Performance of hybrid artificial intelligence in determining candidacy for lumbar stenosis surgery

Raphael Mourad1,2 · Serhii Kolisnyk3 · Yurii Baiun4 · Alessandra Falk2 · Titenkov Yuriy5 · Frolov Valerii5 · Aleksey Kopeev5 · Olga Sulchina6 · Andrey Pospelov2 · Jack Kim2 · Andrej Rusakov2 · Darren R. Lebl7

Received: 9 March 2022 / Revised: 16 May 2022 / Accepted: 24 June 2022 / Published online: 8 July 2022 © The Author(s) 2022

Abstract
Purpose Lumbar spinal stenosis (LSS) is a condition affecting several hundreds of thousands of adults in the United States each year and is associated with significant economic burden. The current decision-making practice to determine surgical candidacy for LSS is often subjective and clinician specific. In this study, we hypothesize that the performance of artificial intelligence (AI) methods could prove comparable in terms of prediction accuracy to that of a panel of spine experts.
Methods We propose a novel hybrid AI model which computes the probability of spinal surgical recommendations for LSS, based on patient demographic factors, clinical symptom manifestations, and MRI findings. The hybrid model combines a random forest model trained from medical vignette data reviewed by surgeons, with an expert Bayesian network model built from peer-reviewed literature and the expert opinions of a multidisciplinary team in spinal surgery, rehabilitation medicine, interventional and diagnostic radiology. Sets of 400 and 100 medical vignettes reviewed by surgeons were used for training and testing.
Results The model demonstrated high predictive accuracy, with a root mean square error (RMSE) between model predictions and ground truth of 0.0964, while the average RMSE between individual doctor’s recommendations and ground truth was 0.1940. For dichotomous classification, the AUROC and Cohen’s kappa were 0.9266 and 0.6298, while the corresponding average metrics based on individual doctor’s recommendations were 0.8412 and 0.5659, respectively.
Conclusions Our results suggest that AI can be used to automate the evaluation of surgical candidacy for LSS with performance comparable to a multidisciplinary panel of physicians.

Keywords Lumbar spinal stenosis · Spinal surgery · Artificial intelligence

Introduction
Lumbar spinal stenosis (LSS) is one of the most common conditions affecting more than 200,000 adults in the United States each year [1]. Over the last 20 years, spine pathologies such as LSS have increased significantly in many Western countries due to an aging and sedentary lifestyle, increasing the need for spine surgery as treatment [2]. While there is a broad body of literature that supports the effectiveness of spinal surgery for late clinical outcomes after one year [3], in some cases, adverse outcomes can occur due to misdiagnosis [4], suboptimal patient selection [5] or subjectivity of clinical and radiographic assessment [6, 7]. The overall failure rate of lumbar spine surgery was estimated to be 10–46% [2, 8]. Another issue encountered is the associated considerable economic burden, with an estimated $40 billion spent on spinal fusions each year in the US [9], which...
provides challenges for both the individual and insurance authorization. Considering these challenges, spinal surgery recommendations should be a careful and judicious consideration taking into account a multidimensional picture of the patient’s health, clinical symptoms, comorbidities, and imaging findings.

Artificial intelligence is currently revolutionizing decision making in many different industries including health care [10]. AI-powered medical solutions have the potential to enable predictive, preventive, personalized, and participatory medicine [10]. Applications in spinal surgery with significant impact are beginning to unfold. Some early applications have included image classification such as the automated detection of vertebral body compression fractures on imaging studies, preoperative risk prediction models, and clinical decision support tools [11, 12].

Here, we propose a novel AI model to compute the probability to recommend spinal surgery for LSS that is in concordance with surgeon decisions. Compared to previously published models predicting complication risks [13, 14], our model has the potential to directly compute the probability to admit a patient to spinal surgery, thus representing an effective augmentation tool to the medical decision making process. The model consists of a hybrid approach combining (i) a random forest model to accurately estimate model parameters from medical vignette data reviewed by surgeons, with (ii) an expert Bayesian network model implementing surgical recommendations from peer-reviewed published literature together with the expert opinions of a multidisciplinary team in the fields of spinal surgery, rehabilitation medicine, interventional and diagnostic radiology. We hypothesized that the performance of our proposed artificial intelligence (AI) methods could prove comparable to that of a panel of spine experts.

