Original Article

AO SPINE

Determining Prior Authorization Approval for Lumbar Stenosis Surgery With Machine Learning

Global Spine Journal
2024, Vol. 14(6) 1753–1759

DOI: 10.1177/21925682231155844
journals.sagepub.com/home/gsj
Sage

Amaury De Barros[1,2], Frederik Abel[3], Serhii Kolisnyk[4], Gaspere C. Geraci[5], Fred Hill[5], Mary Engrav[5], Sundara Samavedi[5], Olga Suldina[6], Jack Kim[5], Andrej Rusakov[5], Darren R. Lebl[3], and Raphael Mourad[5,7]

Abstract

Study Design: Medical vignettes.

Objectives: Lumbar spinal stenosis (LSS) is a degenerative condition with a high prevalence in the elderly population, that is associated with a significant economic burden and often requires spinal surgery. Prior authorization of surgical candidates is required before patients can be covered by a health plan and must be approved by medical directors (MDs), which is often subjective and clinician specific. In this study, we hypothesized that the prediction accuracy of machine learning (ML) methods regarding surgical candidates is comparable to that of a panel of MDs.

Methods: Based on patient demographic factors, previous therapeutic history, symptoms and physical examinations and imaging findings, we propose an ML which computes the probability of spinal surgical recommendations for LSS. The model implements a random forest model trained from medical vignette data reviewed by MDs. Sets of 400 and 100 medical vignettes reviewed by MDs were used for training and testing.

Results: The predictive accuracy of the machine learning model was with a root mean square error (RMSE) between model predictions and ground truth of .1123, while the average RMSE between individual MD's recommendations and ground truth was .2661. For binary classification, the AUROC and Cohen's kappa were .959 and .801, while the corresponding average metrics based on individual MD's recommendations were .844 and .564, respectively.

Conclusions: Our results suggest that ML can be used to automate prior authorization approval of surgery for LSS with performance comparable to a panel of MDs.

Keywords

Lumbar spinal stenosis, spinal surgery, artificial intelligence, machine learning, surgical decision making

[1] Toulouse NeuroImaging Center (ToNIC), University of Toulouse Paul Sabatier-INSERM, Toulouse, France
[2] Neuroscience (Neurosurgery) Center, Toulouse University Hospital, Toulouse, France
[3] Hospital for Special Surgery, New York, NY, USA
[4] Vinnitsa National Medical University, Vinnytsia, Ukraine
[5] Remedy Logic, New York, NY, USA
[6] Cadabra Studio, Dnipro, Ukraine
[7] University of Toulouse, Toulouse, France

Corresponding Author:
Raphael Mourad, PhD, Université Toulouse III Paul Sabatier, 118 Rte de Narbonne, Toulouse 31062, France.
Email: raphael.mourad@univ-tlse3.fr

Introduction

Lumbar degenerative spine disease (DSD) is increasing in developed countries that is linked to multiple factors such as ageing of the population, sedentary lifestyle, or overweight. Among the spectrum of DSD, lumbar spinal stenosis (LSS) represents a condition that has a high incidence estimated between 1700 and 2200 per 100 000 inhabitants in Europe and North America, which most commonly occurs beyond the 5th life decade.[1] Depending on the natural course of LSS, surgery is often offered to patients after well-conducted conservative treatment. Surgical management can vary greatly from minimally invasive decompression techniques to multilevel lumbar arthrodesis without high levels of evidence that favor a particular option for this condition.[2]

In order to reduce the cost of increasingly sophisticated and complex surgical care in the United States, a request for surgery must undergo a prior authorization process reviewed by a medical director (MD) to be considered eligible for coverage by a health plan or insurance.[3] An MD can act as a "safeguard" ensuring both safety for patients and cost-savings for health systems. However, the recommendations of the MDs may also be subject to debate due to their cost and subjectivity.

