MICN Lab

Version 2.0 – 5.7.2021

Development and External Validation of an International, Multicenter Machine Learning Algorithm for Prediction of Outcome After Lumbar Spinal Fusion for Degenerative Disease: The FUSE-ML Study

Victor E. Staartjes[1,2,3], BMed; Carlo Serra[1], MD; Marc L. Schröder[3], MD, PhD;

ClinicalTrials.gov:

ID: FUSE-ML ***Predicting Outcomes After Lumbar Fusion for Degenerative Disease*** ***[NCT ID not yet assigned]***

1: Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, Clinical Neuroscience Center, University Hospital Zurich, Switzerland
2: Amsterdam UMC, Vrije Universiteit Amsterdam, Neurosurgery, Amsterdam Movement Sciences, Amsterdam, The Netherlands
3: Department of Neurosurgery, Bergman Clinics Amsterdam, Amsterdam, The Netherlands

Introduction

Low back pain is one of the top-three causes of disability in Western societies and imposes significant direct and indirect socio-economic costs.[1,2] The etiology of low back pain with or without radiating leg pain is multifactorial, but it is often related to degenerative disc disease (DDD) or to spondylolisthesis.[3–6] The standard treatment for symptomatic spondylolisthesis or progressive DDD in patients who are unresponsive to long-term conservative treatment is interbody fusion, but this is controversial.[7] With some reports showing no benefit compared to conservative treatment, patient selection is vitally important.[6] Various prognostic tests attempt to identify subsets of patients that might benefit most from surgery, but the validity of these tests is unclear.[5,8,9] Ultimately, success in this category of patients should be defined by improved physical symptoms (patient-reported outcome measures [PROMs]) rather than technical success of the procedure. A relevant proportion of patients with intractable, conservative therapy-resistant lumbar degenerative disease do finally profit from lumbar fusion surgery[5] – the difficult question is how to identify them securely and avoid unnecessary, unsuccessful surgery.

In the literature, several subsets of patients with lumbar degenerative disease who may profit more than others from lumbar spinal fusion have been identified.[5,10] Accurate preoperative identification of patients at high risk for unsatisfactory outcome and vice-versa would be clinically advantageous, as it would allow enhanced resource preparation, better surgical decision-making, enhanced patient education and informed consent, and potentially even modification of certain risk factors for unsatisfactory outcome.[11] However, it is often impossible for clinicians to balance the many described single risk factors for each adverse event to arrive at a personalized risk-benefit profile in individual patients.

Machine learning (ML) methods have been extraordinarily effective at integrating many clinical patient variables into one holistic risk prediction tailored to each patient. One multicenter model based on classic statistics has already been described by Khor et al.[12] – However, upon external validation, it proved to be unreliable and rather poorly calibrated.[10] Also, this model was based on a relatively small number of patients for ML. The aim of the FUSE-ML study is to develop and externally validate a robust ML-based prediction tool based on multicenter data from a range of international centers that will provide individualized risk-benefit profiles tailored to each patient undergoing lumbar spinal fusion for degenerative disease.

Methods

Overview

Data will be collected by a range of international centers. Overall, the models will be built and publication will be compiled according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines.[13] One model will be created for each of the relevant outcomes detailed below. University of Zurich (V.E. Staartjes, C. Serra) is the sponsor of this study.

Ethical Considerations

Each center will be responsible for their own ethics board / institutional review board (IRB) approval and for establishing a data transfer agreement (DTA). The sponsor (University of Zurich) will present a standard DTA upon request. They must gain approval for retrospective or prospective data collection and sharing of the completely deidentified data with the sponsor. The sponsor can aid by providing this detailed study protocol. All study procedures will be carried out according to the Declaration of Helsinki and its amendments.

