








Article

The Influence of Baseline Clinical Status and Surgical Strategy on Early Good to Excellent Result in Spinal Lumbar Arthrodesis: A Machine Learning Approach

Pedro Berjano ¹, Francesco Langella ^{1,*} , Luca Ventriglia ², Domenico Compagnone ¹, Paolo Barletta ¹, David Huber ², Francesca Mangili ², Ginevra Licandro ², Fabio Galbusera ¹ , Andrea Cina ¹, Tito Bassani ¹, Claudio Lamartina ¹, Laura Scaramuzza ¹ , Roberto Bassani ¹, Marco Brayda-Bruno ¹ , Jorge Hugo Villafañe ³ , Lorenzo Monti ⁴  and Laura Azzimonti ² 

- ¹ IRCCS Istituto Ortopedico Galeazzi, 20161 Milan, Italy; pberjano@gmail.com (P.B.); compagnone.nico@gmail.com (D.C.); paolo.barletta@grupposandonato.it (P.B.); fabio.galbusera@grupposandonato.it (F.G.); andrea.cina@grupposandonato.it (A.C.); tito.bassani@grupposandonato.it (T.B.); c.lamartina@chirurgiavertebrale.net (C.L.); scaramuzzolaura@gmail.com (L.S.); r.bassani.spine@gmail.com (R.B.); marco.brayda@spinecaregroup.it (M.B.-B.)
- ² Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), USI-SUPSI, 6900 Lugano, Switzerland; luca.ventriglia@idsia.ch (L.V.); david.huber@idsia.ch (D.H.); francesca.mangili@idsia.ch (F.M.); ginevra.licandro@idsia.ch (G.L.); laura.azzimonti@idsia.ch (L.A.)
- ³ IRCCS Fondazione Don Carlo Gnocchi, 20161 Milan, Italy; mail@villafane.it
- ⁴ Orthopedics and Traumatology Unit, Istituto Clinico Villa Aprica, 22100 Como, Italy; lorenzomonti@hotmail.it
- * Correspondence: francesco.langella.md@gmail.com



Citation: Berjano, P.; Langella, F.; Ventriglia, L.; Compagnone, D.; Barletta, P.; Huber, D.; Mangili, F.; Licandro, G.; Galbusera, F.; Cina, A.; et al. The Influence of Baseline Clinical Status and Surgical Strategy on Early Good to Excellent Result in Spinal Lumbar Arthrodesis: A Machine Learning Approach. *J. Pers. Med.* **2021**, *11*, 1377. <https://doi.org/10.3390/jpm11121377>

Academic Editor: Jan Philipp Radtke

Received: 28 November 2021

Accepted: 13 December 2021

Published: 16 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: The study aims to create a preoperative model from baseline demographic and health-related quality of life scores (HRQOL) to predict a good to excellent early clinical outcome using a machine learning (ML) approach. A single spine surgery center retrospective review of prospectively collected data from January 2016 to December 2020 from the institutional registry (SpineREG) was performed. The inclusion criteria were age ≥ 18 years, both sexes, lumbar arthrodesis procedure, a complete follow up assessment (Oswestry Disability Index—ODI, SF-36 and COMI back) and the capability to read and understand the Italian language. A delta of improvement of the ODI higher than 12.7/100 was considered a “good early outcome”. A combined target model of ODI ($\Delta \geq 12.7/100$), SF-36 PCS ($\Delta \geq 6/100$) and COMI back ($\Delta \geq 2.2/10$) was considered an “excellent early outcome”. The performance of the ML models was evaluated in terms of sensitivity, i.e., True Positive Rate (TPR), specificity, i.e., True Negative Rate (TNR), accuracy and area under the receiver operating characteristic curve (AUC ROC). A total of 1243 patients were included in this study. The model for predicting ODI at 6 months' follow up showed a good balance between sensitivity (74.3%) and specificity (79.4%), while providing a good accuracy (75.8%) with ROC AUC = 0.842. The combined target model showed a sensitivity of 74.2% and specificity of 71.8%, with an accuracy of 72.8%, and an ROC AUC = 0.808. The results of our study suggest that a machine learning approach showed high performance in predicting early good to excellent clinical results.

Keywords: artificial intelligence; lumbar fusion; degenerative disc disease; adult spine deformity; scoliosis; spine registry; personalized medicine

1. Introduction

Degenerative spine disorders represent a complex condition that mainly affects the elderly population, with an incidence in healthy people aged over 70 years of up to 68% [1].

Pain and disability represent its main features, leading to a significant clinical and socio-economic impact with an increasing role in daily medical practice. Its dissemination

goes hand in hand with the aging of the population of developed countries. The spinal disorders have a broad spectrum of clinical manifestations: from minimal or asymptomatic to an invalidating condition. The presentation pattern can variably affect segmental, regional, and global alignment. The pain and disability represent the main feature in a way that is comparable with other self-reported chronic conditions in the general population such as congestive heart failure, arthritis, chronic lung disease or diabetes [2].

The therapeutic approach of spinal disorders is challenging in terms of decision making for several causes and symptoms. Furthermore, the decisional process is made even more complicated by aging patients eligible for surgery and different clinical conditions and comorbidities [1].