Artificial intelligence (AI), in particular machine learning models (ML), are increasingly used for complex decision making in medicine. ML-powered medical solutions have the potential to enable predictive, preventive, personalized, and participatory medicine.[4] ML can infer medical expertise directly from experimental data using various algorithms with either classical machine learning (such as random forest)[5] or more recent deep learning (DL) approaches.[6] In spine imaging, ML has already successfully been used for automated spinal segmentation and diagnostic tasks such as vertebral fracture detection.[7,8] Other ML approaches included surgery decision making to identify surgical candidates based on surgeon's recommendation[9] or predicting postoperative outcomes.[10,11]

Here, we propose a novel ML approach to compute the recommendation probability of spinal surgery for LSS based on MD decision making instead of surgeon's recommendation. The model consists of a random forest model trained to accurately estimate model parameters from medical vignette data reviewed by MDs. We hypothesized that the performance of our proposed ML approach is comparable to that of a panel of spine MDs.

Materials and Methods

Medical Vignettes

A set of 66 variables representing clinical symptoms, physical examinations, 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 Table 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 Table 2). The generated vignettes were designed to provide a range of probabilities for surgical recommendation ranging from low to high probability. We assumed the MRI findings, including stenosis, to be determined by a radiologist.

Since the designed vignettes were not from real patients, informed consent and institutional review board were not required.

Review of Vignettes by an Independent Panel of Medical Directors

The 500 medical vignettes were reviewed by an independent panel of four medical directors (MDs) from different medical practices in order to determine the probability of surgical recommendation for each medical vignette. Each MD 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.

Moreover, the MDs were also asked to answer for each vignette the presence (or absence) of (i) inconsistencies between the reported symptoms and the imaging findings, of (ii) inconsistencies between the reported symptoms and the physical examination, and of (iii) inconsistencies between the imaging findings and the physical examination. MDs were asked to globally assess the presence of inconsistencies between sets of variables (for instance between reported symptom variables and imaging finding variables), but not between two particular variables. A vignette can show inconsistencies between the reported symptoms and the imaging findings with, for instance, stenosis found by MRI but without back pain and leg weakness.

Note that this panel of MDs was independent from the multidisciplinary team in spinal surgery, rehabilitation medicine, interventional and diagnostic radiology used to build the vignettes. The panel was composed of physicians specialized in primary care, emergency medicine and geriatric medicine (no surgeons) who had a long experience as medical directors for health insurance companies.

Machine Learning Model of Inconsistencies

Using the medical vignettes reviewed by MDs, three different random forest models were trained from the set of 66 variables. One model was trained to predict the probability of inconsistencies between the reported symptoms and the imaging findings. One model was trained to predict the probability of inconsistencies between the reported symptoms and the physical examination. One model was trained to predict the probability of inconsistencies between the imaging

findings and the physical examination. Vignettes were randomly split into 80% for fine-tuning and training, and 20% for testing predictions.

Hyper-parameters min.node.size = 22, mtry = 3 and sample.fraction = .68 were obtained by fine-tuning with 5-fold cross-validation. Split rule "gini" was used.

Machine Learning Model of Surgery Recommendation

Similarly, a random forest model was trained to predict the probability of surgical recommendation from the set of 66 variables, and from the 3 possible inconsistencies. Vignettes were randomly split into 80% for fine-tuning and training, and 20% for testing predictions.

Hyper-parameters min.node.size = 34, mtry = 3 and sample.fraction = .50 were obtained by fine-tuning with 5-fold cross-validation. Split rule "variance" was used.

Data Analysis

All data analyses, including univariate and bivariate analyses of MDs' feedbacks, random forest, prediction performance metrics and plots were done using R 4.2.1. 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/). R package fastshap was used to compute SHAP values (https://cran.r-project.org/web/packages/fastshap).

Results

Analysis of Medical Directors' Recommendations

An independent panel of four MDs (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 MD recommendations. Overall, MD recommendation probabilities were spread between 0 and 1, whereas for MD 4, recommendations were skewed towards low probabilities. Bivariate analyses were then conducted and revealed that recommendation probabilities were positive overall, but only showed moderate correlation between MDs (Figure 1B). The average pairwise correlation was .4905, while the lowest correlation was .26 between MDs 2 and 4, and the highest correlation was .82 between MDs 1 and 3.