**Materials and methods**

**Medical vignettes**

A set of 36 variables representing clinical symptoms, MRI findings, and patient demographic factors were compiled, using medical literature together with the expert input of a multidisciplinary team of doctors in the fields of spinal surgery, rehabilitation medicine, interventional and diagnostic radiology (Supplementary File 1).

Using these set of variables, a set of 500 vignettes which represent realistic patient profiles were created, while accounting for critical correlations between the variables (Supplementary File 2). The generated vignettes were designed to provide a range of probabilities for surgical recommendation ranging from low to high probability.

**Review of vignettes by an independent panel of doctors**

The 500 medical vignettes were reviewed by an independent panel of five spinal surgeons from different medical practices in order to determine the probability of surgical recommendation for each medical vignette. Each surgeon was asked independently to review each vignette and recommend surgery with a score from 0 (surgery must not be done) to 100 (surgery must definitely be done), and then the score was divided by 100. Note that this panel of surgeons was independent from the multidisciplinary team in spinal surgery, rehabilitation medicine, interventional and diagnostic radiology used to build the vignettes and the Bayesian network (see below).

**Bayesian network (expert model)**

Based on the set of variables from the medical vignettes, a Bayesian network was built to compute the probability to recommend spinal surgery for LSS (Supplementary File 3). For this purpose, GeNiE Modeler from BAYESFUSION was used (https://www.bayesfusion.com/genie/). The Bayesian network structure and parameters were not trained using patient data nor vignettes, but were determined using peer-reviewed medical literature and doctor opinions. The Bayesian network was used to compute the probability of surgical recommendation for each medical vignette.

**Random forest (machine learning model)**

Using the medical vignettes reviewed by doctors, a random forest model was trained to predict the probability of surgical recommendation from the set of 36 variables. Vignettes were randomly split into 70% for fine-tuning and training the random forest, 10% to estimate the hybrid model weights (see below), and 20% for testing predictions.

Variable importance was computed using the mean decrease in accuracy in the out-of-bag sample during training.

Hyper-parameters min.node.size = 3, sample.fraction = 0.88 and mtry = 24 were obtained by fine-tuning with fivefold cross-validation. Split rule “variance” was used.

**Hybrid model**

A hybrid approach was constructed by a weighted average of the predictions from the Bayesian network (expert model) and the random forest (machine learning model). A linear regression was used to stack the predictions. Regression coefficients (after normalizing their sum to 1) were used as
weights for the weighted average, estimated from 10% of the vignettes. The hybrid model therefore combines medical expert knowledge from the Bayesian network with the machine learning findings directly inferred from the data.

Data analysis

All data analyses, including univariate and bivariate analyses of doctors’ feedbacks, random forest, prediction performance metrics and plots, were done using R 3.6.3. R package ranger was used to compute the random forest and the variable importances (https://cran.r-project.org/web/packages/ranger). R package tuneRanger was used for fine-tuning the hyper-parameters (https://cran.r-project.org/web/packages/tuneRanger/).

Source of Funding RM was supported by Université Paul Sabatier and Remedy Logic. SK was supported by Vinnitsa National Medical University and Remedy Logic. OS, JK and AR were supported by Remedy Logic.

