Performance of Artificial Intelligence in Determining Candidacy for Lumbar Stenosis Surgery

Raphaël Mourad[1,2], Serhii Kolisnyk[3], Olga Suldina[1], Jack Kim[1], Andrej Rusakov[1] and Darren Lebl[4]

November 6, 2021

[1] Remedy Logic, 1177 Avenue of the Americas, 5th Floor New York, NY, 10036
[2] University of Toulouse, CNRS, UPS, 31062 Toulouse, France
[3] Vinnitsa National Medical University, Pyrohova St, 56, Vinnytsia, Vinnytsia Oblast, Ukraine, 21018
[4] Hospital for Special Surgery. 535 East 70th Street, New York, NY 10021

Corresponding authors: raphael.mourad@univ-tlse3.fr; drlebl@gmail.com.

Key words: lumbar spinal stenosis; spinal surgery; artificial intelligence.

Abstract

Lumbar spinal stenosis (LSS) is a condition affecting several hundreds of thousands of adults in the US each year. LSS is associated with significant economic burden and a relatively high rate of clinical post-operative failure. We hypothesized that the performance of artificial intelligence (AI) methods could prove comparable to that of a panel of spine experts. We propose a novel AI model which computes the probability of spinal surgery recommendation for LSS based on clinical symptoms, MRI findings, and patient demographic factors. The model demonstrated high prediction accuracy, with a root mean square error (RMSE) between model predictions and ground truth of 0.185, while the average RMSE between individual doctor's recommendations and ground truth was 0.251. For dichotomous classification, the AUROC was 0.899, while the average metric based on individual doctor's recommendations was 0.853. Our results suggest that AI can be used to automate the evaluation of surgical candidacy for LSS with performance comparable to the expert panel.

1 Introduction

Lumbar spinal stenosis (LSS) is one of the most common conditions affecting more than 200,000 adults in the US each year [1]. Over the past 20 years, the use of spinal surgery as a treatment course has increased significantly in many Western countries [2]. LSS is associated with a considerable economic burden, with an estimated $40 billion spent on surgery each year, with a relatively high rate of clinical postoperative failure [3]. For instance, the overall failure rate of lumbar spine surgery was estimated to be $10\% - 46\%$ [4]. Moreover, no more than 30%, 15%, and 5% of the patients experience a successful outcome after the second, third, and fourth surgeries, respectively [5].

Artificial intelligence is currently revolutionizing decision making in many fields and industries, and medicine in particular [6]. AI-powered medical solutions are appealing, as they enable predictive, preventive, personalized, and participatory medicine [6]. Regarding spinal surgery, the use of AI has been a relatively recent development. Most applications include image classification (for instance to automatically detect vertebral body compression fractures on imaging), preoperative risk prediction models and clinical decision support tools [7,8]. Recently, a machine learning approach based on lasso logistic regression was used to predict complications after spinal surgery depending on patient variables with AUROC ranging from 0.7 to 0.76, and found the health insurance provider as the best predictor [9]. Another regression model obtained similar results, with ASA score (physical status score) as the best predictor [10]. Other models were also proposed to predict pain and functional outcomes after surgery [11–14].

Here, we propose a novel AI model to compute the probability to accept or decline spinal surgery for LSS based on a wide range of variables encompassing critical clinical symptoms, MRI findings, and patient demographic factors. Two hundred clinical vignettes were developed for validation. A separate panel of five doctors independently reviewed the vignettes to determine surgery recommendation probability independent of each other. The proposed model demonstrated high prediction accuracy, with a root mean square error (RMSE) between model predictions and ground truth of 0.185, while the average RMSE between individual doctor's recommendations and ground truth was 0.251. For dichotomous classification (with no or weak vs, strong recommendation), the AUROC was 0.899, while the average metric based on individual doctor's recommendations was 0.853. Our results suggest that AI can be used to automate surgical candidacy for LSS with performance comparable to the expert panel.

2 Materials and Methods

2.1 Medical vignettes

We first compiled a set of variables reflecting clinical symptoms, MRI findings, and patient demographic factors, using medical literature together with the expert opinions of a multidisciplinary team of doctors in the fields of spinal surgery, rehabilitation medicine, interventional and diagnostic radiology.