Inclusion and Exclusion Criteria

Patients with the following indications for thoracolumbar pedicle screw placement are considered for inclusion: Degenerative pathologies (one or multiple of the following: spinal stenosis, spondylolisthesis, degenerative disc disease, recurrent disc herniation, failed back surgery syndrome (FBSS), radiculopathy, pseudarthrosis). Patients undergoing surgery for – as the primary indication – infections, vertebral tumors, as well as traumatic and osteoporotic fractures or deformity surgery for scoliosis or kyphosis are not eligible. Patients with moderate or severe scoliosis (Coronal Cobb's >30 degrees / Schwab classification sagittal modifier + or ++) are not eligible. Patients undergoing surgery at more than 6 vertebral levels are also not eligible. Patients with missing endpoint data at 12 months will be excluded. Patients are required to give informed consent. Only patients aged 18 or older are considered for inclusion.

Data Collection

Each center will collect their data either retrospectively, from a prospective registry, or from a prospective registry supplemented by retrospectively collected variables. Each center has to contribute a minimum of 100 patients with complete 12-month follow-up data to be included in the study. A standardized Excel database will be provided by the sponsor for anonymous data entry. The data will be entered in standardized and deidentified form. This Excel database will only contain a study-specific patient number. Each center will keep an internal spreadsheet in which the study-specific patient numbers can be traced back to center-specific patient-numbers, should this be necessary. The deadline for submission of the complete data to the sponsor institution is **13th of August 2021.**

Authorship

Centers will have to contribute at least 100 cases with complete outcome data in total to be included in the study. Each participating center will be able to designate a maximum of four authors to be included in the author list. Any other center-specific contributors will be listed as full members of the FUSE-ML study group and will be granted full PubMed / Medline contributor status. The sponsor institution will have six primary author positions available.

Primary Endpoint Definitions

Several endpoints will be assessed:

- **1. Oswestry Disability Index (ODI) at 12 months.**
- **2. Visual Analogue Scale (VAS-BP, 0 to 100) for back pain at 12 months.** *This can also be a converted numeric rating scale (NRS) from 0-10, or a VAS from 0 to 10 converted to 0 to 100.*
- **3. Visual Analogue Scale (VAS-LP, 0 to 100) for leg pain at 12 months.** *This can also be a converted numeric rating scale (NRS) from 0-10, or a VAS from 0 to 10 converted to 0 to 100.*

These outcomes will be dichotomized using the minimum clinically important difference (MCID) according to Ostelo et al.[14] Thus, a 30% or greater improvement in a specific score compared to baseline will be considered as achievement of MCID (clinical success) in that specific score. If patients presented with zero symptoms initially (in either ODI, NRS-BP or NRS-LP), and remained at zero for that score, this will also be defined as MCID for that score.

Features and Their Definitions

All features are measured or estimated preoperatively. In addition to the endpoints, the following input features will be collected:

- Age (years)
- Gender (m/f)
- Presence of the following indications for surgery (choose all that apply):
  - Spondylolisthesis
  - (Recurrent) disc herniation
  - Radiculopathy
  - Chronic low back pain (CLBP) / Degenerative disc disease (DDD)
  - Failed back surgery syndrome (FBSS)
  - Lumbar spinal stenosis
  - Pseudarthrosis
- Index Level(s) (choose all that apply, T12 – S1)
- Height (cm)

- Weight (kg)
- BMI (kg/m^2)
- Smoking status (active / ceased / never)
- Preoperative (baseline) ODI
- Preoperative (baseline) VAS-BP
- Preoperative (baseline) VAS-LP
- American Society of Anesthesiologists (ASA) Score (1-2 / 3 or higher)
- Preoperative use of opioid pain medication (yes / no)
- Asthma pulmonale as a comorbidity (yes / no)
- Prior thoracolumbar spine surgery (yes / no)
- Race/Ethnicity (Caucasian / Black / Asian / Other)
- Surgical approach (choose all that apply: TLIF / PLIF / ALIF / Lateral)
- Pedicle screw insertion (yes / no)
- Minimally invasive technique (yes / no)

