.63 (0.61, 0.66)	0.72 (0.69, 0.74)	0.68 (0.65, 0.71)			
Sensitivity / Specificity (optimal cut-off)		0.51 / 0.70 (0.61)	0.53 / 0.77 (0.61)	0.76 / 0.49 (0.63)			
R^2					0.156	0.276	0.219
RMSE					2.93	3.00	3.33
Calibration intercept (95% CI)		0.09 (-0.06, 0.24)	0.02 (-0.11, 0.14)	0.08 (-0.03, 0.18)	0.18 (-0.23, 0.59)	0.15 (-0.18, 0.48)	0.49 (0.28, 0.69)
Calibration slope (95% CI)		0.84 (0.59, 1.10)	1.03 (0.81, 1.26)	0.89 (0.74, 1.04)	1.02 (0.90, 1.15)	1.09 (0.97, 1.21)	0.94 (0.89, 0.99)
Hosmer-Lemeshow test		$\chi^2(8) = 14.71$, p = 0.065	$\chi^2(8) = 29.39$, p < 0.001	$\chi^2(8) = 10.54$, p = 0.229			
Pearson's r (observed vs. predicted)					0.40	0.53	0.47

Note. Coefficients in bold indicate significant predictors (95% CI does not include 0). ASA, American Society of Anaesthesiologists morbidity class; AUC, Area Under the Curve; BMI, body mass index; CI, confidence interval; COMI, Core Outcome Measures Index; MCID, Minimal Clinically Important Difference; neuro, neurosurgeon; ortho, orthopaedic surgeon; ref., reference term; RMSE, root mean square error.

The AUC in the development and validation data was bordering on between no-better-than-chance and acceptable discriminability (**Table 1**). COMI MCID model had the worst discrimination, whereas the highest, acceptable, discrimination ability was found for the back pain MCID model in the validation data. Using the optimal ROC cut-offs for each outcome, development models generally had good sensitivity (ability to detect true MCID), while specificity (detecting true no-MCID) oscillated near chance classification. There was a consistent pattern of classification in the validation data for leg pain MCID, however, COMI and back pain MCID classification presented an opposite pattern, with good specificity but poor sensitivity. The ROC curves are presented in **Fig. S13**.

The proportion of explained variation in the development models ranged from 10% for COMI to 17% for back and leg pain intensity. Residual deviance was lower than null deviance, indicating that the included variables allow to predict each outcome better than intercept-only (null) models. Calibration tests indicated good model fit for MCID in COMI and leg pain in the development (**Fig. S14a**) and validation data (**Fig. 1a**), however, according to the Hosmer-Lemeshow test, predicted probabilities of MCID in back pain did significantly deviate from the observed values in both samples. This test is sensitive to sample size and is often significant in large datasets. Considering its results alongside the visual inspection and the fact that in all cases, the intercept of the calibration lines did not significantly differ from 0, and their slope did not significantly differ from 1, we conclude that the models showed good calibration in the validation data.