Using the set of variables, we then generated a set of 200 vignettes which summarized potential realistic patient profiles, while accounting for correlations among the variables. We generated vignettes with probabilities of surgery recommendation ranging from low to high probability.

2.2 Bayesian network

Based on the set of variables from medical vignettes, we developed a Bayesian network to determine the probability to accept or decline spinal surgery for LSS. The Bayesian network was then used to compute the probability of surgery recommendation for each medical vignette.

2.3 Reviewing of vignettes by an independent panel of doctors

The 200 medical vignettes were reviewed by an independent panel of five spinal surgeons from different medical clinics to in order to determine the probability of surgery recommendation for each medical vignette. Each surgeon was asked independently to review each vignette and recommend surgery with a score between 0 (surgery must not be done) to 1 (surgery must be done).

2.4 Data analysis

All data analyses were done using R 3.6.3.

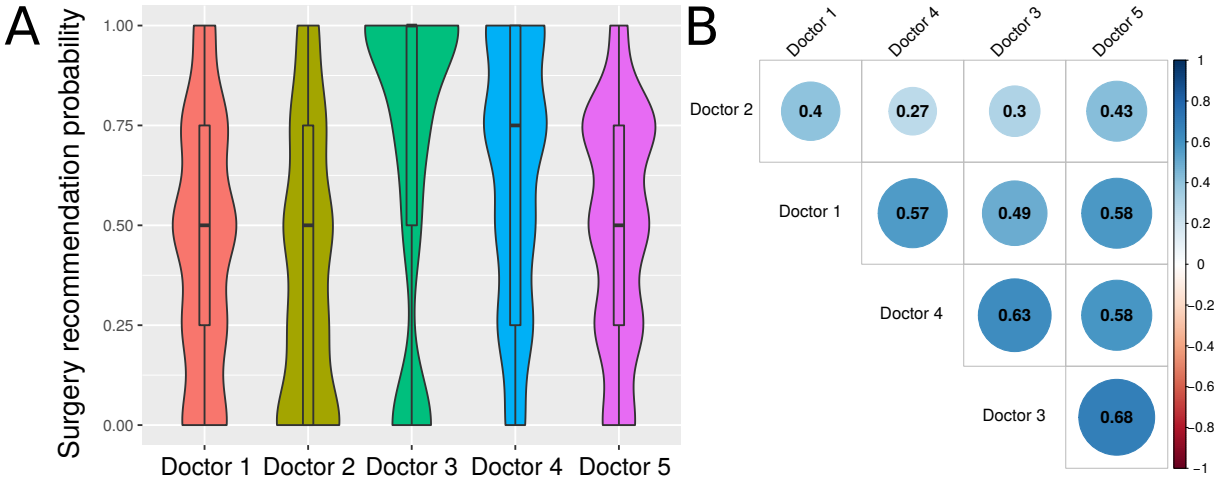

Figure 1. Uni- and bi-variate analyses of doctor recommendations from medical vignettes. A) Violin plot of doctor recommendations. B) Pearson correlations between doctors' recommendations.

3 Results and Discussion

3.1 Analysis of independent doctors' recommendation probabilities

An independent panel of five doctors reviewed 200 medical vignettes to determine the surgery recommendation probability for each vignette. In Figure 1A, we plotted univariate analyses of doctors' recommendations. Overall, we observed that doctors' recommendation probabilities were spread between 0 and 1, and centered around 0.5 for most doctors, except for doctors 3 and 4, whose recommendations were skewed toward high probabilities. We then ran bivariate analyses between doctors and found that doctors' recommendation probabilities were positively but moderately correlated (Figure 1B). The average pairwise correlation was 0.47, the lowest correlation was 0.27 between doctors 2 and 4, while the highest correlation was 0.68 between doctors 3 and 5.

These results thus suggest that, although doctors' recommendations were positively correlated, the agreement between doctors was moderate and some doctors were biased towards high recommendation probabilities, reflecting a high level of heterogeneity in individual doctor recommendations.