Sample Size

While even the largest cohort with millions of patients is not guaranteed to result in a robust clinical prediction model if no relevant input variables are included ("garbage in, garbage out" – do not expect to predict the future from age, gender, and body mass index), the relationship among predictive performance and sample size is certainly directly proportional, especially for some data-hungry ML algorithms. To ensure generalizability of the clinical prediction model, the sample size should be both representative enough of the patient population, and should take the complexity of the algorithm into account. For instance, a deep neural network – as an example of a highly complex model – will often require thousands of patients to converge, while a logistic regression model may achieve stable results with only a few hundreds of patients. In addition, the number of input variables plays a role. Roughly, it can be said that a bare minimum of 10 positive cases are required per included input variable to model the relationships. Often, erratic behavior of the models and high variance in performance among splits is observed when sample sizes are smaller than calculated with this rule of thumb. Of central importance is also the proportion of patients who experience the outcome. For very rare events, a much larger total sample size is consequentially needed. For instance, a prediction based on 10 input features for an outcome occurring in only 10% of cases would require at least 1000 patients including at least 100 who experienced the outcome, according to the above rule of thumb. In general and from personal experience, we do not recommend developing ML models on cohorts with less than 100 positive cases and reasonably more cases in total, regardless of the rarity of the outcome. Also, one might consider the available literature on risk factors for the outcome of interest: If epidemiological studies find only weak associations with the outcome, it is likely that one will require more patients to arrive at a model with good predictive performance, as opposed to an outcome which has several highly associated risk factors, which may be easier to predict. Larger sample sizes also allow for more generous evaluation through a larger amount of patient data dedicated to training or validation, and usually results in better calibration measures.

Between 20% and 40% of patients report no clinically relevant improvement after spinal fusion (minority class).[5,10,12,15] For sample size calculation, we take 20% for a conservative estimate. Consequently, for this study, based on our expertise and on the rules of thumb mentioned above, we estimate that a minimum of 200 patients with a negative outcome (minority class) are required to extract generalizable feature relationships. With an estimated incidence of approximately 20% as explained above, that means that a minimum of around **1000 patients are required for training**. **For adequate evaluation of calibration at external validation, we estimate that another 300 patients will be required** (thus, approximately 60 patients with a positive outcome). Thus, in total, we estimate that **a minimum of 1300 patients** are necessary to arrive at a robust model. More data will likely lead to greater performance and better calibration.

Predictive Modeling

A KNN imputer will be co-trained to impute any missing data that may occur in future application of the model. If there is missing data in the training set, it will be imputed using said KNN imputer.[16] Features or patients with a missingness greater than 25% will be excluded. Data will be standardized and one-hot-encoded. In case of major class imbalance – which is expected for the abovementioned endpoint – random upsampling or synthetic minority oversampling (SMOTE) will be applied to the training set.[17,18] All features will initially be provided to the model for training. If necessary, we will apply recursive feature elimination (RFE) to select input features on the training data.[19]

We will trial the following algorithms for binary classification: Generalized linear model (GLM), generalized additive model (GAM), stochastic gradient boosting machine (GBM), naïve Bayes classifier, simple artificial neural network, support vector machine (SVM), and random forest. Each model will be fully trained and hyperparameter tuned where applicable. The final model will be selected based upon AUC, sensitivity, and specificity, as well as calibration metrics on the resampled training performance. Training will occur in repeated 5-fold cross-validation with 10 repeats.

The one final model will then be assessed on the external validation data only once. 95% confidence intervals for external validation metrics will be derived using the bootstrap.