In the last decades, the rate of spine surgery increased by up to 40%, and several randomized trials demonstrated the positive and significant effects of these procedures [3,4]. Its safety and effectiveness vary widely among patients. In the worse scenarios, the complications rate can be up to 13% [5]. Indeed, there is always room for improvements in terms of the clinical, surgical, and economic points of view [6]. The scientific research in this field targeted to evaluate the improvement in quality of life (QoL) after surgical treatment for spine surgery in relation to patient age, comorbidity and baseline status. With the aim to improve the cost-effect ratio's performance, the rise of predictive models (PM) is continuously increasing. In 2015, McGirt et al. [7] presented a PM for the clinical practice to help patients, providers, and hospital systems. It is based on demographics, patients' reported outcomes, and clinical data. In particular, the baseline patient-specific factors, such as symptom duration, smoking status, preoperative comorbidities and mental and physical conditions, seem to significantly influence outcomes following lumbar surgery. Sinikallio et al. [8], in a prospective analysis, demonstrated that the patients with preoperative depression and those who had continuous depression postoperatively experienced poor post-operative surgical outcomes and can benefit from targeted cognitive behavioral therapy [9]. Patient-specific factors beyond medical comorbidities, surgical indications, and surgical approaches can play a significant role in influencing overall patient outcomes [10]. The impact of lumbar spine surgery on patients' life is commonly evaluated with three patient-reported outcome measures (PROMs): The Oswestry Disability Index (ODI), the Physical Component Score of the Short Form of the Medical Outcomes Study (SF-36 PCS), and pain scales (VAS leg and back). The minimum clinically important difference (MCID) is commonly considered the threshold to measure the effect of the surgery for the single questionnaire. The use of PROMs and their prediction through machine learning approaches represent a milestone in the development of shared, informed, and individualized decision making potentially capable to support the surgeon to choose the right intervention, at the right time, for the right patient [7].

Our study aimed to develop a preoperative machine learning (ML) model to predict a good to excellent early clinical outcome by using baseline demographic and health-related quality of life scores (HRQOL).

2. Materials and Methods

2.1. Clinical and Demographic Data

The study was conducted in a single spine surgery center and it is based on a retrospective review of prospectively collected data from the institutional registry—SpineReg [11]. The inclusion criteria were age ≥ 18 years, both genders, lumbar arthrodesis procedure identified using the ICD-9 code (8106, 8107 or 8108), a follow up assessment (ODI, SF-36 and core outcome measures index—COMI back) and the capability to read and understand the Italian language. A full set of peri-operative and post-operative data along with clinical outcomes from January 2016 to December 2020 were evaluated. Exclusion criteria were a weak degree of baseline disability or pain (ODI $< 20/100$ and COMI back $< 3/10$), number of levels fused not specified and subject not stratified according to Glassman classification.

The study protocol was conducted in accordance with the Helsinki Declaration of 1957 as revised in 2000. The procedures followed the ethical standards of the responsible

committee on human experimentation and was approved by the ethics committee of our institution (second amendment to the SPINEREG protocol 14 issued on 13 April 2016). The project was supported with funds from the Italian Ministry of Health (project code CO-2016-02364645). All patients gave their written informed consent for the participation in the registry. Baseline demographics, BMI, gender, comorbidities collected through the comorbidity Charlson index (CCI) [12], diagnosis according to the Glassman classification [13], number of spinal levels of intervention, spinal level indexed surgery, clinical scores resulting from medical surveys, complications and revision surgeries were collected.

Table 1 shows that the rates of missing data ranged from 0.0% for the baseline scores of the PROMs and for the patient's personal information, to 20.4% for Levels variables. Independent variables with at least one missing value were imputed using predictive mean matching for numerical variables, and binary/multinomial logistic regression or an ordered logit model for categorical variables. Outcome variables were not imputed to avoid the introduction of bias into the results. Only patients with the observed outcome variables were included in the analysis.

Table 1. Table with the features included in the analysis and associated percentage of missing values.

Glassman	Equipe	Age	Gender	BMI	ASA	ODIPre
18.8%	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%
COMIPre	SFPPre	SFMPre	Levels	From Level	To Level	CCI
0.0%	0.0%	0.0%	20.3%	20.4%	20.4%	19.0%

Glassman: Glassman classification data; Equipe: surgical team; BMI: body mass index; ODIPre: pre-operative Oswestry disability index; COMIPre: pre-operative core outcome measures index; SFPPre: pre-operative physical component score of the short form-36; SFMPre: pre-operative mental component score of the short form-36; CCI: Charlson comorbidity index.

2.2. Clinical Outcomes

The primary outcome was the early (six months post op) significant clinical improvement. In particular, value of improvement higher than 12.7 for ODI [14], 6 for SF36-PCS and 2.2 for COMI Back were considered as indicators of significant clinical improvement [15,16].

To classify surgical operations results with a "good early outcome", we defined a delta of improvement of the ODI higher than 12.7/100. On the other hand, to identify surgical operations with an "excellent early outcome" we used a combined target that identifies an excellent clinical result when all three underlying targets, ODI ($\Delta \geq 12.7/100$), SF-36 PCS ($\Delta \geq 6/100$) and COMI back ($\Delta \geq 2.2/10$), showed a relevant improvement; good or excellent early outcomes will be defined as "Outcome +" and predicted good or excellent early outcomes will be defined as "Prediction +". Patients with a Δ below the threshold value 12.7/100 for ODI will be considered as negative results for the good early outcome, while patients with at least one Δ value below the following thresholds, 12.7/100 for ODI, 6/100 for SF-36 PCS or 2.2/10 for COMI back, are considered negative results for the excellent early outcome; we will refer to these cases in the following as "Outcome −" or "Prediction −". Therefore, in the analysis also patients without the minimal clinically relevant improvement as well as patients with clinical worsening were included.