3.2 Comparison between model and individual doctor recommendation probabilities

We then sought to assess the accuracy of the Bayesian network model to predict surgery recommendations, in comparison to individual doctor recommendations. For this purpose, for each vignette, the ground truth probability for surgery recommendation was calculated as the average between the five independent doctors' recommendation probabilities. The model was also used to predict the recommendation probability for the same vignettes. Since the Bayesian network was not trained using the vignettes, we could use the whole set of vignettes for model validation.

The root mean square error (RMSE) between the model prediction and ground truth probabilities was 0.185 (Figure 2A). The Pearson correlation and the R2 were 0.778 and 0.606, respectively. When plotting the linear regression $y = ax + b$ (assuming a linear relation between model prediction and ground truth) with $y = x$ (assuming perfect agreement between model prediction and ground truth), we observed

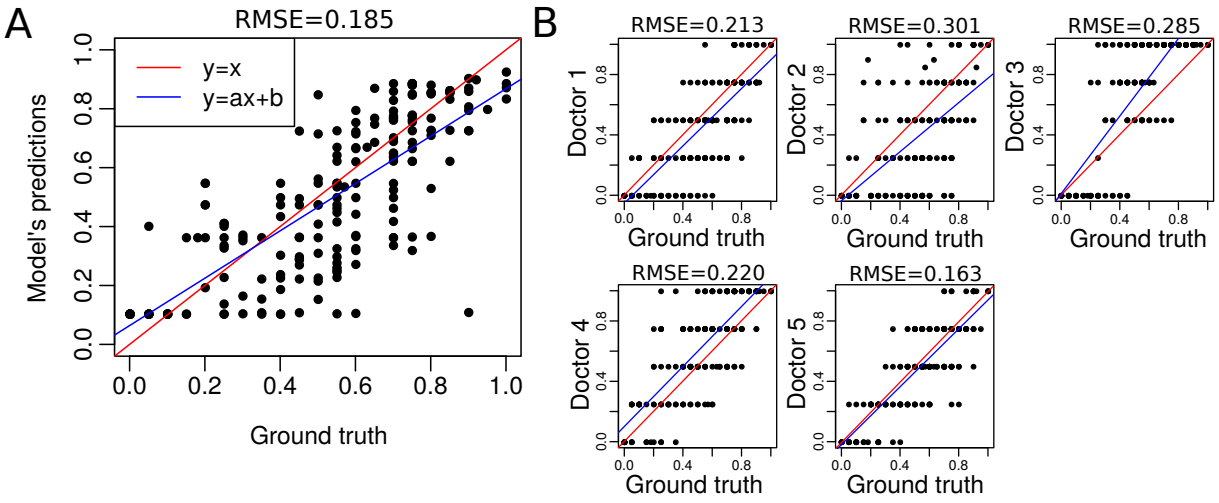

Figure 2. Comparison of prediction performance between the model and individual doctors for recommendation probability. A) Scatter plot between model's recommendation probability and ground truth recommendation probability. B) Scatter plots between individual doctor's recommendation probability and ground truth recommendation probability.

that the model had the tendency to slightly overestimate low ground truth probabilities (when surgery should not be done), while slightly underestimating high ground truth probabilities (when surgery should be done).

The average RMSE between individual doctor recommendations and ground truth was 0.251 (Figure 2B). The average Pearson correlation and the average R2 were 0.771 and 0.594, respectively. When plotting the linear regression $y = ax+b$ (assuming a linear relation between individual doctor recommendation and ground truth) with $y = x$ (assuming perfect agreement between individual doctor recommendation and ground truth), we observed that the doctors 1 and 2 were globally underestimating the ground truth probabilities contrary to doctors 3 and 4.

When predicting surgery recommendation probabilities, our validation performed on 200 vignettes revealed that the AI model we built performed comparably to individual doctor recommendations.

3.3 Comparison between model and individual doctor recommendations in dichotomous classification

We then setup the probability predictions as a dichotomous classification, where we combined the predictions to two classes, no or weak recommendation vs strong recommendation, with a probability threshold of 0.66.