Results

Analysis of spinal surgeons’ recommendations

An independent panel of five spinal surgeons (fellowship trained spinal surgeons with more than 5 years of experience in practice) was set up. The panel reviewed the 500 medical vignettes to determine the surgical recommendation probability for each vignette (recommendations ranging from 0 to 1). Figure 1A plots the univariate analyses of doctor recommendations. Overall, we observe that doctor recommendation probabilities were spread between 0 and 1, whereas for doctors 3, recommendations were skewed towards high probabilities. Bivariate analyses were then conducted between doctors which found that doctors’ recommendation probabilities were positively but only moderately correlated (Fig. 1B). The average pairwise correlation was 0.4957, the lowest correlation was 0.36 between doctors 1 and 2, and between doctors 1 and 5, while the highest correlation was 0.72 between doctors 3 and 4. Pairwise Cohen’s kappa between doctors also revealed moderate agreements between doctors (Supplementary File 4A). The standard deviations of recommendations were moderate, revealing good consistency of individual doctor recommendations (Supplementary File 4B).

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

Model predictions of surgical recommendation probabilities

An assessment of the accuracy of our hybrid model to predict surgical recommendations was conducted, in comparison to individual doctor recommendations. For this purpose, for each vignette, the ground truth probability for surgical recommendation was calculated as the average between the five independent doctors’ recommendation probabilities. We removed vignettes showing a very high disagreement between doctors (top 10% highest variance). The model
was used to compute the recommendation probability for the same vignettes. The vignettes were randomly split into 70% of vignettes to train the random forest, 10% of vignettes for hybrid model weight estimation and 20% vignettes to estimate prediction accuracy (note that model training was irrelevant for the Bayesian network which was not trained through data). The root mean square error (RMSE) between the model prediction and ground truth probabilities was 0.0964 (Fig. 2A). The Pearson correlation and the R2 were 0.9093 and 0.8268, 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 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). In the hybrid model, the relative weights for the random forest and the Bayesian network were 0.85 and 0.15, putting more weights to machine learning. Random forest slightly overestimated low ground truth probabilities, but globally was performing better than the Bayesian network, explaining the higher weight of the former (Supplementary File 5). Lower performance of the Bayesian network was expected, since it was developed without any training from data.

The average RMSE between individual doctor recommendations and ground truth was 0.1940 (Fig. 2B). The average Pearson correlation and the average R2 were 0.7846 and 0.6155, respectively. When plotting the linear regression $y = ax + b$ with $y = x$, we observed that the doctor 3 was globally overestimating the ground truth probabilities.

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

**Variable importance**

We next assessed which variables were the best predictors of surgical recommendation. For this purpose, we computed variable importance from the random forest model to identify the best predictors of surgical recommendation. Variables related to radiologic findings ranked among the top predictors, including “Imaging showing stenosis”, “Imaging showing disc herniation” and “Imaging showing segmental instability”. Moreover, certain clinical symptoms including “Motor deficit as reported by doctors”, “Back pain” and “Leg weakness as reported from patient” were also very influential (Fig. 3).

**Model predictions of surgical recommendation as binary decision**

Surgical recommendations were also analyzed as a dichotomous classification, by discriminating between two classes: no or weak recommendation class vs. strong recommendation class, with a probability threshold of 0.66.

The AUROC between model and ground truth recommendations was 0.9266 (Fig. 4A), while the sensitivity and specificity were 0.8 and 0.8298, revealing good accuracy metrics. The Cohen’s kappa for interrater agreement was 0.6298. In comparison, the average AUROC based on individual doctor’s recommendations was 0.8412 (Fig. 4B), and the sensitivity and specificity were 0.7850 and 0.7830, respectively. Average Cohen’s kappa was 0.5659, showing similar agreement.
In a dichotomous classification setting, these results reveal that our model performed comparably to individual doctors.

Discussion

Artificial intelligence (AI) is a rapidly expanding field of research and one which has demonstrated the capabilities to improve decision processes across multiple domains. Its applications, which encompass a broad range of human activities, make it possible in particular to improve the quality of care [10]. In spinal surgery, early applications include image classification such as automating the detection of compression fractures of the vertebral body on imaging studies, preoperative risk prediction models, and clinical decision support tools [11, 12].