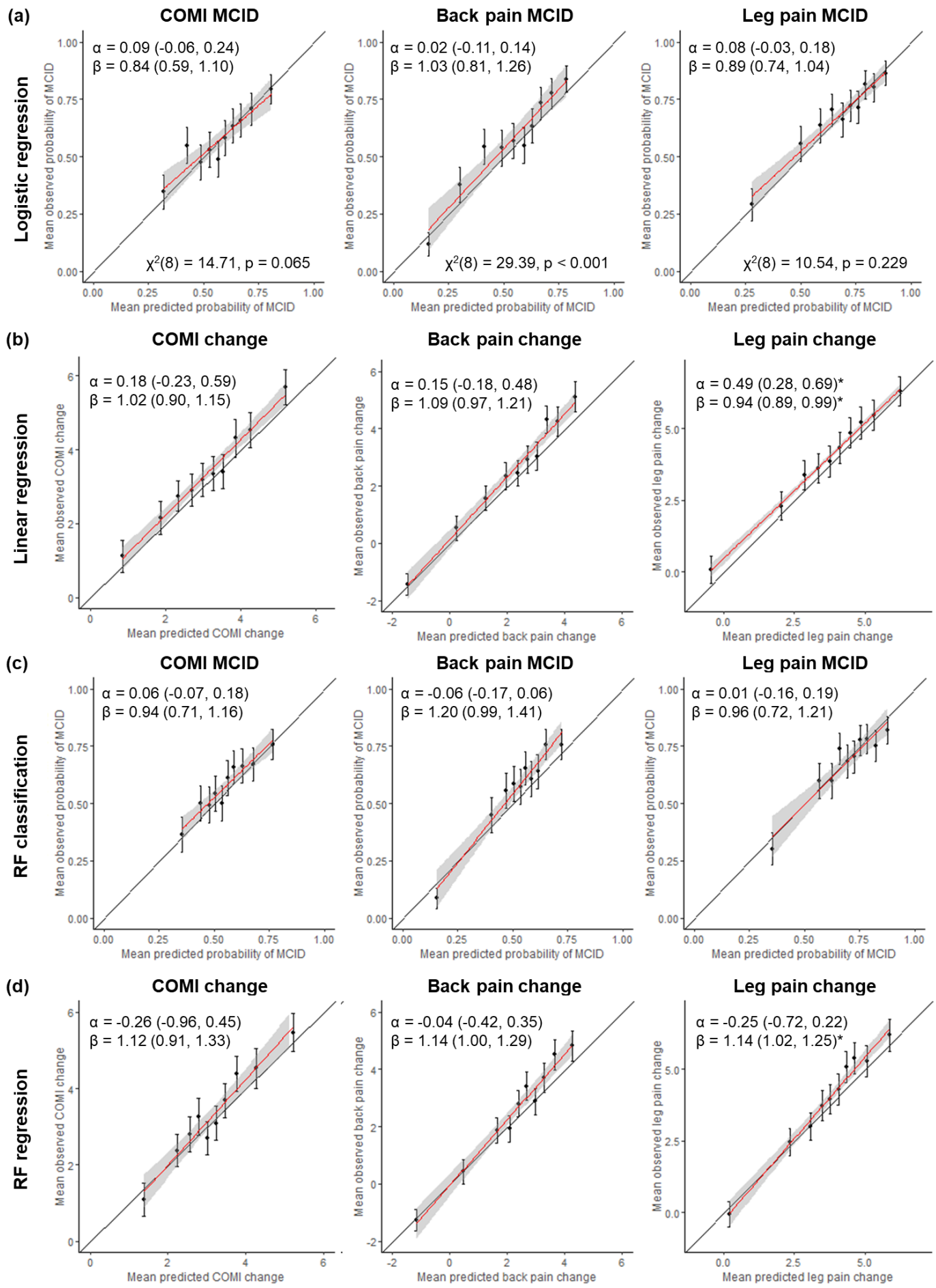
(a)
Logistic regression
COMI MCID
α = 0.09 (-0.06, 0.24)
β = 0.84 (0.59, 1.10)
χ2(8) = 14.71, p = 0.065
Back pain MCID
α = 0.02 (-0.11, 0.14)
β = 1.03 (0.81, 1.26)
χ2(8) = 29.39, p < 0.001
Leg pain MCID
α = 0.08 (-0.03, 0.18)
β = 0.89 (0.74, 1.04)
χ2(8) = 10.54, p = 0.229
Mean observed probability of MCID
Mean predicted probability of MCID
(b)
Linear regression
COMI change
α = 0.18 (-0.23, 0.59)
β = 1.02 (0.90, 1.15)
Mean observed COMI change
Mean predicted COMI change
Back pain change
α = 0.15 (-0.18, 0.48)
β = 1.09 (0.97, 1.21)
Mean observed back pain change
Mean predicted back pain change
Leg pain change
α = 0.49 (0.28, 0.69)*
β = 0.94 (0.89, 0.99)*
Mean observed leg pain change
Mean predicted leg pain change
(c)
RF classification
COMI MCID
α = 0.06 (-0.07, 0.18)
β = 0.94 (0.71, 1.16)
Back pain MCID
α = -0.06 (-0.17, 0.06)
β = 1.20 (0.99, 1.41)
Leg pain MCID
α = 0.01 (-0.16, 0.19)
β = 0.96 (0.72, 1.21)
(d)
RF regression
COMI change
α = -0.26 (-0.96, 0.45)
β = 1.12 (0.91, 1.33)
Back pain change
α = -0.04 (-0.42, 0.35)
β = 1.14 (1.00, 1.29)
Leg pain change
α = -0.25 (-0.72, 0.22)
β = 1.14 (1.02, 1.25)*

Fig. 1. Calibration plots for (a) logistic regression, (b) linear regression, (c) random forest (RF) classification, and (d) RF regression models when fitted to the validation data for each outcome. Points correspond to the mean predicted and observed probabilities of MCID or change scores in each decile with 95% confidence intervals (CI) of the mean observed probabilities or change scores. The results of the Hosmer-Lemeshow goodness of fit tests are presented in the bottom-right corner of each plot for logistic regression (a). Calibration lines of best fit are plotted in red in 95% CI in grey and their intercept (α) and slope (β) estimates with 95% CIs are presented in the top-left corner of each plot. *95% CI of the intercept do not include 0, or the 95% CI of the slope do not 1, indicating significant deviation from the perfect fit.