An exploratory analysis was performed. Patients were classified as having or not the binomial risk factor—(Risk +) or (Risk −), respectively. Different scenarios were simulated to verify the performance of each method of calculation in three age categories. For each scenario, a 2×2 table was built (Good Outcome +/− vs. Risk +/−). A Chi-squared test was used for statistical association between reaching the good outcomes and the presence of risk factors. The odds-ratios with their 95% confidence intervals, and point estimations of the sensitivity and specificity of the alignment rules to discriminate patients with final good or poor clinical outcome, the positive (PPV) and negative (NPV) predictive values and positive and negative likelihood ratios (LR +, LR −, respectively) were calculated. Differences between preoperative and postoperative clinical outcomes were tested with the two-tailed Student's t test for paired samples. The Mann–Whitney U-test was used in the

cases of abnormally distributed variables. Normality was verified with the Kolmogorov–Smirnov test. The threshold for statistical significance was set at $p < 0.05$ in all of the tests.

2.3. Machine Learning Approach

The available dataset is characterized by the non-negligible presence of missing values. Therefore, data imputation of independent variables was first performed to exploit all of the instances and obtain more stable and reliable results [17]. In the dataset, different data types are available and each of them was treated with dedicated techniques. In the case of numerical variables, for each instance with at least one missing value a small subset of complete instances similar to the instance under investigation was selected. From this set, a randomly sampled instance was used to replace the missing values. Discrete variables were instead imputed with ad hoc models: Logistic regression models for binary variables, multinomial logistic regression models for unordered categorical variables and ordered logit models (or proportional odds models) for categorical variables with ordered categories. The data imputation was implemented with the “mice” R package [18].

A multivariate classification model was used to predict the target variables: (1) Single ODI improvement or (2) combined ODI + SF36 PCS + COMI Back. For both targets, we used a random forest (RF) classification method to predict the outcome of the surgical operation. RF is an ensemble model composed of multiple decision trees, each of them trained independently on randomly sampled subsets of variables. The single outputs of the multiple decision tree models are then combined with a majority vote to obtain the final decision of RF. This ensemble helps in improving the predictive performance of the individual decision tree models. Indeed, RF has been recognized as one of the best performing classifiers in extensive classification studies [19], and the R implementation provided by the “randomForest” package is empirically more accurate than other implementations [20]. We thus train a RF model using the default settings of the “randomForest” R package for both targets. To evaluate the most important features used by RF to classify instances, we used the mean decrease Gini index, which measures the contribution of each variable to the homogeneity of internal and leaf nodes of the tree.

We trained RF in cross-validation (five folds) and selected the classification threshold for each fold by optimizing the geometric mean of sensitivity and specificity in a nested cross-validation loop. The proposed nested cross-validation allows to robustly estimate the optimal classification threshold and assess RF performance, while balancing sensitivity and specificity. This is particularly relevant since both the target variables are slightly unbalanced (70.7% of the available data are associated with a good surgical outcome and 43.3% of the available data are associated with an excellent surgical outcome).

The model for predicting ODI (good early outcome) at 6 months’ follow up (FU) makes use of the following features: Classification of the patient’s clinical state (Glassman), equipte operating (Equipte), age, gender, body mass index (BMI), ASA code (ASA), pre-operative medical PROMs (ODI, COMI, SF-36 Physical and SF-36 Mental), number of vertebrae stabilized during the operation (Levels), start and end points of the stabilized vertebrae (From_Level, To_Level) and comorbidity Charlson Index (Charlson).

The performance of the model is evaluated in terms of sensitivity, i.e., true positive rate (TPR), specificity, i.e., true negative rate (TNR), accuracy and area under the receiver operating characteristic curve (AUC ROC). The exploratory and further machine learning analysis were performed in R [21].

The entire study was performed according to the TRIPOD guideline for the development of multivariate models for individual prognosis or diagnosis [22].

3. Results

A total of 1243 patients who underwent lumbar arthrodesis surgery were included in this study. Out of them, 9.5% were disc pathologies, 38.4% were disc collapse, 32.8% were spondylolysis or spondylolisthesis, 7.6% were degenerative scoliosis, 0.1% were facet pathologies, 11.1% were non-union, 0.3% were cancer and 0.2% were infection. The rate of

early good outcome was 70.7% (n = 879). A total of 43.3% (n = 538) of patients reached an “excellent” early outcome. The patients had a median age of 56 (interquartile range: 22) years and 771 (62.0%) were female. The mean baseline disability of the study population was ODI 47.3 ± 17.1, the mean pain score was COMI back 7.7 ± 1.7 and the mean quality of life was SF-36 PCS 32.7 ± 6.9 and SF-36 MCS 45.5 ± 11.8.

Since univariate exploratory analysis did not collect significant results, multivariate classification models were used to identify both surgical operations with good and excellent early outcome; the results of these analyses are reported in the Supplementary Materials Tables S1–S11.

Machine Learning Model

The model showed a good balance between sensitivity (74.3%) and specificity (79.4%), while providing a good accuracy (75.8%) with ROC AUC = 0.842.

This combined target model (excellent early outcome) makes use of the same features used for the good early outcome model. The excellent early outcome model showed a sensitivity of 74.2% and specificity of 71.8%, with an accuracy of 72.8%, and a ROC AUC = 0.808. Furthermore, both models for predicting good and excellent clinical outcomes showed a good balance between sensitivity (74.3% for good and 74.2% for excellent outcomes) and specificity (79.4% for good and 71.8% for excellent outcomes), while providing a good accuracy (75.8% for good and 72.8% for excellent outcomes). Details are reported in Tables 2 and 3.

Table 2. Good clinical outcome AI predictions.