In this article, we propose a novel artificial intelligence (AI) model to predict surgical recommendations based on
variables reflecting clinical symptoms, MRI findings, and patient demographic factors. The proposed model demonstrated high prediction accuracy, with a prediction error as measured by the root mean square error (RMSE) between model predictions and ground truth of 0.0964, while the average RMSE between individual doctors’ recommendations and ground truth was 0.1940. In a dichotomous classification setting, the prediction error as measured by AUROC was 0.9266, with a Cohen’s kappa value of 0.6298, while the corresponding average metrics based on individual doctor’s recommendations were 0.8412 and 0.5659, respectively. The model thus shows surgical recommendation accuracy metrics that are comparable to recommendations from an independent expert panel.

In a previous application of AI for a preoperative risk prediction model, a machine learning approach based on logistic regression was used to predict complications after spinal surgery depending on patient variables with AUROC ranging from 0.7 to 0.76 [13]. Another regression model obtained similar results [14]. Machine learning models were also proposed to predict pain and functional outcomes after surgery [15–18]. For example, Karhade et al. used various machine learning algorithms including random forests, support vector machines, and logistic regression to predict the minimal clinically important difference after surgery as a successful outcome [19]. Compared to previously published models predicting complication risks, pain reduction or functional outcomes, our model directly computes the probability to recommend spinal surgery for a potential LSS patient, thus representing an effective augmentation tool to the medical decision making process. Another caveat of previous machine learning models is that, if the training data is biased toward certain cohort characteristics such as gender or ethnicity, the model will be biased accordingly [20]. Unlike previous models, the proposed hybrid model relies on both a data-driven machine learning model as well as an expert model. On one hand, the machine learning model (random forest) has the advantage to learn complex nonlinear relationships and cross-effects from the data, which is hard for medical experts to elicit. On the other hand, an expert model (a Bayesian network) explicitly implements surgical recommendations from the peer-reviewed published literature combined with the expert opinions of a multidisciplinary team in the fields of spinal surgery, rehabilitation medicine, interventional and diagnostic radiology. The hybrid model combines the advantages of both sides.

We found that variables from imaging analyses are the most important variables to consider surgical recommendations, which is in line with a recent machine learning model showing that spinal surgery candidacy may be predicted using imaging only [21]. However, we also found as important predictors certain specific clinical symptoms including motor deficit reported by the doctor, back pain and leg weakness as reported from the patient. These variables differ greatly from previous studies aiming to predict surgery complications which found the health insurance provider [13] or the ASA score (physical status score) as the best predictors [14].

One limitation of the present study was the utilization of 500 medical vignettes as opposed to recorded patient data. Future studies will be carried out using patient data from a large cohort for improved validation. Second, building a model using expert knowledge requires extensive efforts for expert elicitation and for obtaining a consensus among doctors. Third, the model was focused on spinal surgery for LSS, and in its current state, is not generalizable to the broader population of patients with lumbar spine issues that may have concomitant diagnoses (e.g. spondylolisthesis) that may require distinct interventions from patients with isolated LSS. Future research will further extend the model to predict surgical recommendations for other spinal conditions, such as lumbar disc herniation, and segmental instability.

In conclusion, the results suggest that AI can be used to bring efficiency and automation to the decision-making process for determining surgical candidacy for LSS, with comparable performance to physicians. Considering physicians and other health care providers must obtain advance approval from a health plan before a specific service, our model could fill in as a significant instrument for fast and efficient decisions at limited cost. Moreover, in the current study, we found that imaging combined with certain clinical variables such as motor deficit and pain are the key predictors of surgery candidacy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00586-022-07307-7.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Wu AM, Zou F, Cao Y, Xia DD, He W, Zhu B, Chen D, Ni WF, Wang XY, Kwan KY (2017) Lumbar spinal stenosis: an update on the epidemiology, diagnosis and treatment. AME Med J 2(5):1–14
2. Grotle M, Smastuen MC, Fjeld O, Gryvle L, Helgeland J, Storheim K, Solberg TK, Zwart J-A (2019) Lumbar spine surgery across 15 years: trends, complications and reoperations in a longitudinal observational study from Norway. BMJ Open 9(8):e028743