3.2.2. Continuous outcomes

Results of the linear regression models on change in COMI, back pain, and leg pain are presented in **Table 1** (unstandardised regression coefficients) and **Fig. S12** (standardised coefficients). After adjusting for other factors included in the models, older age, male gender, and decompression with fusion (moderate effect size) were associated with greater improvement after surgery across all outcomes. Additionally, higher baseline COMI, back pain, and leg pain predicted greater improvement in their corresponding outcomes (moderate-large effects). On the contrary, patients with spinal stenosis and disc herniation with stenosis, history of previous surgeries, currently smoking (moderate effect), with higher morbidity class (moderate effect), and longer hospital stay had less improvement after surgery across all outcomes. Additionally, higher baseline COMI predicted less improvement in back and leg pain, higher baseline back pain predicted less improvement in COMI and leg pain, and operation time >3 hours (moderate effect) predicted less improvement in COMI and back pain. Adjusted effect sizes of these predictors were small (*betas* <0.25), unless specified otherwise.

Discrimination performance in the development models ranged from 13% (COMI change) to 22% (back and leg pain change) and the included set of predictors explained each outcome better than intercept-only models (*F*-test *ps* <0.001) (**Table 1**). Unadjusted portion of explained variance was higher in the validation than development data for COMI and back pain change, but lower for leg pain change outcome. Prediction accuracy was slightly worse in the validation compared to the development data with approximately 0.1 higher RMSEs across all outcomes.

Model calibration was very good in the development data (**Fig. S14b**), and plots in the validation data also showed close agreement between mean observed and predicted outcomes, although there was some degree of underestimation of the predicted changes in back and leg pain (**Fig 1b**). However, only the leg pain model calibration line significantly deviated from the perfect fit.

Individual predictions of outcomes based on the developed logistic and linear regression models can be made according to the equations provided in **Results S3**.

Compared to the primary analyses on the imputed data, sensitivity complete case analyses presented in **Results S4** and **Table S5** showed similar or worse model performance in the development and validation datasets, and no systematic differences in significant predictors except for the current smoking status which did not significantly predict any outcomes in the sensitivity analyses.

3.3. Development and validation of random forest models

Table 2 provides an overview of the RF performance measures across MCID (classification) and continuous change (regression) in COMI, back pain, and leg pain outcomes in the development and validation data.

Table 2. Random forests performance measures in the development and validation data across all outcomes.

	Classification outcomes			***Regression outcomes***		
	COMI MCID	*Back pain MCID*	*Leg pain MCID*	*COMI change*	*Back pain change*	*Leg pain change*
Development						
AUC (95% CI)	0.62 (0.60, 0.64)	0.66 (0.64, 0.68)	0.67 (0.64, 0.69)			
Sensitivity / Specificity (optimal cut-off)	0.51 / 0.68 (0.58)	0.67 / 0.56 (0.53)	0.71 / 0.53 (0.67)			
OOB error / RMSE	39.84%	38.72%	29.77%	2.89	2.94	3.33
pseudo-R^2				0.112	0.207	0.186
Calibration intercept (95% CI)	0.11 (0.04, 0.18)	0.01 (-0.08, 0.10)	0.01 (-0.12, 0.14)	0.04 (-0.46, 0.54)	-0.06 (-0.33, 0.22)	-0.18 (-0.64, 0.28)
Calibration slope (95% CI)	0.83 (0.71, 0.95)	1.00 (0.83, 1.16)	0.95 (0.76, 1.13)	0.98 (0.84, 1.13)	1.02 (0.92, 1.13)	1.05 (0.93, 1.17)
Pearson's r (observed vs. predicted)				0.33	0.45	0.43
Validation						
AUC (95% CI)	0.62 (0.60, 0.65)	0.68 (0.66, 0.71)	0.66 (0.63, 0.69)			
Sensitivity / Specificity (optimal cut-off)	0.56 / 0.64 (0.56)	0.86 / 0.39 (0.47)	0.75 / 0.50 (0.66)			
RMSE				2.94	3.04	3.33
pseudo-R^2				0.152	0.252	0.220
Calibration intercept (95% CI)	0.06 (-0.07, 0.18)	-0.06 (-0.17, 0.06)	0.01 (-0.16, 0.19)	-0.26 (-0.96, 0.45)	-0.04 (-0.42, 0.35)	-0.25 (-0.72, 0.22)
Calibration slope (95% CI)	0.94 (0.71, 1.16)	1.20 (0.99, 1.41)	0.96 (0.72, 1.21)	1.12 (0.91, 1.33)	1.14 (1.00, 1.29)	1.14 (1.02, 1.25)
Pearson's r (observed vs. predicted)				0.39	0.5	0.47

AUC, Area Under the Receiver-Operating Characteristic Curve; CI, confidence interval; MCID, Minimal Clinically Important Difference; OOB error, out-of-bag misclassification rate; RMSE, root mean square error.