Overall, the data revealed positive but moderate correlation between MD recommendations and that one MD was biased towards very low recommendation probabilities, reflecting a high level of heterogeneity between individual MD recommendations.

Machine Learning Predictions of Inconsistencies Between Reported Symptoms, Physical Examination and Imaging Findings

When reviewing a request for surgery, medical directors usually not only reject surgery based on reported symptoms, physical examination and imaging findings separately, but also look for inconsistencies between them. For instance, they usually check whether symptoms are supported by imaging findings, or if reported symptoms are consistent with physical examination.

Hence, in order to gain more insight into the prior authorization process, we sought to predict the probability of inconsistencies between reported symptoms, physical examination and imaging findings. For this purpose, for each vignette, we considered the presence of an inconsistency when at least one MD identified an inconsistency. For the three possible inconsistencies, the AUROC were .902 (reported symptoms and imaging findings), .913 (reported symptoms

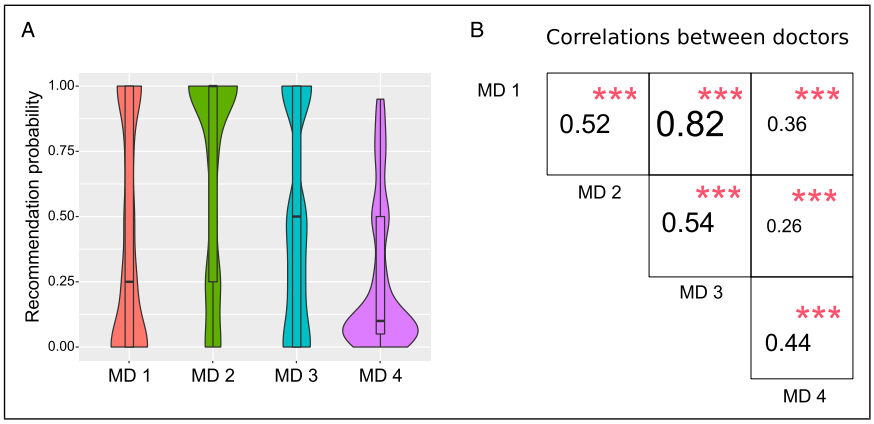

Figure 1. Uni- and bivariate analyses of medical director's (MD's) recommendation. (A) Violin plots of surgery recommendations (univariate analysis). (B) Pearson correlations for each pair of medical directors (MDs), together with statistical significance with red stars (bivariate analysis).

and physical examination) and .899 (imaging findings and physical examination), respectively (Figure 2).

Inconsistencies between reported symptoms, physical examination and imaging findings could thus be predicted with good accuracy, and thus further used to support surgery recommendation.

Model Predictions of Surgical Recommendation Probabilities

The accuracy of our random forest model to predict surgical recommendations was assessed and compared with individual doctor recommendations. For this purpose, for each vignette, the ground truth probability for surgical recommendation was calculated as the average between the four independent MDs' recommendation probabilities. The model was used to compute the recommendation probability for the same vignettes. The vignettes were randomly split into 80% of vignettes to train the random forest, and 20% vignettes to estimate prediction accuracy.

The root mean square error (RMSE) between the model prediction and ground truth probabilities was .1123 (Figure 3A). The Pearson correlation and the R^2 were .9258 and .8571, respectively. When plotting the linear regression y = ax + b (assuming a linear relation between model prediction and ground truth) with y = x (assuming perfect