3. Xin-long M, Xing-wen Z, Jian-xiong M, Fei L, Yin W, Bin L (2017) Effectiveness of surgery versus conservative treatment for lumbar spinal stenosis: a system review and meta-analysis of randomized controlled trials. Int J Surg 44:329–338

4. Yanamadala V, Leveque J-C, Sethi RK (2017) Misdiagnosis is a prevalent cause of inappropriate lumbar spine surgery. Spine J 17(10):S195

5. Yanamadala V, Kim Y, Buchlak QD, Wright AK, Babington J, Friedman A, Mecklenburg RS, Farrokhi F, Leveque J-C, Sethi RK (2017) Multidisciplinary evaluation leads to the decreased utilization of lumbar spine fusion: an observational cohort pilot study. Spine 42(17):E1016–E1023

6. Vagaska E, Litavcova A, Srotova I, Vlckova E, Kerkovsky M, Jarkovsky J, Bednarik J, Adamova B (2019) Do lumbar magnetic resonance imaging changes predict neuropathic pain in patients with chronic non-specific low back pain? Medicine 98(17):e15377

7. Zeifang F, Schiltenwolf M, Abel R, Moradi B (2008) Gait analysis does not correlate with clinical and MR imaging parameters in patients with symptomatic lumbar spinal stenosis. BMC Musculoskelet Disord 9(1):89

8. Osti OL, Daniel JR (2018) Failed back surgery syndrome: a review article. Asian Spine J 12(2):372–379

9. Deyo RA (2015) Fusion surgery for lumbar degenerative disc disease: still more questions than answers. Spine J 15(2):272–274

10. Briganti G, Le Moine O (2020) Artificial intelligence in medicine: Today and tomorrow. Front Med 7:27

11. Chang M, Canseco JA, Nicholson KJ, Patel N, Vaccaro AR (2020) The role of machine learning in spine surgery: the future is now. Front Surg 7:54

12. Labeilski D, Hersh A, Azad TD, Ehresman J, Pennington Z, Lerner K, Sciuabba DM (2021) Prediction models in degenerative spine surgery: a systematic review. Global Spine J 11(1):79S-88S (PMID: 33890803)

13. Han SS, Azad TD, Suarez PA, Ratliff JK (2019) A machine learning approach for predictive models of adverse events following spine surgery. Spine J 19(11):1772–1781

14. Zehnder P, Heid U, Pigott T, Luca A, Loibl M, Reitmeir R, Fekete T, Haschtmann D, Mannion AF (2021) Development of a model to predict the probability of incurring a complication during spine surgery. Eur Spine J 30(5):1337–1354

15. Gilmore SJ, Jahne AJ, Davidson M, McClelland JA (2019) 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

16. Khor S, Lalavalle D, Cizik AM, Bellabarba C, Chapman JR, Howe CR, Dawei L, Alex Mohit A, Oskouian RJ, Roh JR, Shonnard N, Dagal A, Flum DR (2018) Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. JAMA Surg 153(7):634–642

17. Kim H-J, Park J-H, Kim J-W, Kang K-T, Chang B-S, Lee C-K, Yeom JS (2014) Prediction of postoperative pain intensity after lumbar spinal surgery using pain sensitivity and preoperative back pain severity. Pain Med 15(12):2037–2045

18. McGirt MJ, Sivaganesan A, Asher AL, Devin CJ (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. Neurosurg Focus 39(6):E13

19. Karhade AV, Thio QCBS, Ogink PT, Shah AA, Bono CM, Oh KS, Saylor PJ, Schoenfeld AJ, Shin JH, Harris MB, Schwab JH (2018) Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. Neurosurgery 85(1):E83–E91

20. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, Mavridis N (2020) Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 3(1):81

21. Wilson B, Gaonkar B, Yoo B, Salehi B, Attiah M, Villaroman D, AHN C, Edwards M, Laiwalla A, Ratnaparkhi A, Li I, Cook K, Beckett J, Macyszyn L (2021) Predicting spinal surgery candidacy from imaging data using machine learning. Neurosurgery 89(1):116–121

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.