3.3.1. Binary outcomes (MCID)

Misclassification rates ranged from 30% for leg pain to 40% for back pain and COMI MCID outcomes. Discrimination performance was in a similar range for development and validation data, with AUC consistently over 0.60, but still below acceptable discrimination across all outcomes, and consistently lower than AUC values obtained from logistic regression models. Outcome classification at the optimal probability threshold in the development and validation data across back and leg pain outcomes appeared to be biased towards higher sensitivity at the expense of specificity which oscillated near or below chance classification of no-MCID cases. This pattern was reversed for COMI outcome, where classification showed better specificity but near-chance sensitivity. The ROC curves are presented in **Fig. S15**.

Calibration plots demonstrated good agreement between mean observed and predicted MCID probabilities in the development data, however, the calibration line for the COMI outcome significantly deviated from the perfect fit suggesting some underestimation of the predicted MCID (**Fig. S14c**). Calibration of the COMI and back pain models in the validation data also suggested a small degree of underestimation of predicted outcome probabilities, but without significant deviations from the perfect fit (**Fig. 1c**). Overall, calibration plots supported good agreement between the observed and predicted probabilities of MCID for most quantiles in the validation data, similar to the calibration of the logistic regression models.

3.3.2. Continuous outcomes

The proportion of explained variance in change in COMI, back pain and leg pain ranged from 11% for COMI to 21% for back pain in the development data, and increased in the validation data (15-25%). These values were lower compared to R^2 from the linear regression models in the development data, but did not differ in the validation data except for the back pain model. RMSEs were only minimally higher in the RF regression, but with less discrepancy between the development and validation data.

RF regression models showed very good calibration in the development data (**Fig. S14d**). While there was also good agreement between mean observed and predicted outcomes in the validation data, calibration plots showed some degree of underestimation of the predicted relative to the observed changes in outcomes, in particular in the higher deciles, with the slope of the leg pain calibration line significantly deviating from the perfect fit (**Fig. 1d**). Tendency to underestimate predicted changes in the validation data was similar to that in the calibration of linear regression models, although underestimation of leg pain change was more pronounced in lower deciles. Correlation coefficients between ungrouped observed and predicted outcomes in the RF models were overall higher in the validation than development data, and lower compared to linear regression models.

3.3.3. Variable importance

Highest variable importance in RFs was generally assigned to the baseline scores on the corresponding outcome measures (except for COMI MCID), for instance, baseline back pain was most important for classifying back pain MCID and predicting change in back pain intensity (**Fig. S16**). While for pain intensity outcomes, these baseline scores appeared to be the sole most relevant predictors, COMI outcomes showed broader distribution of importance over different predictors. Across all outcomes, relatively high importance was also attributed to the duration of hospital stay, age, baseline scores on other outcome measures, current smoking status, type of degenerative disease, history of previous spinal surgeries, and morbidity. The same factors were found to have significant prognostic effects in the logistic and linear regression models.

4. Discussion

We developed and externally validated multivariate regression and RF models to predict patient-reported outcomes 3-24 months after lumbar spine surgery based on prospectively recorded medical and patient data. The models demonstrated good calibration in the temporal validation data, while their discrimination ability oscillated between acceptable and no-better-than-chance. Linear and logistic regression models performed better than RF algorithms, both in the development and validation data. The most important predictors included age, baseline COMI and pain scores, type of degenerative disease, previous surgeries, smoking, morbidity, and hospital stay.

This study brings a novel contribution to the field by assessing and comparing performance of linear and logistic regression models versus RF regression and classification algorithms, and validating them on external data, in order to further develop our ability to predict more precisely individual surgery outcomes. High number of participants and events per variable, which were limited in previous clinical prediction models (Christodoulou et al., 2019), add to the strength of the present work. Comparable performance and consistency in identified predictors demonstrate the robustness and generalisability of our models across different patient-reported outcomes (COMI, back, and leg pain MCID and continuous change scores) and modelling approaches.

4.1. Model performance

There was no substantial decrease in the models' performance on the new data relative to the development data, indicating no overfitting issues. Regression models predicting changes in back pain showed the best external validity, with acceptable discrimination (0.72) and 28% explained variance, followed by the models predicting leg pain outcomes (AUC 0.68, 22% explained variance). These metrics are comparable or better than in similar externally validated models predicting pain-related outcomes (AUC 0.52-0.83; 6-19%; Janssen et al., 2021; Quddusi et al., 2020; Staub et al., 2020). COMI models showed poorer discrimination (AUC 0.63, 16% explained variance), although comparable with external validity of another model relying on the same measure (17%; Staub et al., 2020), suggesting that composite outcomes like COMI may be more difficult to predict than, for instance, specific disability measures (AUC 0.71; Quddusi et al., 2020).