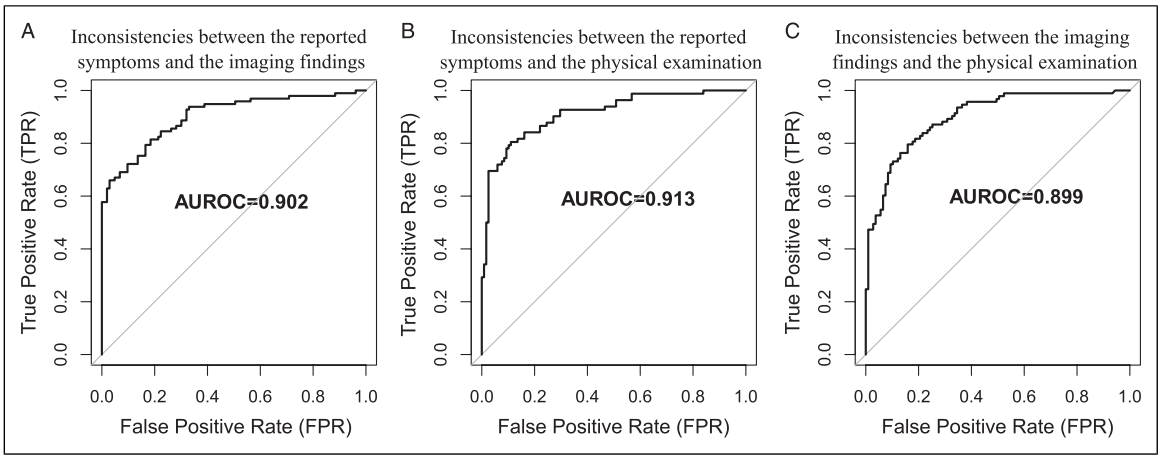

Figure 2. Performance of the prediction of inconsistencies. (A) Receiver operating characteristic curve (ROC) of the prediction of inconsistencies between the reported symptoms and the imaging findings. Area under the ROC (AUROC) is plotted. (B) ROC of the prediction of inconsistency between reported symptoms and the physical examination. (C) ROC of the prediction of inconsistency between the imaging findings and the physical examination.

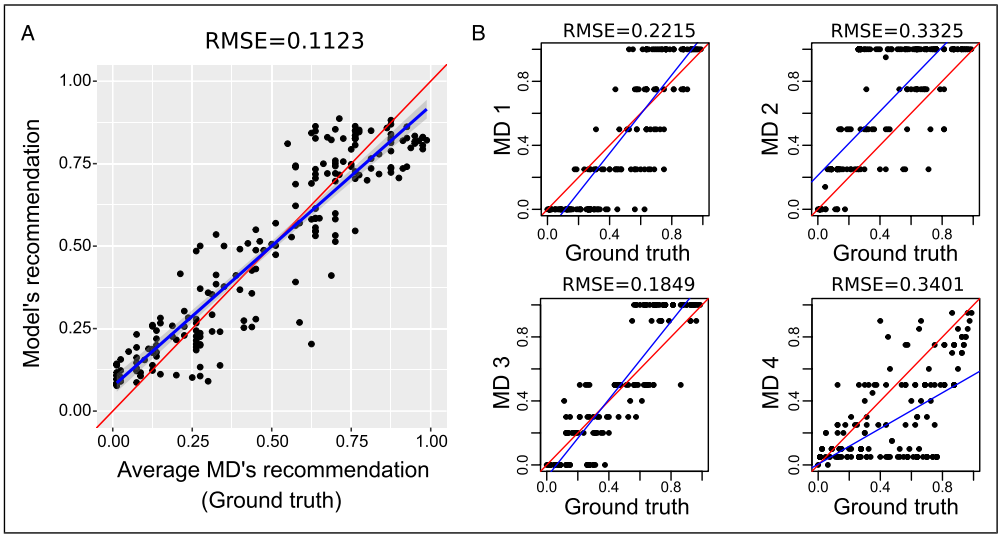

Figure 3. Comparison of prediction performance between the model and individual MDs for surgery recommendation probability. (A) Scatter plot between model's recommendation probability and ground truth recommendation probability. (B) Scatter plots between individual medical director's recommendation probability and ground truth recommendation probability.

agreement between model prediction and ground truth), we globally observed a good fit to the data, and a slight overestimation of low ground truth probabilities (when surgery should not be done), and a slight underestimation of high ground truth probabilities (when surgery should be done). In binary classification (no or weak recommendation class vs strong recommendation class), the prediction error as measured by AUROC was .959, the sensitivity was .914, the specificity was .916, and the Cohen's kappa value was .801 (Supplementary Figure 1A).