	Outcome +	Outcome –	Total		
Prediction +	653	75	728	Sensitivity	74.3%
Prediction –	226	289	515	Specificity	79.4%
Total	879	364	1243	PPV	89.7%
				NPV	56.1%
				Accuracy	75.8%
				AUC ROC	0.842

The table shows the performance evaluation of the AI model predicting the “good clinical outcome”—ODI at 6 months FU. Outcome +: patients with ODI Δ ≥ 12.7/100; Outcome –: patients with ODI Δ < 12.7/100; Prediction +: model’s predictions of patients with ODI Δ ≥ 12.7/100; Outcome –: model’s predictions of patients with ODI Δ < 12.7/100; PPV: Positive Predictive Values; NPV: Negative Predictive Values.

Table 3. Excellent clinical outcome AI predictions.

	Outcome +	Outcome –	Total		
Prediction +	399	199	598	Sensitivity	74.2%
Prediction –	139	506	645	Specificity	71.8%
Total	538	705	1243	PPV	66.7%
				NPV	78.4%
				Accuracy	72.8%
				AUC ROC	0.808

The table shows the performance evaluation of the AI model predicting the “Excellent Clinical Outcome”—ODI—SF36—COMI Back at 6 months FU. Outcome +: patients with ODI Δ ≥ 12.7/100 and SF-36 PCS (Δ ≥ 6/100) and COMI back (Δ ≥ 2.2/10); Outcome –: patients with at least one of the following conditions ODI Δ < 12.7/100 or SF-36 PCS Δ < 6/100 or COMI back Δ < 2.2/10; Prediction +: model’s predictions of patients with ODI Δ ≥ 12.7/100 and SF-36 PCS Δ ≥ 6/100 and COMI back Δ ≥ 2.2/10; Outcome –: model’s predictions of patients with at least one of the following conditions—ODI Δ < 12.7/100 or SF-36 PCS Δ < 6/100 or COMI back Δ < 2.2/10; PPV: Positive Predictive Values; NPV: Negative Predictive Values

The models also showed a good discriminatory capacity of the two classes (ROC AUC = 0.842 for good and ROC AUC = 0.808 for excellent outcome). See Figure 1.

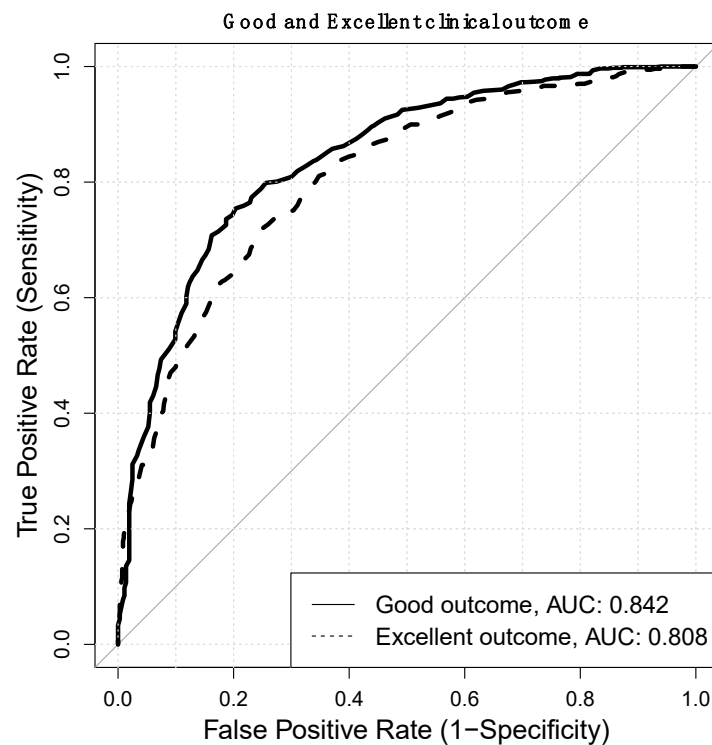


Figure 1. ROC CURVE for the ODI model and the combined model.

According to the mean decreasing Gini index of the random forest model, the top five predictors of both good and excellent clinical outcomes were SF-36 PCS, SF-36 MCS, ODI, BMI and age at baseline (the weights of machine learning models for good clinical outcomes were: SFMPre = 73.20, SFPPre = 70.80, ODIPre = 66.77, BMI = 62.97 and Age = 61.12; for excellent clinical outcomes they were: SFPPre = 90.34, SFMPre = 87.13, BMI = 78.61, Age = 69.92 and ODIPre = 66.21 (Table 4).

Table 4. Mean decreased Gini weights in good and excellent clinical outcomes.

Predictive Variables	Good Clinical Outcome	Predictive Variables	Excellent Clinical Outcome
SFMPre	73.20	SFPPre	90.34
SFPPre	70.80	SFMPre	87.13
ODIPre	66.77	BMI	78.61
BMI	62.97	Age	69.92
Age	61.12	ODIPre	66.21
COMIPre	31.00	COMIPre	43.74
Glassman	28.30	Glassman	29.90

Glassman: Glassman classification data; BMI: body mass index; ODIPre: pre-operative Oswestry disability index; COMIPre: pre-operative core outcome measures index; SFPPre: pre-operative physical component score of the short form-36; SFMPre: pre-operative mental component score of the short form-36.

The graphs of mean decreased Gini for weights of machine learning models, as well as odds ratios of the explorative analysis, are included in the Supplementary Materials Figures S1 and S2.

4. Discussion

In spine surgery, multiple factors can influence clinical outcomes. According to our results and the explorative analysis, there is not a single risk factor capable of influencing or predicting early clinical outcomes. In our registry, the machine learning (ML) approach

predicts the likelihood of good or excellent early clinical results. Our ML model showed good performances of post-operative prediction if based on patients' demographic data and pre-operative self-reported degree of disability and quality of life. In the last decades, the machine learning approach in predictive models has gained interest in clinical practice.