Similar studies relying on internal validation generally reported better discrimination (AUC 0.64-0.84; 23-49%; Abbott, 2011; Cobo Soriano et al., 2010; Khor et al., 2018; McGirt et al., 2015, 2017), highlighting a potential degree of over-optimism when model performance is only assessed on resampled or randomly-split data. Furthermore, model calibration was not always assessed (Abbott et al., 2011; Janssen et al., 2021; McGirt et al., 2017), but accurate prediction of outcomes can be particularly problematic in external data (Quddusi et al., 2020; Staub et al., 2020). Our models showed good calibration in the validation data across all outcomes, although there was a mild tendency to underestimate back pain MCID and leg pain reduction.

We found that RF did not outperform linear and logistic regression models. RF showed similarly good calibration in the validation data, yet none of the models reached acceptable discriminability in the validation (or development) data (0.62-0.68). This may reflect the difficulty of RF algorithms to extrapolate to new, untrained data, although previous relevant studies only achieved RF discrimination of 0.64-0.72 in internal validation (Karhade et al., 2021; Siccoli et al., 2019). Various machine learning classification approaches (e.g. elastic net penalised regression, deep neural networks, extreme gradient boosting, RF) have previously shown superior predictive performance compared to logistic regression (Couronné et al., 2018; Siccoli et al., 2019; Staartjes et al., 2019). However, consistent with our findings, a recent meta-analysis concluded that based on low risk of bias studies, performance of machine learning clinical prediction algorithms, including RF, does not

differ from logistic regression (their advantage was only found in high risk of bias studies) (Christodoulou et al., 2019). Our results extend this conclusion to linear regression versus RF regression.

There could be several reasons why RFs did not outperform statistical regression in the present study. Previous work demonstrating an advantage of machine learning over logistic regression did not cover external validation (Couronné et al., 2018; Siccoli et al., 2019; Staartjes et al., 2019), while regression models are likely to have better generalisability. Furthermore, machine learning works best for problems with high signal-to-noise ratio, which rarely characterises clinical data. Finally, since RFs show improved performance on data with nonlinear and nonadditive effects, any nonlinearities in the present data were likely not severe enough to be detrimental to statistical regression. Therefore, RFs might not show superior performance on large enough datasets satisfying the regression assumptions.

4.2. Relevant predictors

According to the regression models, greater odds of achieving MCID and larger reduction in COMI, back, and leg pain were significantly associated with older age, higher baseline score on the respective outcome measure, having decompression surgery with fusion, no stenosis, no history of previous spinal surgeries, lower morbidity class, not smoking, and shorter hospital stay. The same factors (except for surgical measures) were the most important predictors in RF analyses, with preoperative COMI, back, or leg pain scores leading across all models. Relevance of several of the identified predictors was also supported by previous research on patient-reported outcomes from the Spine Tango registry in other countries (Sobottke et al., 2017; Staub et al., 2020). Our results are also consistent with systematic reviews supporting prognostic value of age, preoperative pain intensity and disability, type of spinal pathology, previous surgeries, and smoking (Dorow et al., 2017; Halicka et al., *under review*; Rushton et al., 2018; Wilhelm et al., 2017; Wilson et al., 2016). In contrast, we did not find any effect of symptom duration, here recorded indirectly as treatment duration.

Low back pain and spinal surgery are complex clinical issues where multifactorial data is necessary to make accurate individualised predictions of treatment outcomes. Suboptimal model performance suggests that additional factors to those already recorded in registries such as Spine Tango might be needed to improve the predictive accuracy. For instance, other predictors identified in the above-mentioned systematic reviews, but not available in our data, included education level, compensation, duration of sick leave, sensory loss, comorbidities, and psychological pain-related and affective factors. Previous prediction models which achieved better discrimination (at least in internal validation) incorporated additional predictors, such as unemployment, medical insurance (although not applicable in the UK context), opioid use, antidepressants, mental functioning, optimism, control over pain, catastrophising, and postoperative psychomotor therapy (Abbott et al., 2011; Cobo Soriano et al., 2010; Karhade et al., 2021; Khor et al., 2018).

4.3. Limitations

The present study is not without limitations. Class imbalance is a common problem for machine learning classification algorithms, where selecting the outcome occurring more frequently increases overall classification accuracy, which may still be poor for the less frequent outcome (Staartjes & Schröder, 2018). This could potentially account for the sensitivity/specificity trade-off apparent in some of the MCID models, although not specific to RF.