The threshold for binary classification will either be identified on the training data alone using the AUC-based "closest-to-(0,1)-criterion" or Youden's index to optimize both sensitivity and specificity, or will be optimized on the training set based on clinical significance (rule-out model). All analyses will be carried out in R Version 4.0.2 or more recent.[20]

Evaluation

The performance of classification models can roughly be judged along two dimensions: Model discrimination and calibration.[46] The term *discrimination* denotes the ability of a prediction model to correctly classify whether a certain patient is going to or is not going to experience a certain outcome. Thus, discrimination described the accuracy of a binary prediction – yes or no. *Calibration*, however, describes the degree to which a model's predicted probabilities (ranging from 0% to 100%) correspond to the actually observed incidence of the binary endpoint (true posterior). Many publications do not report calibration metrics, although these are of central importance, as a well-calibrated predicted probability (e.g. your predicted probability of experiencing a complication is 18%) is often much more valuable to clinicians – and patients! – than a binary prediction (e.g. you are likely not going to experience a complication).[21]

Resampled training performance as well as performance on the external validation set will be assessed for discrimination and calibration.[21,22] In terms of discrimination, we will evaluate AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 Score. In terms of calibration, we will assess the Brier score, expected-observed (E/O)-ratio, calibration slope and intercept, the Hosmer-Lemeshow goodness-of-fit test, as well as visual inspection of calibration plots for both datasets, which will also be included in the publication.

Interpretability

The degree and choice of methods for interpretability will depend on the finally chosen algorithm. Some algorithms can natively provide explanations as to which factors influence the outcome in what way. Thus, in case e.g. a GLM, GAM, or naïve Bayes classifier is chosen, the parameters / partial dependence values will be provided. For simple decision trees, diagrams of the decision-making process can be provided. Other models with higher degrees of complexity, such as neural networks or stochastic gradient boosting machines cannot natively provide such explanations. In that case, we will provide both AUC-based variable importance as well as model-agnostic local interpretations of variable importance using the LIME principle.[23,24]

Expected Results

We expect to arrive at a generalizable model based on multicenter international data that is likely to predict consistently with an AUC of at least 0.70 and that is well-calibrated. A web-based prediction tool will also be created for each of the two models using the *shiny*[25] environment, much akin to e.g. https://neurosurgery.shinyapps.io/impairment (Also see for example: Staartjes et al., Journal of Neurosurgery, 2020 [26]). This web-based app will be available for free on any internet-capable device (mobile or desktop), and should be stable on most devices due to the server-based computing. The costs for maintaining the server will be carried by the sponsor. The collected data will be stored by the sponsor for 10 years. The large dataset will be open to further analysis and will be provided to any of the contributing centers at reasonable request and after approval by all other centers. The goal is to enable other analyses using the collected dataset. If any additional analyses lead to publication, all contributors will be included as co-authors and all co-authors will have the opportunity to review said manuscript beforehand. Any contributing study center has the right to veto publication of any subsequent analyses containing their own data.

References

1. Lambeek LC, van Tulder MW, Swinkels ICS, Koppes LLJ, Anema JR, van Mechelen W. The trend in total cost of back pain in The Netherlands in the period 2002 to 2007. *Spine*. 2011;36(13):1050-1058. doi:10.1097/BRS.0b013e3181e70488

2. GBD 2015 DALYs and HALE Collaborators. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Lond Engl*. 2016;388(10053):1603-1658. doi:10.1016/S0140-6736(16)31460-X

3. Ghogawala Z, Dziura J, Butler WE, et al. Laminectomy plus Fusion versus Laminectomy Alone for Lumbar Spondylolisthesis. *N Engl J Med*. 2016;374(15):1424-1434. doi:10.1056/NEJMoa1508788

4. Försth P, Ólafsson G, Carlsson T, et al. A Randomized, Controlled Trial of Fusion Surgery for Lumbar Spinal Stenosis. *N Engl J Med*. 2016;374(15):1413-1423. doi:10.1056/NEJMoa1513721

5. Staartjes VE, Vergroesen P-PA, Zeilstra DJ, Schröder ML. Identifying subsets of patients with single-level degenerative disc disease for lumbar fusion: the value of prognostic tests in surgical decision making. *Spine J*. 2018;18(4):558-566. doi:10.1016/j.spinee.2017.08.242

6. Mannion AF, Brox J-I, Fairbank JC. Consensus at last! Long-term results of all randomized controlled trials show that fusion is no better than non-operative care in improving pain and disability in chronic low back pain. *Spine J Off J North Am Spine Soc*. 2016;16(5):588-590. doi:10.1016/j.spinee.2015.12.001

7. Manchikanti L, Abdi S, Atluri S, et al. An update of comprehensive evidence-based guidelines for interventional techniques in chronic spinal pain. Part II: guidance and recommendations. *Pain Physician*. 2013;16(2 Suppl):S49-283.