4.1. Predictive Models of Surgical Improvement Based on Clinical Data

Surgical treatment for degenerative spine disorders has been shown to improve the quality of life and reduce disability in patients most severely affected [23]. Nevertheless, the associations between demographic baseline factors and overall complication rates are still unclear.

Thanks to the anesthesiological and surgical implementations, a large population can safely be a spinal surgery candidate. The increase in the use of mini-invasive surgery [24] and the improvements of pre-operative planning methods [25] has allowed enlarging the cohort of patients eligible for surgery and capable of obtaining significant results [26,27].

One of the most relevant demographic indicators is BMI (body mass index). High BMI values are known to be risk factors for many diseases and are directly correlated with the complication rate after spinal surgery, even if in the literature, BMI's role is still debated in the prediction of functional outcomes.

According to Mulvanay et al. [28], an increased BMI is associated with decreased effectiveness of one- to three-level elective lumbar fusion, despite the absence of surgical complications. A BMI value higher than 30 is considered a risk factor for surgical complications and poor spine surgery results. According to our data, a low BMI seems to present a relevant role in predicting good clinical outcomes. Despite several studies suggesting that weight does not represent a major impact on patients' health-related quality of life after surgery [29], obesity has a relevant impact on intraoperative blood loss, length of surgery and complication rate. It seems that BMI should always be kept in mind when planning spinal fusion. Several clinical indicators of postoperative success are continuously analyzed to improve the surgical outcome in terms of complication rate and patients' satisfaction. The aging of the population and the relative increase of the comorbidities can challenge the surgical decision.

According to Daniels et al. [30], the upgrading of some surgical methods increased the performances of therapeutic strategies. In particular, in a retrospective analysis of surgical cases enrolled between 2009 and 2016, the complication rates decreased over time, despite an increasingly elderly, medically compromised, and obese patient population. As a critical point, the authors identified the evolution of surgical strategies that resulted in an overall improvement of the treatment quality.

4.2. Predictive Models of Surgical Improvement Based on PROMs

The proper estimation of the pre-operative degree of disability and quality of life is mandatory when surgery is required. In spine surgery, considerable controversy exists regarding spinal arthrodesis' risk–benefit ratio where surgery itself creates a permanent fusion of vertebral bodies. Nevertheless, several studies demonstrated a significant improvement after spinal arthrodesis in cases of degenerative spinal disorders [26,31,32]. A combination of scales is often used in clinical studies to assess multiple aspects of human health.

The indicators of quality of life and disability progressively gained attention, becoming the gold standard to measure the success rate after spine surgery. The post-operative clinical improvement can be evaluated based on patients' reported outcomes such as ODI. Although post-operative improvement may be statistically significant, it is not necessarily clinically relevant. For this reason, several studies have defined the values used to indicate a difference that is clinically meaningful to the patient (MCID) [33]. In particular, Monticone et al. defined as significant a cut-off value of MCID at a 12.7 ODI unit score of improvement [14].

4.3. Predictive Models' Performances

Predictive models for patient-reported outcomes can improve the surgical strategies when deciding to opt for surgery or not or potentially to adapt the surgical approach.

Despite the significant variability in the population affected by a common clinical condition, lumbar disc herniation, Staartjes et al. [34] proposed a predictive model based on deep learning-based analytics. Out of a population of 422 patients, the deep learning and logistic regression attained AUC values of 0.84 and 0.72, and accuracies of 75% and 59%, respectively. The greatest discrepancy in performance measures regarded the models predicting back pain improvement. This could reflect the model's weakness or the inherent difficulty of the outcome to be measured.

Models based on naïve Bayes machine learning to predict hospitalization days and indications for discharge (for example, admission to rehabilitation facilities or back to home and hospitalization costs) showed high performances. In particular, the system proposed by Karnuta et al. revealed a predictive accuracy of 0.800 for costs of recovery, 0.874 for length of stay (LOS), and 0.878 for disposition with AUC for hospitalization costs (0.880), an excellent AUC for LOS (0.941), and an excellent AUC for discharge disposition (0.906) [35]. The disease variability, combined with the psychological influencing factors and patients' expectations related to the surgeries, challenges the accuracy of clinical predictions. Siccoli et al. [36] evaluated the feasibility of short- and long-term PROMs and reoperation rate using an ML approach in patients affected by lumbar stenosis. According to this study, the models were able to predict the endpoints, providing accurate information. Despite a progressive increase in the use of prediction models in spine surgery, little is known in spinal arthrodesis for two to four levels surgery.

Our results seem to provide comparable or higher predictive performances than other studies on spine surgery. Thanks to the recent advances in technologies, AI can involve the application of mathematical algorithms that continuously learn and make observations from existing data. The aim is to create a more accurate predictive model based on these data [37].

4.4. Influence of ML Predictions on Therapeutic Strategy

With the widening of the modern dataset, the use of ML will progressively become the gold standard and the primary candidate for the data analysis. Future application in diagnosis, prognosis and decision-making processes is desirable and will soon become an essential spine physician tool. Khan et al. introduced the application in the clinical management of cervical myelopathy and nontraumatic spinal cord injury to predict the risk of neurological impairment at one year [38]. These tools allowed the physicians to predict individual patient outcome after surgery for degenerative cervical myelopathy [39] and to apply preventive strategies such as targeted physiotherapy and the timing of psychological counselling. With the application of ML techniques, several studies demonstrated the possibility to predict clinical outcomes. Ames et al. predicted patients' responses to SRS-22R (questionnaire) item per item up to 86.9% AUROC at 1 and 2 years following surgical treatment for ASD. The main clinical application is to aid surgical decision making during preoperative counselling [40]. In complex surgery, this approach will be capable of implementing already available surgical decision making [41].