Furthermore, there was missing data on several predictors, and although imputation diagnostics did not indicate any biases, complete case sensitivity analysis was inconsistent with respect to the

predictive value of smoking status, which had the highest missingness rate. This suggests that the imputed data may not accurately reflect the true smoking status in the population of interest. Thus, the missingness of the smoking status could potentially be related to other unmeasured variables. However, the significant prognostic value of smoking is consistent with several previous prediction models for spinal surgery outcomes (Janssen et al., 2021; Khor et al., 2018; McGirt et al., 2017; Staub et al., 2020).

Finally, although the developed models performed relatively well in the temporal validation, geographic validation is often more problematic (König et al., 2007). Thus, further research could include external validation of the prediction models across different neurosurgery centres.

4.4. Implications

The developed models demonstrated good ability to predict spinal surgery outcomes from new data, thus in practice, they could help identify patients at risk of poor outcomes. Such patients could be considered for additional interventions to improve their chance of recovery (Abbott et al., 2011). While all models appeared to be well-calibrated, and those predicting change in back pain showed the best performance on external validation, the discrimination ability of the leg pain and COMI models could be further improved, for instance, by including factors that previously demonstrated important contributions to spinal surgery outcomes. Modifiable preoperative predictors could be particularly useful for prospectively maximising the treatment benefit. The proposed models can therefore serve as a benchmark to inform future studies aimed at improving the accuracy of individual outcome prediction and potential revision of routinely collected information for spinal surgery registries.

4.5. Conclusions

We found comparable performance and consistent predictors across different outcomes, modelling approaches, and datasets. Regression models showed good calibration and acceptable to no-better-than-chance discrimination in the validation data. For similar datasets (with comparable set of predictors, sufficient sample size, and satisfying regression assumptions), RFs do not appear to outperform statistical regression. A strong advantage of statistical regression is its explanatory value and more easily interpretable prediction rules readily applicable in the clinical context. RFs, however, allow to establish relative predictor importance, which may assist in prioritising complex multifactorial data. Nonetheless, there is still room for improvement in terms of recorded predictor data.

Author contributions

Conceptualization: CB, MW, RD, MH; Data Curation: MH; Formal Analysis: MH; Funding Acquisition: CB, MW, RD; Investigation: MH; Methodology: MH, CB; Project Administration: MH, CB; Resources: MH, MW; Software: MH; Supervision: CB; Validation: MH, CB; Visualization: MH; Writing – Original Draft Preparation: MH, CB; Writing – Review & Editing: MH, CB, MW, RD. All authors discussed the results and commented on the manuscript.

References

Abbott, A. D., Tyni-Lenné, R., & Hedlund, R. (2011). Leg pain and psychological variables predict outcome 2–3 years after lumbar fusion surgery. *European Spine Journal*, *20*(10), 1626–1634. https://doi.org/10.1007/s00586-011-1709-6

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cobo Soriano, J., Sendino Revuelta, M., Fabregate Fuente, M., Cimarra Díaz, I., Martínez Ureña, P., & Deglané Meneses, R. (2010). Predictors of outcome after decompressive lumbar surgery and instrumented posterolateral fusion. *European Spine Journal*, *19*(11), 1841–1848. https://doi.org/10.1007/s00586-010-1284-2

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

de Campos, T. F. (2017). Low back pain and sciatica in over 16s: Assessment and management NICE Guideline [NG59]. *Journal of Physiotherapy*, *63*(2), 120.

Deyo, R. A., Battie, M., Beurskens, A., Bombardier, C., Croft, P., Koes, B., Malmivaara, A., Roland, M., Von Korff, M., & Waddell, G. (1998). Outcome measures for low back pain research: A proposal for standardized use. *Spine*, *23*(18), 2003–2013.

Dorow, M., Lobner, M., Stein, J., Konnopka, A., Meisel, H. J., Gunther, L., Meixensberger, J., Stengler, K., Konig, H. H., & Riedel-Heller, S. G. (2017). Risk Factors for Postoperative Pain Intensity in Patients Undergoing Lumbar Disc Surgery: A Systematic Review. *PLoS One*, *12*(1), e0170303. https://doi.org/10.1371/journal.pone.0170303

Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., Kerns, R. D., Stucki, G., Allen, R. R., & Bellamy, N. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain*, *113*(1), 9–19.

Farrar, J. T., Young, J. P., LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*, *94*(2), 149–158.

Ferrer, M., Pellisé, F., Escudero, O., Alvarez, L., Pont, A., Alonso, J., & Deyo, R. (2006). Validation of a Minimum Outcome Core Set in the Evaluation of Patients With Back Pain. *Spine*, *31*(12), 1372–1379. https://doi.org/10.1097/01.brs.0000218477.53318.bc

Haji-Maghsoudi, S., Haghdoost, A., Rastegari, A., & Baneshi, M. R. (2013). Influence of pattern of missing data on performance of imputation methods: An example using national data on drug injection in prisons. *International Journal of Health Policy and Management*, *1*(1), 69.