The average RMSE between individual doctor recommendations and ground truth was .2661 (Figure 3B). The average Pearson correlation and the average R2 were .7843 and .6151, respectively. When plotting the linear regression $y = ax + b$ with $y = x$, we observed that the doctor 2 was globally overestimating the ground truth probabilities and the doctor 4 underestimated high ground truth probabilities. In binary classification, the average AUROC was .844, the average sensitivity was .780, the average specificity was .820, and the average Cohen's kappa value was .564 (Supplementary Figure 1B).

When predicting surgical recommendation probabilities, our validation performed on vignettes revealed that the ML model has higher accuracy compared to individual MD recommendations.

Model Explainability

When approving or denying prior authorization of surgery for a patient, it is critical for a medical director to justify the decision. Hence, to better understand the contributions of each variable to approve surgery for a patient, SHAP (SHapley Additive exPlanations) values were computed from the random forest model used to predict surgical recommendations. To illustrate the model exploitability, we picked two vignettes (SHAP values are in Supplementary Table 3). The first vignette was predicted to be denied for surgery (prob = .195; Figure 4A). SHAP values revealed that denial was mainly due to the absence of stenosis identified by MRI, the absence of activity limitation and participation restriction, and the lack of physical therapy and epidural steroid injection. The second vignette was predicted to be accepted for surgery (prob = .851; Figure 4B). For this vignette, SHAP values showed that approval was due to a severe stenosis shown by MRI with a severity score of .895 and a cross-sectional area of the dural sac of 45 mm^2, no inconsistency between reported symptoms and imaging findings, activity limitation and participation restriction and segmental instability shown by MRI.

Using SHAP values, the top-10 contributions of variables could be used to explain the decision to approve or deny surgery for a given vignette, providing insight in the decision process.

Discussion

Machine learning (ML) is a rapidly expanding field of research used nowadays in many different industries to improve decision making. In particular, ML is being increasingly used in medicine, in order to improve the quality of care by providing a greater autonomy and a more personalized treatment to the patients.[4]

In this article, we propose a novel ML model to predict surgical recommendations from a consensus of MDs based on variables reflecting clinical symptoms, physical examinations, MRI findings, and patient demographic factors. The ML model showed high prediction accuracy, as measured by a low root mean square error (RMSE) of .1123 between model predictions and ground truth, compared to the average RMSE of .2661 between individual MDs' recommendations and ground truth. In binary classification, the prediction error as measured by AUROC was .959, with a Cohen's kappa value of .801, while the corresponding average metrics based on individual MD's recommendations were .844 and .564, respectively. The model thus shows MD recommendation accuracy metrics comparable or better than those from an independent group of experts.

In a previous article, Mourad et al. found similar results for prediction error (AUROC and Cohen's kappa were .9266 and .6298 respectively) using a hybrid AI model in determining candidates for lumbar surgery in LSS based on surgeon decisions on medical vignettes.[10] Similarly, another study showed similar prediction error (AUROC = .90) with a cohort of 387 patients.[11] Interestingly, Wilson et al. obtained good predictions with ML on MRI imaging only, with AUROC = .88.[12] Our Top 10 variable contributions according to SHAP-values emphasize the preponderance of MRI findings such as the presence of stenosis and the stenosis grade, but also point out the role of some clinical and previous therapeutic history data such as activation limitation and participation restriction, and inconsistencies between reported symptoms and imaging for the decision process. According to the comment of Fourney,[13] predicting surgery only on MRI findings expose to the risk of "excessive referrals to surgery" due to inconsistencies between imaging and clinical repercussion.

Other ML models have been used for the prediction of postoperative outcomes, which can be an indirect aid in surgical decision-making.[14-16] For example, Azimi et al., developed an Artificial Neural Network based on MRI, clinical and demographic data for predicting 2-year surgical satisfaction after LSS surgery with AUROC = .80.[17]

One limitations of our approach could be the use of medical vignettes that do not refer to real patient cases. However, unlike the prediction from a patient cohort, this may limit the biases related to the demographic characteristics of the cohort used for ML.[18] Another limitation of the study is that our ML approach focused specifically on LSS and extrapolation of the results to other types of spinal pathologies is limited. However, considering other spinal pathologies was out of the scope of this study and future studies will aim to develop novel models adapted to other spinal degenerative conditions, such as disc herniation, or application on other spine locations.