4.5. Methodological Consideration and Limitations

The result of our study comes with several limitations that we have to take into account. First, the term follow-up indicators for 6 months can be considered only a preliminary result. External and prospective validations are necessary to support this methodology further so as to improve the knowledge acquired. Furthermore, the lower performance in terms of PPV in "excellent outcomes predictions" and NPV in "good outcomes predictions" can be explained by low numbers of positive and negative events, respectively. Indeed, the PPV and NPV values for the two problems show that both models perform better on the majority class (positive for good clinical outcome and negative for excellent clinical

outcome). This is especially true for the “good outcome prediction”, where the imbalance is higher. In the “excellent outcome prediction”, the PPV, although lower than NPV, still highlights the model’s ability to predict the positive class. These results highlight that some predictions, i.e., negative good clinical outcome and positive excellent clinical outcome, are more difficult than the opposites, i.e., positive good clinical outcome and negative excellent clinical outcome. The predictive ability can be improved by combining the two models to obtain a classification of patients in three categories: “Excellent”, “Good” and “Not good”. This classification of patients can be used to support clinicians in making personalized and patient-specific decisions.

Although a significant role of different surgical approaches has not been identified, further studies are needed to clarify the role of different surgical techniques on medium- and long-term clinical outcomes.

5. Conclusions

The results of our study suggest that a machine learning approach showed high performance in predicting early good to excellent clinical results. In particular, our data suggest that with a worse score of preoperative indicators of disability and quality of life, younger or healthier patients should expect a significant clinically relevant improvement. On the other hand, older patients and patients with higher BMI, comorbidities (higher ASA and Charlson score) with higher SF-36 scores and lower ODI scores would experience less clinically relevant improvements by following the path of lumbar spine surgery. These results must be seen in light of the study’s limitations—first, the mid-term follow-up indicators, six months. A potential improvement or worsening in the PROMS results could occur later. The latter was not the focus of the study.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/jpm11121377/s1>, Figure S1: Mean Decrease Gini for Good Clinical Outcome, Figure S2: Mean Decrease Gini for Excellent Clinical Outcome, Table S1: Table 2 × 2 for Good Clinical Outcome, Table S2: Odds ratio for Good Clinical Outcome, Table S3: Chi-square test for Good Clinical Outcome, Table S4: Summary table of the chi-square test for Good Clinical Outcome, Table S5: Table 2 × 2 for Excellent Clinical Outcome, Table S6: Odds ratio for Excellent Clinical Outcome, Table S7: Chi-square test for Excellent Clinical Outcome, Table S8: Summary table of the chi-square test for Excellent Clinical Outcome, Table S9: Paired *t*-test for Good Clinical Outcome, Table S10: Paired *t*-test for Excellent Clinical Outcome, Table S11: Wilcoxon test for Good and Excellent Clinical Outcome.

Author Contributions: Conceptualization, F.G., F.L., G.L. and P.B. (Pedro Berjano); methodology, F.L. and L.A.; software, L.A., L.V. and D.H.; validation, A.C., F.G., T.B. and L.A.; formal analysis, L.V., D.H., F.M., G.L. and L.A.; investigation, P.B. (Paolo Barletta) and J.H.V.; resources, P.B. (Paolo Barletta); data curation, P.B. (Paolo Barletta); writing—original draft preparation, F.L., D.C., L.M., P.B. (Pedro Berjano) and L.A.; writing—review and editing, L.S., R.B. and M.B.-B.; visualization, P.B. (Pedro Berjano); supervision, P.B. (Pedro Berjano); project administration, F.G., L.M. and P.B. (Pedro Berjano); funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Italian Ministry of Health (Project Code: CO-2016-02364645). The funders had no involvement in study design, collection, analysis and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication. The data that were used for this study were collected from the SpineReg electronic register (IOG SpineReg version 1.7.4, Deloitte, Milan, Italy).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of IRCCS Istituto Ortopedico Galeazzi (Fourth Amendment to the SPINEREG Protocol, Issued on 10 October 2019).

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets used and/or analyzed in the present study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no competing interest.