The AUROC between model and ground truth recommendations was 0.899, while the sensitivity and specificity were 0.712 and 0.874, revealing good accuracy metrics. The Cohen's kappa was 0.595, nearly substantial. In comparison, the average AUROC based on individual doctor's recommendations was 0.853, and the sensitivity and specificity were 0.764 and 0.743, respectively. Most doctors showed an AUROC over 0.85, except one who presented an AUROC of 0.812. Cohen's kappa was 0.490, showing moderate agreement.

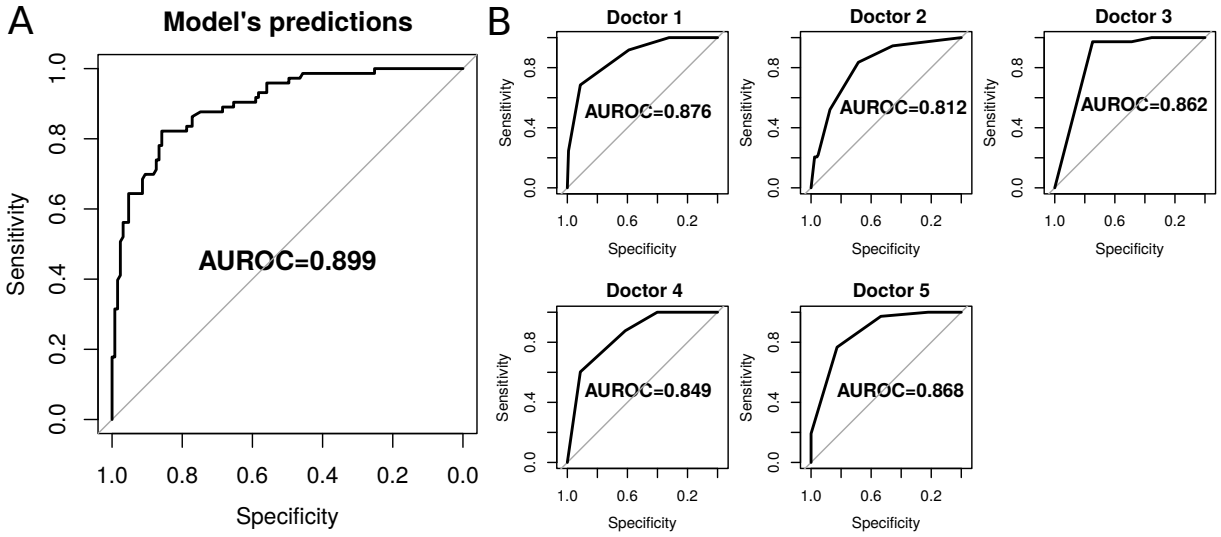

Figure 3. Comparison of prediction performance between the model and individual doctors, in a dichotomous classification setting. A) Receiver operating characteristic curve (ROC) between model's recommendation and ground truth recommendation to classify between no or weak recommendation versus strong recommendation. The area under the ROC (AUROC) is plotted. B) ROC curves between individual doctor's recommendation and ground truth recommendation. AUROC is plotted.

In a dichotomous classification setting, these results reveal that our model performed comparably to individual doctors.

4 Conclusion

In this article, we propose a novel artificial intelligence (AI) model to predict surgery recommendation based on variables reflecting clinical symptoms, MRI findings and patient demographic factors. The model shows surgery recommendation accuracy metrics that are comparable to recommendations from an independent expert panel. These results suggest that AI can be used to bring efficiency and automation to the decision making process for determining surgical candidacy for LSS, with similar performance to physicians.

Compared to previously published models predicting complication risks, pain reduction or functional outcomes, our model directly computes the probability to admit a patient to spinal surgery, thus representing a more straightforward solution in the medical decision process.

Despite the merits of this study, there are some limitations. First, the study was assessed on a small set of 200 medical vignettes, and not on patient data. Future studies will be carried out using patient data from a large cohort for more rigorous validation. Second, building a model using expert knowledge requires extensive efforts for expert elicitation and for obtaining a consensus among doctors, whereas alternative machine learning methods have the advantage of training the model directly from data. Third, the model was focused on spinal surgery for LSS. Future research will further extend the model to predict surgery recommendation for other spinal conditions, such as lumbar disc herniation and spinal instability.