Halicka, M., Duarte, R., Catherall, S., Maden, M., Coetsee, M., Wilby, M., & Brown, C. (under review). *Predictors of pain and disability outcomes following spinal surgery for chronic low back and radicular pain: A systematic review*.

Harrell, F. E. (2015). Multivariable modeling strategies. In *Regression modeling strategies* (pp. 63–102). Springer.

Hegarty, D., & Shorten, G. (2012). Multivariate prognostic modeling of persistent pain following lumbar discectomy. *Pain Physician*, *15*(5), 421–434.

Hosmer, D. W. J., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Janssen, E. R., Osong, B., Soest, J. van, Dekker, A., Meeteren, N. L. van, Willems, P. C., & Punt, I. M. (2021). Exploring Associations of Preoperative Physical Performance With Postoperative Outcomes After Lumbar Spinal Fusion: A Machine Learning Approach. *Archives of Physical Medicine and Rehabilitation*, *102*(7), 1324-1330.e3. https://doi.org/10.1016/j.apmr.2021.02.013

Karhade, A. V., Fogel, H. A., Cha, T. D., Hershman, S. H., Doorly, T. P., Kang, J. D., Bono, C. M., Harris, M. B., Schwab, J. H., & Tobert, D. G. (2021). Development of prediction models for clinically meaningful improvement in PROMIS scores after lumbar decompression. *The Spine Journal*, *21*(3), 397–404. https://doi.org/10.1016/j.spinee.2020.10.026

Khor, S., Lavallee, D., Cizik, A. M., Bellabarba, C., Chapman, J. R., Howe, C. R., Lu, D., Mohit, A. A., Oskouian, R. J., Roh, J. R., Shonnard, N., Dagal, A., & Flum, D. R. (2018). Development and Validation of a Prediction Model for Pain and Functional Outcomes After Lumbar Spine Surgery. *JAMA Surgery*, *153*(7), 634–642. https://doi.org/10.1001/jamasurg.2018.0072

König, I. R., Malley, J. D., Weimar, C., Diener, H.-C., & Ziegler, A. (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine*, *26*(30), 5499–5511. https://doi.org/10.1002/sim.3069

Maniadakis, N., & Gray, A. (2000). The economic burden of back pain in the UK. *PAIN*, *84*(1), 95–103. https://doi.org/10.1016/S0304-3959(99)00187-6

Mannion, A. F., Elfering, A., Staerkle, R., Junge, A., Grob, D., Semmer, N. K., Jacobshagen, N., Dvorak, J., & Boos, N. (2005). Outcome assessment in low back pain: How low can you go? *European Spine Journal*, *14*(10), 1014–1026.

Mannion, A. F., Porchet, F., Kleinstück, F. S., Lattig, F., Jeszenszky, D., Bartanusz, V., Dvorak, J., & Grob, D. (2009). The quality of spine surgery from the patient's perspective: Part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index. *European Spine Journal*, *18*(3), 374–379.

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology*, *10*(1), 7. https://doi.org/10.1186/1471-2288-10-7

McGirt, M. J., Bydon, M., Archer, K. R., Devin, C. J., Chotai, S., Parker, S. L., Nian, H., Harrell, F. E., Jr., Speroff, T., Dittus, R. S., Philips, S. E., Shaffrey, C. I., Foley, K. T., & Asher, A. L. (2017). An analysis from the Quality Outcomes Database, Part 1. Disability, quality of life, and pain outcomes following lumbar spine surgery: Predicting likely individual patient outcomes for

shared decision-making. *Journal of Neurosurgery. Spine*, *27*(4), 357–369. https://doi.org/10.3171/2016.11.SPINE16526

McGirt, M. J., Sivaganesan, A., Asher, A. L., & Devin, C. J. (2015). Prediction model for outcome after low-back surgery: Individualized likelihood of complication, hospital readmission, return to work, and 12-month improvement in functional disability. *Neurosurgical Focus*, *39*(6), E13. https://doi.org/10.3171/2015.8.FOCUS15338

Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*, *162*(1), W1–W73. https://doi.org/10.7326/M14-0698

Moons, K. G. M., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. *PLoS Medicine*, *11*(10), e1001744. https://doi.org/10.1371/journal.pmed.1001744

Quddusi, A., Eversdijk, H. A. J., Klukowska, A. M., de Wispelaere, M. P., Kernbach, J. M., Schröder, M. L., & Staartjes, V. E. (2020). External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. *European Spine Journal*, *29*(2), 374–383. https://doi.org/10.1007/s00586-019-06189-6

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rice, A. S., Smith, B. H., & Blyth, F. M. (2016). Pain and the global burden of disease. *Pain*, *157*(4), 791–796.