References

- Schwab, F.; Dubey, A.; Gamez, L.; El Fegoun, A.B.; Hwang, K.; Pagala, M.; Farcy, J.-P. Adult Scoliosis: Prevalence, SF-36, and Nutritional Parameters in an Elderly Volunteer Population. *Spine* **2005**, *30*, 1082–1085. [[CrossRef](#)] [[PubMed](#)]
- Pellis e, F.; Vila-Casademunt, A.; Ferrer, M.; Domingo-S abat, M.; Bago, J.; P erez-Grueso, F.J.S.; Alanay, A.; Mannion, A.F.; Acaroglu, E. Impact on health related quality of life of adult spinal deformity (ASD) compared with other chronic conditions. *Eur. Spine J.* **2014**, *24*, 3–11. [[CrossRef](#)]
- Weinstein, J.N.; Lurie, J.D.; Tosteson, T.D.; Zhao, W.; Blood, E.A.; Tosteson, A.N.; Birkmeyer, N.J.O.; Herkowitz, H.N.; Longley, M.C.; Lenke, L.G.; et al. Surgical Compared with Nonoperative Treatment for Lumbar Degenerative Spondylolisthesis. four-year results in the Spine Patient Outcomes Research Trial (SPORT) randomized and observational cohorts. *J. Bone Jt. Surg. Am. Vol.* **2009**, *91*, 1295–1304. [[CrossRef](#)]
- Dagenais, S.; Caro, J.; Haldeman, S. A systematic review of low back pain cost of illness studies in the United States and internationally. *Spine J.* **2008**, *8*, 8–20. [[CrossRef](#)] [[PubMed](#)]
- Zanirato, A.; Damilano, M.; Formica, M.; Piazzolla, A.; Lovi, A.; Villafa e, J.H.; Berjano, P. Complications in adult spine deformity surgery: A systematic review of the recent literature with reporting of aggregated incidences. *Eur. Spine J.* **2018**, *27*, 2272–2284. [[CrossRef](#)]
- Ferguson, T.B., Jr. The Institute of Medicine Committee Report “Best Care at Lower Cost: The Path to Continuously Learning Health Care”. *Circ. Cardiovasc. Qual. Outcomes* **2012**, *5*, e93–e94. [[CrossRef](#)] [[PubMed](#)]
- McGirt, M.J.; Sivaganesan, A.; Asher, A.L.; Devin, C.J. 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* **2015**, *39*, E13. [[CrossRef](#)]
- Sinikallio, S.; Aalto, T.; Airaksinen, O.; Lehto, S.M.; Kr oger, H.; Viinam aki, H. Depression Is Associated with a Poorer Outcome of Lumbar Spinal Stenosis Surgery: A two-year prospective follow-up study. *Spine* **2011**, *36*, 677–682. [[CrossRef](#)]
- Archer, K.R.; Devin, C.J.; Vanston, S.W.; Koyama, T.; Phillips, S.E.; George, S.Z.; McGirt, M.J.; Spengler, D.M.; Aaronson, O.S.; Cheng, J.S.; et al. Cognitive-Behavioral-Based Physical Therapy for Patients with Chronic Pain Undergoing Lumbar Spine Surgery: A Randomized Controlled Trial. *J. Pain* **2016**, *17*, 76–89. [[CrossRef](#)] [[PubMed](#)]
- McGirt, M.J.; Bydon, M.; Archer, K.R.; Devin, C.J.; Chotai, S.; Parker, S.L.; Nian, H.; Harrell, F.E.; Speroff, T.; Dittus, R.S.; et al. An analysis from the Quality Outcomes Database, Part 1. Disability, quality of life, and pain outcomes following lumbar spine surgery: Predicting likely individual patient outcomes for shared decision-making. *J. Neurosurg. Spine* **2017**, *27*, 357–369. [[CrossRef](#)]
- Langella, F.; Barletta, P.; Baroncini, A.; Agarossi, M.; Scaramuzza, L.; Luca, A.; Bassani, R.; Peretti, G.M.; Lamartina, C.; Villafa e, J.H.; et al. The use of electronic PROMs provides same outcomes as paper version in a spine surgery registry. Results from a prospective cohort study. *Eur. Spine J.* **2021**, *30*, 2645–2653. [[CrossRef](#)]
- Charlson, M.E.; Pompei, P.; Ales, K.L.; MacKenzie, C.R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **1987**, *40*, 373–383. [[CrossRef](#)]
- Glassman, S.D.; Carreon, L.Y.; Anderson, P.A.; Resnick, D.K. A diagnostic classification for lumbar spine registry development. *Spine J.* **2011**, *11*, 1108–1116. [[CrossRef](#)] [[PubMed](#)]
- Monticone, M.; Baiardi, P.; Vanti, C.; Ferrari, S.; Pillastrini, P.; Mugnai, R.; Foti, C. Responsiveness of the Oswestry Disability Index and the Roland Morris Disability Questionnaire in Italian subjects with sub-acute and chronic low back pain. *Eur. Spine J.* **2011**, *21*, 122–129. [[CrossRef](#)] [[PubMed](#)]
- Copay, A.G.; Glassman, S.D.; Subach, B.R.; Berven, S.; Schuler, T.C.; Carreon, L.Y. Minimum clinically important difference in lumbar spine surgery patients: A choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and Pain Scales. *Spine J.* **2008**, *8*, 968–974. [[CrossRef](#)] [[PubMed](#)]
- Mannion, A.F.; Porchet, F.; Kleinst uck, F.S.; Lattig, F.; Jeszenszky, D.; Bartanusz, V.; Dvorak, J.; Grob, D. The quality of spine surgery from the patient’s perspective: Part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index. *Eur. Spine J.* **2009**, *18*, 374–379. [[CrossRef](#)]
- Harel, O.; Mitchell, E.M.; Perkins, N.; Cole, S.R.; Tchetgen, E.J.T.; Sun, B.; Schisterman, E. Multiple Imputation for Incomplete Data in Epidemiologic Studies. *Am. J. Epidemiol.* **2018**, *187*, 576–584. [[CrossRef](#)]
- Van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [[CrossRef](#)]
- Fern andez-Delgado, M.; Cernadas, E.; Barro, S.; Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **2014**, *15*, 3133–3181. Available online: <http://hdl.handle.net/10347/17792> (accessed on 9 December 2021).
- Bagnall, A.; Cawley, G.C. On the Use of Default Parameter Settings in the Empirical Evaluation of Classification Algorithms. *arXiv* **2017**, arXiv:1703.06777, in preprint.
- Bunn, M.K.A. An Introduction to dpIR. *Ind. Commer. Train.* **2008**, *10*, 11–18.
- Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G.M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Med.* **2015**, *13*, 1. [[CrossRef](#)] [[PubMed](#)]