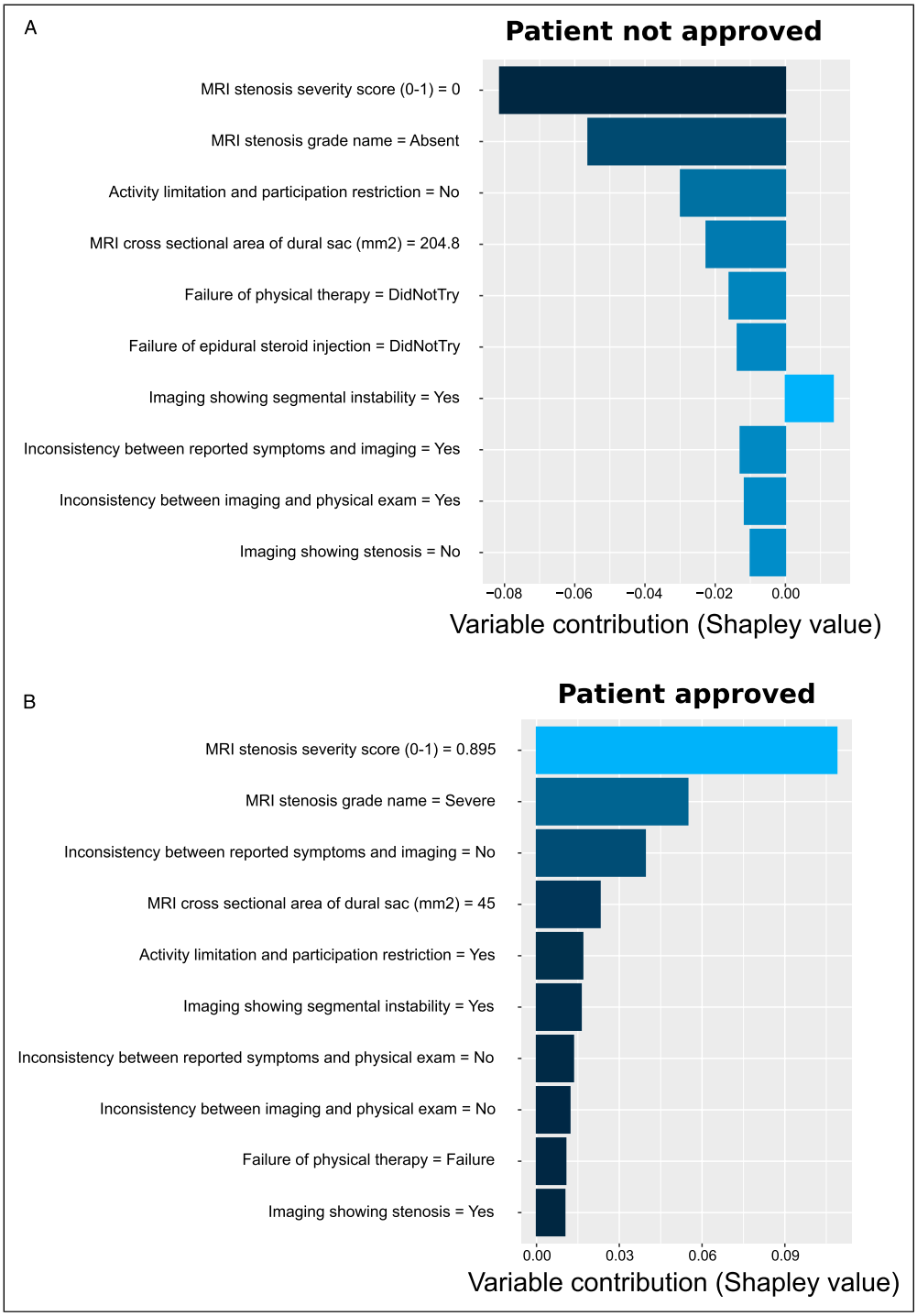

Figure 4. Illustration of model's exploitability with two vignettes. (A) Top-10 variable contributions as measured by SHAP values for a vignette corresponding to a patient whose surgery was denied. (B) Top-10 variable contributions as measured by SHAP values for a vignette corresponding to a patient whose surgery was approved.