Röder, C., Chavanne, A., Mannion, A. F., Grob, D., & Aebi, M. (2005). SSE Spine Tango—Content, workflow, set-up. Www.eurospine.org-Spine Tango. *European Spine Journal*, *14*(10), 920–924. https://doi.org/10.1007/s00586-005-1023-2

Röder, C., El-Kerdi, A., Grob, D., & Aebi, M. (2002). A European spine registry. *European Spine Journal*, *11*(4), 303–307. https://doi.org/10.1007/s00586-002-0453-3

Rushton, A., Zoulas, K., Powell, A., & Staal, J. B. (2018). Physical prognostic factors predicting outcome following lumbar discectomy surgery: Systematic review and narrative synthesis. *BMC Musculoskeletal Disorders*, *19*(1), 326. https://doi.org/10.1186/s12891-018-2240-2

Siccoli, A., de Wispelaere, M. P., Schröder, M. L., & Staartjes, V. E. (2019). Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurgical Focus*, *46*(5), E5. https://doi.org/10.3171/2019.2.FOCUS18723

Siontis, G. C. M., Tzoulaki, I., Castaldi, P. J., & Ioannidis, J. P. A. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, *68*(1), 25–34. https://doi.org/10.1016/j.jclinepi.2014.09.007

Sobottke, R., Herren, C., Siewe, J., Mannion, A. F., Röder, C., & Aghayev, E. (2017). Predictors of improvement in quality of life and pain relief in lumbar spinal stenosis relative to patient age: A study based on the Spine Tango registry. *European Spine Journal*, *26*(2), 462–472. https://doi.org/10.1007/s00586-015-4078-8

Staartjes, V. E., de Wispelaere, M. P., Vandertop, W. P., & Schröder, M. L. (2019). Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: Feasibility of center-specific modeling. *The Spine Journal*, *19*(5), 853–861. https://doi.org/10.1016/j.spinee.2018.11.009

Staartjes, V. E., & Schröder, M. L. (2018). Letter to the Editor. Class imbalance in machine learning for neurosurgical outcome prediction: Are our models valid? *Journal of Neurosurgery: Spine*, *29*(5), 611–612. https://doi.org/10.3171/2018.5.SPINE18543

Staub, L. P., Aghayev, E., Skrivankova, V., Lord, S. J., Haschtmann, D., & Mannion, A. F. (2020). Development and temporal validation of a prognostic model for 1-year clinical outcome after decompression surgery for lumbar disc herniation. *European Spine Journal*, *29*(7), 1742–1751. https://doi.org/10.1007/s00586-020-06351-5

van der Ploeg, T., Nieboer, D., & Steyerberg, E. W. (2016). Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *Journal of Clinical Epidemiology*, *78*, 83–89. https://doi.org/10.1016/j.jclinepi.2016.03.002

Weir, S., Samnaliev, M., Kuo, T.-C., Choitir, C. N., Tierney, T. S., Cumming, D., Bruce, J., Manca, A., Taylor, R. S., & Eldabe, S. (2017). The incidence and healthcare costs of persistent postoperative pain following lumbar spine surgery in the UK: a cohort study using the Clinical Practice Research Datalink (CPRD) and Hospital Episode Statistics (HES). *BMJ Open*, *7*(9), e017585.

Wilhelm, M., Reiman, M., Goode, A., Richardson, W., Brown, C., Vaughn, D., & Cook, C. (2017). Psychological Predictors of Outcomes with Lumbar Spinal Fusion: A Systematic Literature Review. *Physiotherapy Research International*, *22*(2). https://doi.org/10.1002/pri.1648

Wilson, C. A., Roffey, D. M., Chow, D., Alkherayf, F., & Wai, E. K. (2016). A systematic review of preoperative predictors for postoperative clinical outcomes following lumbar discectomy. *The Spine Journal*, *16*(11), 1413–1422. https://doi.org/10.1016/j.spinee.2016.08.003

Wright, M. N., Wager, S., & Probst, P. (2021). *ranger: A Fast Implementation of Random Forests* (0.13.1) [Computer software]. https://CRAN.R-project.org/package=ranger

Zweig, T., Enke, J., Mannion, A. F., Sobottke, R., Melloh, M., Freeman, B. J., Aghayev, E., & Spine Tango, C. (2017). Is the duration of pre-operative conservative treatment associated with the clinical outcome following surgical decompression for lumbar spinal stenosis? A study based on the Spine Tango Registry. *European Spine Journal*, *26*(2), 488–500. https://doi.org/10.1007/s00586-016-4882-9