23. Reid, D.B.; Daniels, A.H.; Ailon, T.; Miller, E.; Sciubba, D.M.; Smith, J.S.; Shaffrey, C.I.; Schwab, F.; Burton, D.; Hart, R.A.; et al. Frailty and Health-Related Quality of Life Improvement Following Adult Spinal Deformity Surgery. *World Neurosurg.* **2018**, *112*, e548–e554. [[CrossRef](#)]
24. Berjano, P.; Lamartina, C. Far lateral approaches (XLIF) in adult scoliosis. *Eur. Spine J.* **2012**, *22*, 242–253. [[CrossRef](#)] [[PubMed](#)]
25. Langella, F.; Villafañe, J.H.; Damilano, M.; Cecchinato, R.; Pejrona, M.; Ismael, M.; Berjano, P. Predictive Accuracy of Surgimap Surgical Planning for Sagittal Imbalance: A Cohort Study. *Spine* **2017**, *42*, E1297–E1304. [[CrossRef](#)] [[PubMed](#)]
26. Berjano, P.; Langella, F.; Ismael, M.-F.; Damilano, M.; Scopetta, S.; Lamartina, C. Successful correction of sagittal imbalance can be calculated on the basis of pelvic incidence and age. *Eur. Spine J.* **2014**, *23*, 587–596. [[CrossRef](#)]
27. Yamato, Y.; Hasegawa, T.; Kobayashi, S.; Yasuda, T.; Togawa, D.; Arima, H.; Oe, S.; Iida, T.; Matsumura, A.; Hosogane, N.; et al. Calculation of the Target Lumbar Lordosis Angle for Restoring an Optimal Pelvic Tilt in Elderly Patients with Adult Spinal Deformity. *Spine* **2016**, *41*, E211–E217. [[CrossRef](#)]
28. Mulvaney, G.; Rice, O.M.; Rossi, V.; Peters, D.; Smith, M.; Patt, J.; Pfortmiller, D.; Asher, A.L.; Kim, P.; Bernard, J.; et al. Mild and Severe Obesity Reduce the Effectiveness of Lumbar Fusions: 1-Year Patient-Reported Outcomes in 8171 Patients. *Neurosurgery* **2020**, *88*, 285–294. [[CrossRef](#)]
29. Lingutla, K.K.; Pollock, R.; Benomran, E.; Purushothaman, B.; Kasis, A.; Bhatia, C.K.; Krishna, M.; Friesem, T. Outcome of lumbar spinal fusion surgery in obese patients: A systematic review and meta-analysis. *Bone Jt. J.* **2015**, *97-B*, 1395–1404. [[CrossRef](#)]
30. Tan, G.H.; Goss, B.G.; Thorpe, P.J.; Williams, R.P. CT-based classification of long spinal allograft fusion. *Eur. Spine J.* **2007**, *16*, 1875–1881. [[CrossRef](#)]
31. Berjano, P.; Langella, F.; Damilano, M.; Pejrona, M.; Buric, J.; Ismael, M.; Villafañe, J.H.; Lamartina, C. Fusion rate following extreme lateral lumbar interbody fusion. *Eur. Spine J.* **2015**, *24*, 369–371. [[CrossRef](#)]
32. Cecchinato, R.; Langella, F.; Bassani, R.; Sansone, V.; Lamartina, C.; Berjano, P. Variations of cervical lordosis and head alignment after pedicle subtraction osteotomy surgery for sagittal imbalance. *Eur. Spine J.* **2014**, *23*, 644–649. [[CrossRef](#)]
33. Carragee, E.J.; Cheng, I. Minimum acceptable outcomes after lumbar spinal fusion. *Spine J.* **2010**, *10*, 313–320. [[CrossRef](#)] [[PubMed](#)]
34. Staartjes, V.E.; de Wispelaere, M.P.; Vandertop, W.P.; Schröder, M.L. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: Feasibility of center-specific modeling. *Spine J.* **2019**, *19*, 853–861. [[CrossRef](#)]
35. Karnuta, J.M.; Golubovsky, J.L.; Haeberle, H.; Rajan, P.V.; Navarro, S.; Kamath, A.F.; Schaffer, J.L.; Krebs, V.E.; Pelle, D.W.; Ramkumar, P.N. Can a machine learning model accurately predict patient resource utilization following lumbar spinal fusion? *Spine J.* **2020**, *20*, 329–336. [[CrossRef](#)]
36. Siccoli, A.; De Wispelaere, M.P.; Schröder, M.L.; Staartjes, V.E. Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurg. Focus* **2019**, *46*, E5. [[CrossRef](#)] [[PubMed](#)]
37. Lee, M.S.; Grabowski, M.M.; Habboub, G.; Mroz, T.E. The Impact of Artificial Intelligence on Quality and Safety. *Glob. Spine J.* **2020**, *10*, 99S–103S. [[CrossRef](#)]
38. Khan, O.; Badhiwala, J.H.; Witiw, C.D.; Wilson, J.R.; Fehlings, M.G. Machine learning algorithms for prediction of health-related quality-of-life after surgery for mild degenerative cervical myelopathy. *Spine J.* **2021**, *21*, 1659–1669. [[CrossRef](#)] [[PubMed](#)]
39. Merali, Z.G.; Witiw, C.D.; Badhiwala, J.H.; Wilson, J.R.; Fehlings, M.G. Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. *PLoS ONE* **2019**, *14*, e0215133. [[CrossRef](#)] [[PubMed](#)]
40. Ames, C.P.; European Spine Study Group; Smith, J.S.; Pellisé, F.; Kelly, M.; Gum, J.L.; Alanay, A.; Acaroğlu, E.; Pérez-Grueso, F.J.S.; Kleinstück, F.S.; et al. Development of predictive models for all individual questions of SRS-22R after adult spinal deformity surgery: A step toward individualized medicine. *Eur. Spine J.* **2019**, *28*, 1998–2011. [[CrossRef](#)] [[PubMed](#)]
41. Obeid, I.; Berjano, P.; Lamartina, C.; Chopin, D.; Boissière, L.; Bourghli, A. Classification of coronal imbalance in adult scoliosis and spine deformity: A treatment-oriented guideline. *Eur. Spine J.* **2019**, *28*, 94–113. [[CrossRef](#)] [[PubMed](#)]