Third, algorithms were developed as a surgical decision aid but do not presume the type of surgery performed (eg minimal invasive or open surgery, decompression with or without performed fusion, etc.). Taking the type of surgery into account as variable similar to clinical or imaging data, may further improve the accuracy of surgical decision support algorithms for a given pathology.

In conclusion, the results suggest that ML can be used as a support tool for surgical decision-making in the context of LSS for prior authorization for coverage by a health plan or

insurance similar to MDs. In addition, the use of ML may reduce heterogeneity and subjectivity of decision-making by MDs along with its cost and time-consuming nature.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

AB and RM were 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.

ORCID iDs

Amaury De Barros https://orcid.org/0000-0001-9724-8616
Frederik Abel https://orcid.org/0000-0001-8988-6917
Serhii Kolisnyk https://orcid.org/0000-0001-9424-0037
Raphael Mourad https://orcid.org/0000-0001-6700-5728

Supplemental Material

Supplemental material for this article is available online.

References

1. Ravindra VM, Senglaub SS, Rattani A, et al. Degenerative lumbar spine disease: Estimating global incidence and worldwide volume. *Global Spine J*. 2018;8(8):784-794.
2. Yoo RIJ, Harris IA, Pinheiro MB, Koes BW, van Tulder MW. Surgical options for lumbar spinal stenosis. *Cochrane Database Syst Rev*. 2016;11.
3. Epstein NE. Lumbar stenosis surgery: Spine surgeons not insurance companies should decide when enough is better than too much. *Surg Neurol Int*. 2017;8:247.
4. Briganti G, Le Moine O. Artificial intelligence in medicine: Today and tomorrow. *Front Med*. 2020;7, 27.
5. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
6. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016.
7. Hornung AL, Hornung CM, Mallow G, et al. Artificial intelligence in spine care: Current applications and future utility. *Eur Spine J*. 2022;31(8):2057-2081.
8. Stephens ME, O'Neal CM, Westrup M, et al. Utility of machine learning algorithms in degenerative cervical and lumbar spine disease: A systematic review. *Neurosurg Rev*. 2021;45(2):965-978.
9. Siccoli A, de Wispelaere P, Schröder, Staartjes ML, Staartjes VE. Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurg Focus*. 2019; 46(5):E5.
10. Mourad R, Kolisnyk S, Baiun Y, et al. Performance of hybrid artificial intelligence in determining candidacy for lumbar stenosis surgery. *Eur Spine J*. 2022;31(8):2149-2155.
11. Xie N, Wilson PJ, Reddy R. Use of machine learning to model surgical decision-making in lumbar spine surgery. *Eur Spine J*. 2022;31(8):2000-2006.
12. Wilson B, Gaonkar B, Yoo B, et al. Predicting spinal surgery candidacy from imaging data using machine learning. *Neurosurgery*. 2021;89(1):116-121.
13. Fourney R. Commentary: Predicting spinal surgery candidacy from imaging data using machine learning. *Neurosurgery*. 2021; 89(1):E16.
14. Kalagara S, Eltorai E, Durand M, DePasse WM, Daniels J, Daniels AH. Machine learning modeling for predicting hospital readmission following lumbar laminectomy. *J Neurosurg Spine*. 2019;30(3):344-352.
15. Karhade V, Thio CBS, Ogink PT, et al. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. *Neurosurgery*. 2018;85(1):E83.
16. Ogink PT, Karhade AV, Thio Q, et al. Development of a machine learning algorithm predicting discharge placement after surgery for spondylolisthesis. *Eur Spine J*. 2019;28(8): 1775-1782.
17. Azimi P, Benzel EC, Shahzadi S, et al. Use of artificial neural networks to predict surgical satisfaction in patients with lumbar spinal canal stenosis. *J Neurosurg Spine*. 2014;20(3):300-305.
18. Cirillo D, Catuara-Solarz S, Morey C, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *Npj Digital Medicine*. 2020;3(1):81.