

Machine Learning versus Logistic Regression for Predicting Patient-Reported Outcomes After Lumbar Spinal Stenosis Surgery:

Insights from the Swespine Registry

Eric Brisby Enger (Gothenburg, Suède), Alexander Berman (Gothenburg, Suède),
Fredrik Johansson (Gothenburg, Suède), Catharina Parai (Gothenburg, Suède)

Introduction

Lumbar spinal stenosis is a common indication for surgery, yet a substantial proportion of patients do not improve postoperatively. One way of improving outcome might be the use of a clinical decision support system to assist doctors and patients in their decision to operate or not to operate. This study aimed to compare logistic regression (LR), which is used in an existing tool called the Swespine Dialogue Support to the more advanced machine learning (ML) algorithm XGBoost. XGBoost was selected as it has previously shown a better fit than other ML algorithms.

Materials and Methods

Data were retrieved from the Swedish national spine registry (Swespine) for patients undergoing lumbar spinal stenosis surgery between 2016 and 2023 ($n = 32,075$). Outcomes were one-year global assessment of leg pain (GA) and satisfaction (SAT). Twenty-six baseline covariates were included. Missing values in continuous variables were multiply imputed; categorical missingness was coded as indicators. Models were trained on 90% of the data and evaluated on a 10% holdout set, with hyperparameters tuned via cross-validation. Performance was assessed using AUROC, calibration plots and accuracy.

Results

After preprocessing, 20,385 observations were available for GA and 21,601 for SAT. For GA, LR achieved an AUROC of 0.68 (95% CI: 0.65–0.70) and accuracy of 0.65, compared with XGB AUROC 0.69 (95% CI: 0.67–0.71) and accuracy 0.66. For SAT, LR reached AUROC 0.66 (95% CI: 0.63–0.68) and accuracy 0.67, compared with XGB AUROC 0.67 (95% CI: 0.64–0.69) and accuracy 0.67. Calibration plots demonstrated a high degree of concordance in both models. Both models maintained stable performance even when restricted to approximately 6–8 key variables, with ODI, pain duration, prior surgery, and age among the most influential predictors.

Conclusion

XGBoost provided no clear improvement in predictive performance compared to logistic regression, both models achieved only moderate discrimination. The findings suggest that current registry data may be insufficient for precise individual-level predictions but also indicate that simplified models using as few as 6–8 variables perform comparably to the existing tool. In clinical practice, outcome prediction tool—such as the Dialogue Support—should be regarded as complements to, rather than replacements for, the surgeon's judgment and shared decision-making with patients.