References

[1] Ai-Min Wu, Fei Zou, Yong Cao, Dong-Dong Xia, Wei He, Bin Zhu, Dong Chen, Wen-Fei Ni, Xiang-Yang Wang, and Kenny Kwan. Lumbar spinal stenosis: an update on the epidemiology, diagnosis and treatment. *AME Medical Journal*, 2(5), 2017.

[2] Margreth Grotle, Milada Cvancarova Småstuen, Olaf Fjeld, Lars Grøvle, Jon Helgeland, Kjersti Storheim, Tore K Solberg, and John-Anker Zwart. Lumbar spine surgery across 15 years: trends, complications and reoperations in a longitudinal observational study from Norway. *BMJ Open*, 9(8), 2019.

[3] Daniell James R. and Osti Orso L. Failed back surgery syndrome: A review article. *Asian Spine Journal*, 12(2):372–379, 2018.

[4] Simon Thomson. Failed back surgery syndrome – definition, epidemiology and demographics. *British Journal of Pain*, 7(1):56–59, 2013. PMID: 26516498.

[5] Alf L Nachemson. Evaluation of results in lumbar spine surgery. *Acta Orthopaedica Scandinavica*, 64(sup251):130–133, 1993.

[6] Giovanni Briganti and Olivier Le Moine. Artificial intelligence in medicine: Today and tomorrow. *Frontiers in Medicine*, 7:27, 2020.

[7] Michael Chang, Jose A. Canseco, Kristen J. Nicholson, Neil Patel, and Alexander R. Vaccaro. The role of machine learning in spine surgery: The future is now. *Frontiers in Surgery*, 7:54, 2020.

[8] Daniel Lubelski, Andrew Hersh, Tej D. Azad, Jeff Ehresman, Zachary Pennington, Kurt Lehner, and Daniel M. Sciubba. Prediction models in degenerative spine surgery: A systematic review. *Global Spine Journal*, 11(1_suppl):79S–88S, 2021. PMID: 33890803.

[9] Summer S. Han, Tej D. Azad, Paola A. Suarez, and John K. Ratliff. A machine learning approach for predictive models of adverse events following spine surgery. *The Spine Journal*, 19(11):1772–1781, November 2019.

[10] Pascal Zehnder, Ulrike Held, Tim Pigott, Andrea Luca, Markus Loibl, Raluca Reitmeir, Tamás Fekete, Daniel Haschtmann, and Anne F. Mannion. Development of a model to predict the probability of incurring a complication during spine surgery. *European Spine Journal*, 30(5):1337–1354, May 2021.

[11] Ho-Joong Kim, Joon-Hee Park, Jang-Woo Kim, Kyoung-Tak Kang, Bong-Soon Chang, Choon-Ki Lee, and Jin S. Yeom. Prediction of postoperative pain intensity after lumbar spinal surgery using pain sensitivity and preoperative back pain severity. *Pain Medicine*, 15(12):2037–2045, 12 2014.

[12] Matthew J. McGirt, Ahilan Sivaganesan, Anthony L. Asher, and Clinton J. Devin. 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 FOC*, 39(6):E13, 2015.

[13] Sara Khor, Danielle Lavallee, Amy M. Cizik, Carlo Bellabarba, Jens R. Chapman, Christopher R. Howe, Dawei Lu, A. Alex Mohit, Rod J. Oskouian, Jeffrey R. Roh, Neal Shonnard, Armagan Dagal, and David R. Flum. Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. *JAMA Surgery*, 153(7):634–642, 07 2018.

[14] Sarah J. Gilmore, Andrew J. Hahne, Megan Davidson, and Jodie A. McClelland. Predictors of substantial improvement in physical function six months after lumbar surgery: is early post-operative walking important? A prospective cohort study. *BMC Musculoskeletal Disorders*, 20(1):418, Sep 2019.