8. Willems PC, Elmans L, Anderson PG, Jacobs WCH, van der Schaaf DB, de Kleuver M. The value of a pantaloon cast test in surgical decision making for chronic low back pain patients: a systematic review of the literature supplemented with a prospective cohort study. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc*. 2006;15(10):1487-1494. doi:10.1007/s00586-006-0121-0

9. Carragee EJ, Don AS, Hurwitz EL, et al. 2009 ISSLS Prize Winner: Does discography cause accelerated progression of degeneration changes in the lumbar disc: a ten-year matched cohort study. *Spine*. 2009;34(21):2338-2345. doi:10.1097/BRS.0b013e3181ab5432

10. Quddusi A, Eversdijk HAJ, Klukowska AM, et al. External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc*. Published online October 22, 2019. doi:10.1007/s00586-019-06189-6

11. Steinmetz MP, Mroz T. Value of Adding Predictive Clinical Decision Tools to Spine Surgery. *JAMA Surg*. Published online March 7, 2018. doi:10.1001/jamasurg.2018.0078

12. Khor S, Lavallee D, Cizik AM, et al. Development and Validation of a Prediction Model for Pain and Functional Outcomes After Lumbar Spine Surgery. *JAMA Surg*. 2018;153(7):634-642. doi:10.1001/jamasurg.2018.0072

13. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.

14. Ostelo RWJG, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine*. 2008;33(1):90-94. doi:10.1097/BRS.0b013e31815e3a10

15. Mannion AF, Brox JI, Fairbank JCT. Comparison of spinal fusion and nonoperative treatment in patients with chronic low back pain: long-term follow-up of three randomized controlled trials. *Spine J Off J North Am Spine Soc*. 2013;13(11):1438-1448. doi:10.1016/j.spinee.2013.06.101

16. Templ M, Kowarik A, Alfons A, Prantner B. *VIM: Visualization and Imputation of Missing Values*.; 2019. Accessed January 5, 2020. https://CRAN.R-project.org/package=VIM

17. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953

18. Staartjes VE, Schröder ML. Letter to the Editor. Class imbalance in machine learning for neurosurgical outcome prediction: are our models valid? *J Neurosurg Spine*. 2018;29(5):611-612. doi:10.3171/2018.5.SPINE18543

19. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom Intell Lab Syst*. 2006;83(2):83-90. doi:10.1016/j.chemolab.2006.01.007

20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020. https://www.R-project.org/

21. Staartjes VE, Kernbach JM. Letter to the Editor. Importance of calibration assessment in machine learning-based predictive analytics. *J Neurosurg Spine*. Published online February 21, 2020:1-2. doi:10.3171/2019.12.SPINE191503

22. Collins GS, Ogundimu EO, Le Manach Y. Assessing calibration in an external validation study. *Spine J*. 2015;15(11):2446-2447. doi:10.1016/j.spinee.2015.06.043

23. Kuhn M. Building Predictive Models in *R* Using the **caret** Package. *J Stat Softw*. 2008;28(5). doi:10.18637/jss.v028.i05

24. Tulio Ribeiro M, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *ArXiv E-Prints*. 2016;1602:arXiv:1602.04938.

25. Chang W, Cheng J, Allaire JJ, et al. *Shiny: Web Application Framework for R*.; 2020. Accessed May 28, 2020. https://CRAN.R-project.org/package=shiny

26. Staartjes VE, Broggi M, Zattra CM, et al. Development and external validation of a clinical prediction model for functional impairment after intracranial tumor surgery. *J Neurosurg*. 2020;1(aop):1-8. doi:10.3171/2020.4.JNS20643